Sampling bias in baseline endemicity classification surveys for lymphatic filariasis in sub-Saharan Africa

Kevin Kwong

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2019

Committee:

Elizabeth Cromwell

David Pigott

Program Authorized to Offer Degree:

Global Health

University of Washington

Abstract

Sampling bias in baseline endemicity classification surveys for lymphatic filariasis in sub-Saharan Africa

Kevin Kwong

Chair of the Supervisory Committee:

Elizabeth Cromwell

Department of Health Metrics Sciences

Mapping surveys were conducted prior to the scale up of national programs to eliminate lymphatic filariasis in order to define populations-at-risk. Usually, relatively few locations were surveyed for this process and some sites were purposively sampled from areas with high LF transmission risk. This survey design potentially led to sampling bias in mapping data that, when used in secondary analyses, may lack representativeness. We present the first quantitative assessment of sampling bias in LF mapping surveys across multiple countries. First, a list of 17 socio-ecological covariates commonly used in secondary analyses was compiled. Then, simulations were performed to determine covariate distributions characteristic of a simulated set of randomly selected communities. These were compared to distributions from observed mapping survey locations to test for evidence of systematic sampling bias. Finally, using country mapping datasets, the probability a location was sampled was modelled as a function of socio-ecological covariates using boosted regression trees. We found statistically significant covariate pattern differences potentially indicative of sampling bias in Benin, Cote d'Ivoire, Ghana, and Togo. Non-random differences in sampling along these commonly used covariates suggest

these locations may not be representative of inhabited areas in these countries and use of these data in secondary analyses should explore the potential impact of bias.

Introduction

Lymphatic filariasis (LF) is a tropical, parasitic disease that can, with chronic infection, lead to irreversible damage to lymphatic systems. At its most severe, LF disfigures and disables in the form of hydrocele or lymphedema. Historically, LF transmission occurred across much of sub-Saharan Africa, Southeast Asia, and Oceania. It is caused by three species of thread-like filarial worms. Of these, *Wuchereria bancrofti* is the most common, accounting for ~90% of cases across all regions. *Brugia malayi* and *Brugia timori* account for the remainder but are not circulated in sub-Saharan Africa. Prior to the Global Programme to Eliminate LF (GPELF), the WHO estimated 120 million people were affected globally by LF [1].

In 2000, when the GPELF began, data on whether interventions were required was limited or out of date in many parts of the world, particularly in sub-Saharan Africa [1]. To facilitate planning of baseline surveys, the World Health Organization (WHO) and other partners recommended guidelines for “mapping” areas to determine where LF was endemic [2]. This process was used to define the geographic extent of LF transmission across implementation units (IU) (typically second or third order administrative units such as districts or counties), enabling national programs to quantify populations at risk eligible for preventative chemotherapy (PC) via mass drug administration (MDA).

These guidelines recommended purposive sampling of relatively few survey sites with high LF transmission risk in place of a geographically representative strategy [2,3]. This design choice made the process of mapping endemicity relatively cheap and simple in order to encourage national elimination programs to scale up quickly [2,3]. In many countries, survey administrators collected and consulted historical data, administrative data, and expert opinion during survey site selection. Data collected after 1990 and judged to be of sufficient quality were used to differentiate IUs as clearly endemic, clearly non-endemic, and inconclusive. New surveys were typically conducted in IUs for which endemicity status was inconclusive. Usually two locations

were surveyed per IU, with a sample size of around 100 individuals per site. The WHO recommended that at least one of these sites be chosen in villages previously identified by local health administrators as at high risk of LF transmission. If measured prevalence exceeded 1% in any of the surveyed villages, an IU was considered “endemic” and eligible for MDA. IUs below this threshold were considered “non-endemic”.

Since 2000, this strategy has been used successfully to map LF endemicity in thousands of districts, rapidly delineating districts in need of MDA [4]. Mapping surveys also generated valuable information on baseline prevalence that has since been used in many secondary analyses: these data have been used to model country-specific environmental correlates of LF transmission [5], to create maps of environmental suitability [6], and to model LF prevalence [7,8,9]. Mapping data has also been used to estimate baseline prevalence and population-at-risk in order to simulate the impact of different MDA scenarios [10,11].

Secondary analyses differ in their consideration of potential sampling bias in the mapping data. Sampling bias occurs when locations chosen for surveys are not representative of the unit of inference (such as the IU in this context). Large cluster random survey designs are commonly used to decrease the risk of sampling bias. The LF endemicity mapping design differs from this approach, relying on selection of a few communities per IU, and may be at higher risk of sampling bias.

Skewed or spurious results can arise in secondary analyses if sampling bias exists and sampling probability is associated with the independent variables (ex: temperature) and the dependent variable (ex: LF prevalence). Consider a hypothetical positive association between temperature and LF prevalence. If a simple ordinary least squares (OLS) model is applied to data biased such that data from locations where temperature is low is systematically under-represented in the dataset, both estimates of slope coefficients and intercept may be biased and can lead to inaccurate predictions. This example is visualized below. In it, modelled intercept is

higher in the biased sample and modelled coefficient is smaller. The model would over-predict prevalence at low temperature locations.

Figure 1:

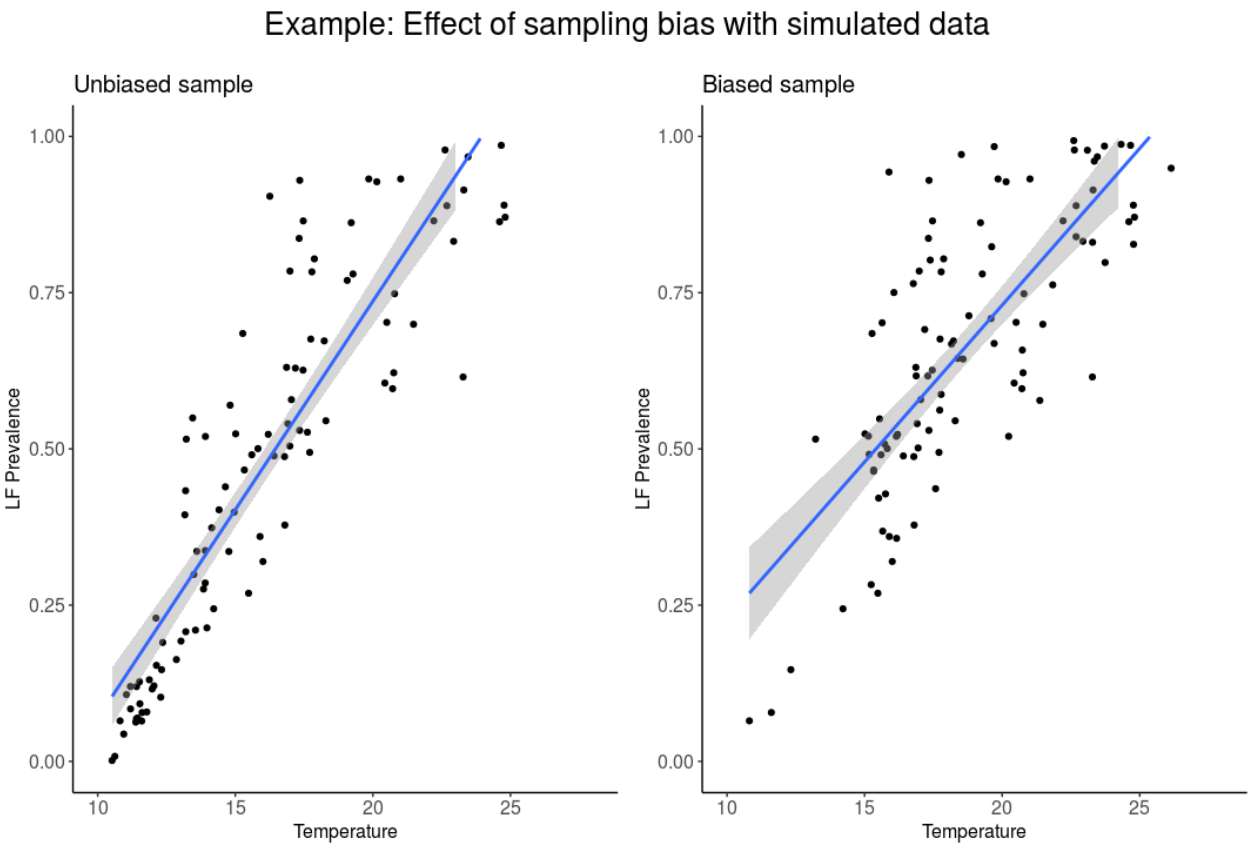

Figure 1: An illustration of how inference & prediction may be affected when using data with sampling bias. Both plots have 100 observations from the same hypothetical joint distribution. The biased sample was created by omitting a percentage of observations at low temperatures and drawing new observations at higher temperatures.

By definition, purposive sampling aims to correlate sampling probability with prevalence since high LF transmission areas are more likely to be sampled. Thus, secondary analyses using biased mapping data may lead to results distorted by sampling bias. Sampling bias also decreases generalizability of measurements. When used to calculate baseline IU prevalence &

population-at-risk for example, the result would be an overestimate. This subsequently can lead to inaccurate simulations of intervention effects.

Some researchers recognize the potential of sampling bias to skew their results and implemented adjustments such as through simulation of pseudo-absence data in areas known to be unsuitable for mosquito breeding or non-endemic for LF [8]. Others do not factor potential bias into the design or interpretation of their research. At present, no studies have systematically tested country mapping datasets for the presence of bias.

In this study, we present the first attempt to identify and characterize sampling bias in LF endemicity mapping across 19 sub-Saharan countries. In each country, we analyzed the national representativeness of mapping survey locations across a variety of socio-ecological dimensions to (1) identify country mapping datasets with clear evidence of sampling bias; (2) explicitly model sampling probability as a function of socio-ecological factors, producing descriptive maps of sampling bias where present. This analysis may aid researchers when designing secondary analyses using LF mapping data to identify settings where adjustments for sampling bias is necessary.

Methods

Our analytical approach has two stages. First, for each country, we conduct a simulation study in which communities are randomly selected and compared to reported survey locations to test for the presence of clear differences in the distribution of socio-ecological covariates. Consistent differences are potentially indicative of a biased sampling process, which resulted in sampled locations unrepresentative of IUs' entire socio-ecological spectrum. We then modeled the probability a location could have been sampled for mapping surveys as a function of socio-ecological covariates using an ensemble boosted regression trees (BRT) approach. Finally, modelled predictions and covariate relationships were analyzed by country and across countries to identify correlates of sampling probability.

Data

Geo-positioned, Sub-Saharan Africa mapping survey data were downloaded from the online portal created and maintained by the Expanded Special Project for Elimination of Neglected Tropical Disease (ESPEN) [4]. With the exception of Benin, Burkina Faso, Ghana, and Togo which coordinated and carried out mapping surveys in collaboration [7], this mapping survey data was generated by each country independently and made publically available via the ESPEN. IU boundaries used in this analysis were also downloaded from the ESPEN. These boundaries are fundamental to the definition of sampling bias because they define the unit of inference. We would conclude sampling bias in mapping data if the locations chosen for mapping surveys are not representative of all locations that could have been sampled within an IU.

In order to simulate randomly selected locations for mapping, we required a complete sampling frame of inhabited communities. We chose to use a large database maintained by the United Nations Office for the Coordination of Humanitarian Affairs - Regional Office for West & Central Africa (UN OCHA ROWCA), available through the Humanitarian Data Exchange [12] The method by which the UN OCHA database was assembled is not documented online and is unknown.

Geospatial, gridded socio-ecological variables from a variety of published and online sources were used in this analysis at a 5km X 5km grid resolution (Table 1).

Table 1:

Description	**Source**
Travel time to the nearest settlement >50,000 inhabitants	Malaria Atlas Project, Big Data Institue, Nuffield Department of Medicine, University of Oxford [21]
Aridity Index derived from CRUTS precipitation and evapotranspiration data	Climatic Research Unit Time-Series (CRUTS) 2017). (2014) [22, 23]
CRUTS mean daily temperature	Climatic Research Unit Time-Series (CRUTS) [22, 23]
Distance to river >= 25m wide	Natural Earth Data (derived) [24]
Elevation	NOAA/NCEI [25]
Enhanced Vegetation Index (EVI)	MODIS [26, 27, 28, 29]

Diurnal difference in land surface temperature	MODIS [30]
Population	WorldPop [31]
Precipitation (Multi-source Weighted Ensemble Precipitation)	Princeton Climate Analytics [32]
Distance from roads	Center for International Earth Science Information Network (CIESIN) Columbia University [33]
Tasseled cap brightness index (TCB)	MODIS [26, 27, 28, 29]
Tasseled cap wetness index (TCW)	MODIS [26, 27, 28, 29]
Growing season	FAO [34]
Percent equipped for irrigation	University of Frankfurt [35]
Night time lights	AVHRR [36]
GHSL Urbanicity	European Commission/GHS [37]

Table 1 Data sources for geospatial, gridded variables. A uniform resolution of 5km X 5km was used.

The list of variables included in this analysis were chosen based on relevance to common secondary analyses [5,6,7,8,9]. Ecological variables such as temperature, precipitation, elevation, and Enhanced Vegetation Index (EVI); and socio-demographic variables such as population, access, and nighttime lights; are widely used in LF prevalence and environmental suitability modelling due to known or hypothesized relationships to LF transmission risk. Sampling biases along these variables may increase the possibility of skewed or spurious results in secondary analyses.

Identifying sampling bias

For each country, a simulation of survey site selections was done without replacement by IU, drawing a sample of size N equivalent to the number of mapping surveys present in the dataset for each IU. While the WHO guidelines stipulate selection of 1 or 2 locations per IU, in some settings, survey administrators elect to sample more sites. This is reflected in our simulation. To create socio-ecological profiles of the two samples (simulated and actual), covariate values for 17 geospatial, gridded variables such as population, temperature, and accessibility were extracted at each sampled location (Table 2). The samples' joint multivariate distributions were then compared via multivariate nonparametric hypothesis testing using the cross-match statistical test [13].

The cross-match statistical test is a multivariate alternative to the commonly used Kolmogorov-Smirnov (K-S) test, which has been used in various previous studies on detecting sampling bias in geospatial datasets but does not extend well to multivariate data [14,15]. Both tests are used for two-sample hypothesis testing, assessing the likelihood that two continuous probability distributions (reference and comparison) come from the same underlying distribution. In the cross-match statistical test, reference and comparison data points are combined and paired up such that inter-point Mahalanobis distances across all variables are minimized, creating the "optimal non-bipartite matching". The number of pairs containing observations from both the reference and comparison samples, the cross-match statistic, is calculated. This statistic is low when two samples are very different and high when they are similar. Probability of accepting the null hypothesis, that the samples are drawn from the same distribution, is calculated based on the number of cross-matches and N. The Mahalanobis transformation of inter-point distances accounts for covariance and different scales across variables [13].

We set alpha to be .05. When the p-value falls below this cutoff for a cross-match test, we conclude that mapping survey sites are significantly different from sites selected via simulated random selection, and that sampling bias is potentially present.

To capture uncertainty in our simulation methods, we use a bootstrap approach, repeating this process 1000 times per country. We anticipate greater uncertainty in countries with relatively few mapping surveys, and countries with relatively large IUs each containing a high number of settlements. The latter factor likely increase the variability of socio-ecological distributions in different permutations of randomly selected sites, leading to greater variability in the resultant cross-match statistics. Rejection of the null hypothesis in some bootstraps is insufficient evidence to conclude sampling bias in a country mapping dataset. We only conclude clear sampling bias when a country mapping dataset has a mean p-value across 1000 bootstrap

cross-match tests below .05. This indicates a systematic difference exists in the socio-ecological profile of mapped locations versus the profile expected if IUs were random sampled.

BRT model to predict locations included in LF mapping

For countries where we have found evidence of systematic sampling bias in LF mapping data, we then model country-specific sampling bias using an ensemble boosted regression trees (BRT) method. BRT modelling is commonly used in geospatial analyses, such as for ecological niche modeling in which a species' distribution over space is predicted using a combination of presence and absence records [16,17]. The model exploits differences between the ecological distributions of presence versus absence records to create an index describing the relative probability of the species' presence at any given location. An important feature of BRTs is gradient boosting, a process by which regression tree models are iteratively improved upon by minimizing the variation in residual unexplained variable in the response variable [17]. BRT models have demonstrated success in fitting complex non-linear relationships and is less susceptible to overfitting than comparable approaches [17,18,19].

Our BRT models exploit socio-ecological differences between locations that were sampled ("presence" of sampling) versus randomly simulated locations (pseudo-"absence" of sampling) to predict an index describing a given location's relative probability to have been sampled for mapping surveys. For each country, we fitted 100 BRT models, each using a different draw from simulated pseudo-absence records. We then predicted out relative sampling probability surface from individual BRT models using the 5kmX5km gridded covariate rasters, and calculating the mean ensemble prediction surface.

Each BRT model was fit using the 'dismo' package in the R statistical programming environment with the 'gbm.step' function [39]. This function selects for the optimal number of trees that maximizes cross-validation prediction accuracy. Default values were used for the algorithm's

tuning parameters (tree complexity = 4, learning rate = 0.005, bag fraction = 0.75, step size = 10, cross-validation folds = 10) [17]. If the model fails to converge, learning rate is decreased by 1/10 and model fitting is attempted again.

To better understand sampling biases found, covariate partial dependence functions in the ensemble BRT model were visualized. These show the effect of a covariate on the response variable after accounting for the average effect of all other variables in the model [17].

An ensemble receiver operating characteristic curve (ROC) was calculated to determine the optimal threshold to convert ensemble model predictions to binary classifications for clear illustration of areas most likely to have been surveyed/represented in LF country endemicity mapping processes. The threshold chosen maximizes the true positive rate (correct classification of locations that were sampled) and minimizes the false positive rate (incorrect classification of pseudo-absence records).

Sensitivity Analyses

To assess the potential impact of covariate selection on the outcome of cross-match tests, a sensitivity analysis was completed by which the cross-match analysis was iteratively rerun 10 times in country mapping datasets concluded to have sampling bias; each time with a random selection of 10 covariate from the full list. We expect absence of influential covariates – covariates in which values at surveyed and simulated locations differ greatly – may result in vastly different cross-match statistics. However, reruns should ideally yield comparable results to the analysis run with the full covariate list. This would indicate that our conclusions are robust and not entirely contingent upon covariates selection.

We further validated results from the cross-match analysis by attempting the ensemble BRT modelling process on a subset of countries with average cross-match statistics greater than .05. Theoretically, success in model convergence should be correlated with the cross-match statistic

since both methods are predicated on finding differences in joint covariate distributions, with high cross-match statistic bootstraps more likely to result in BRT model convergence failures.

Results

The ESPEN LF dataset contains 9845 records of which 8645 are labeled mapping surveys. Historical data prior to 2000 were excluded from this analysis. The number of surveys by country used in this analysis and the years of the surveys are described in Table 2.

Table 2:

Country	Total # of locations surveyed	Mapping survey years	# of settlements (# of IUs)	5km coverage of mapping survey locations	Average # of settlements per IU
Benin	151	2000	6265 (77)	95.3%	82
Burkina Faso	286	2000-2002	10656 (70)	96.2%	152
Cameroon	539	2003, 2009, 2013	23753 (189)	97.4%	126
Central African Republic	48	2008	7481 (17)	97.9%	441
Chad	64	2015	17157 (100)	86.5%	171
Cote d'Ivoire	215	2000-2001	8369 (83)	95.5%	101
Democratic Republic of Congo	660	2010-2016	26641 (516)	83.9%	54
Gabon	233	2008, 2014	4600 (51)	87.5%	89
The Gambia	50	2001-2002	3066 (44)	100%	73
Ghana	402	2000-2002	11440 (216)	94.4%	53
Guinea	86	2005	2244 (38)	81.7%	62
Guinea Bissau	28	2004	5354 (118)	96.3%	47
Liberia	38	2006, 2012	14022 (15)	98.1%	929
Mauritania	143	2015	11242 (54)	92.2%	221
Niger	66	2002	29905 (42)	98.6%	713
Nigeria	1107	2000-2013	58272 (774)	93.5%	75
Senegal	95	2003, 2010	23759 (76)	99.2%	312
Sierra Leone	76	2005	9581 (14)	100%	681
Togo	156	2000-2002	5393 (40)	98.2%	134

Table 2: Summary of mapping survey and inhabited places data availability by country. Completeness of the UN OCHA locations databases is assessed by calculating the proportion of mapping survey locations located within 5km of locations in the UN OCHA database. The total number of settlements and IUs present in the locations database and IU shapefile are reported here, not just for the IUs where mapping surveys occurred.

The IU shapefile from ESPEN contains 2534 total IUs for the above 19 countries. The UN OCHA database contains 272,200 settlements from which random selection was simulated. The number of IUs per country is described in Table 1, along with the number of settlements and average number of settlements within IUs by country.

Presence/Absence of sampling bias

For most countries, the difference in socio-ecological profiles between mapping survey sites and randomly selected survey sites was insufficient to conclude presence of sampling bias, with mean cross-match statistics generally well above alpha = 0.05 (Figure 1). While the null hypothesis was rejected in some bootstraps, on average, across 1000 simulated draws and comparisons, we were unable to conclude the presence of sampling bias for 15 countries. We conclude that in these countries, mapping survey sites are not systematically different from randomly selected locations according to the covariates tested.

Figure 2:

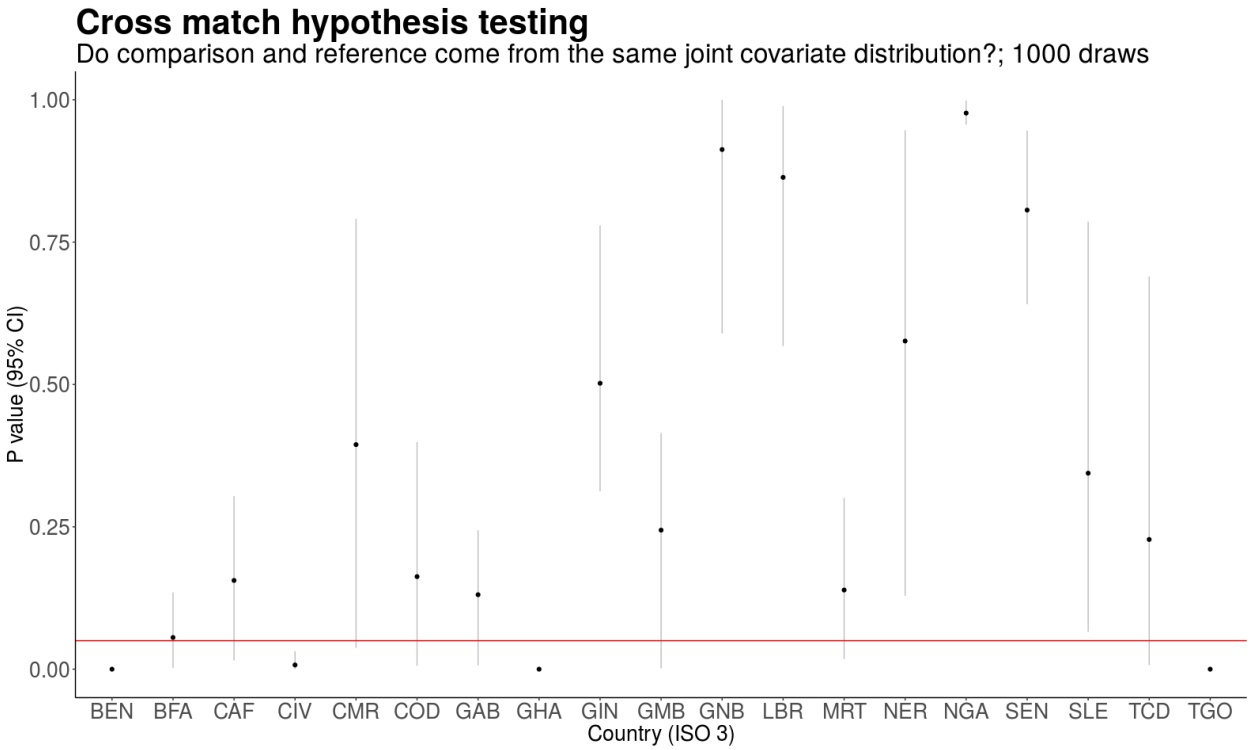

Figure 2: Distribution of cross-match statistics by country. Vertical lines indicate the 5th and 95th percentile p-value from 1000 bootstraps. Red horizontal line marks alpha = 0.05. Points indicate the mean p-value. Clear sampling bias in mapping data is concluded for countries where this value falls below alpha.

In Benin, Côte d'Ivoire, Ghana, and Togo, our results suggest systematic differences in covariate values at locations selected for mapping compared to locations selected via simulation. Joint covariate distributions at mapping survey sites were different to a degree that is statistically significant from almost all joint distributions obtained through simulated random selection. In these countries, variance in cross-match statistics across bootstraps is low (Figure 2). This meant that socio-ecological characteristics at simulated site selection were significantly different from all permutations of randomly selected locations. The probability that mapping survey locations were derived through a random process, and are representative of IUs, is likely low.

Variance in cross-match statistics does not clearly correlate with the average number of settlements per IU, but a relationship does appear to exist with the number of mapping surveys. Countries with less than 100 mapping surveys such as The Gambia, Sierra Leone, Chad, and Niger, tend to have wider confidence intervals. Meanwhile, Nigeria, which has the highest number of mapping surveys in studied countries, has low variance. Larger samples likely yielded less variability in the covariate distributions of different simulations, resulting in similar p-values when compared to the reference covariate distribution.

BRT models to predict locations included in LF mapping

Covariate partial dependence functions provide a means by which fitted covariate response in BRT models can be interpreted. Each plot illustrates a covariate's modelled relationship to the response variable after accounting for the mean effect of all other variables in the model. The response variable in this context is a location's relative probability of being sampled for mapping surveys. Covariate partial dependence functions for the three most influential covariates in each country-specific ensemble BRT model are illustrated in Figure 3.

Figure 3:

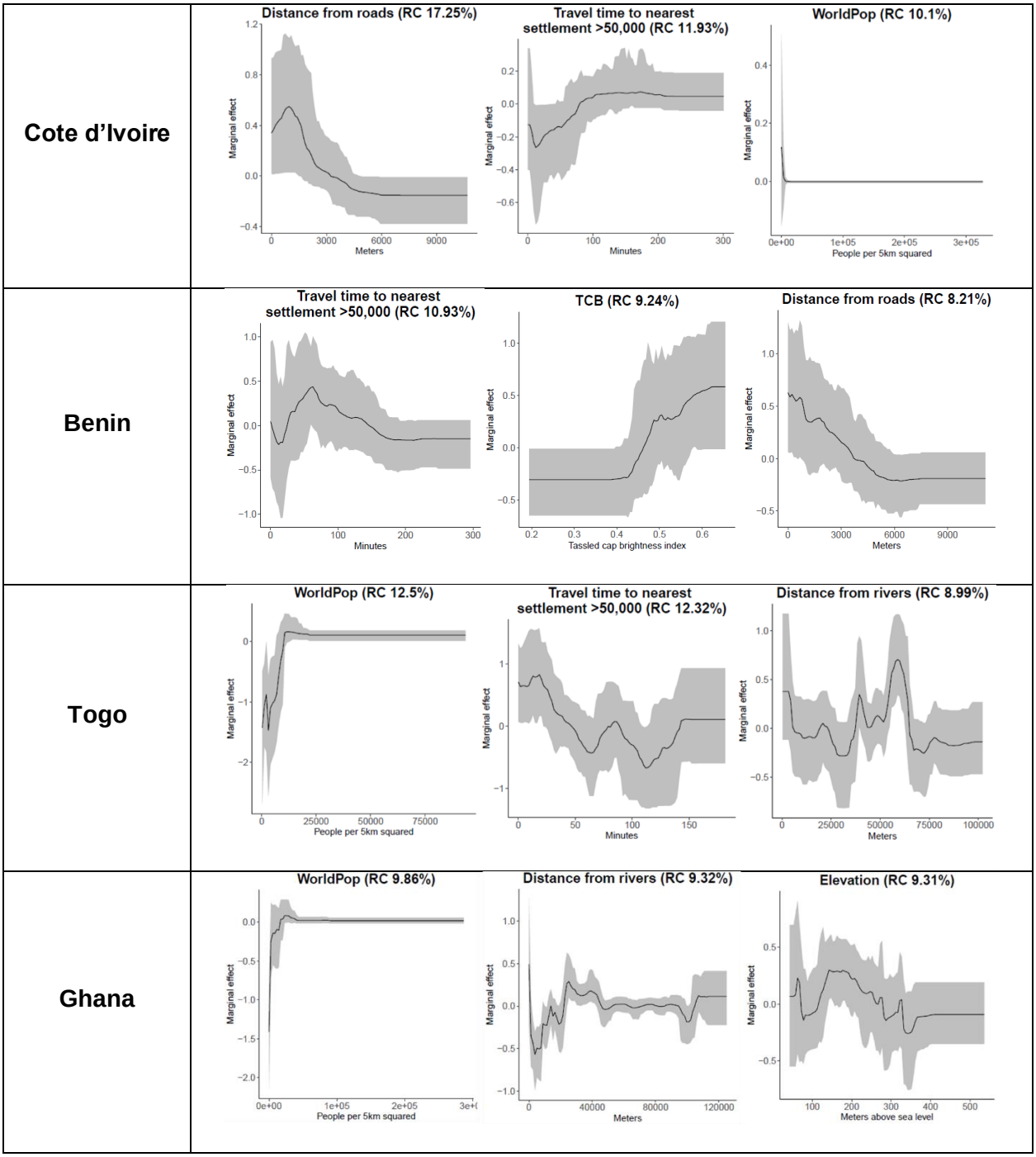

Figure 3: Covariate partial dependence functions for the three most influential covariates in each country-specific ensemble BRT model. Each line plot illustrates a covariate's modelled relationship to the response variable (relative sampling probability for mapping surveys) after the mean effect of all other variables are accounted for. RC = relative contribution; a metric of covariate influence in the ensemble model. The gray area around the mean trend illustrate variability in covariate relationships across 100 bootstrap models.

Côte d'Ivoire's covariate partial dependence functions suggest that, compared to random selection, mapping survey sites are more likely to be near major roads, less likely to near densely populated settlements, and more likely to be sparsely populated. Similarly, the Benin ensemble BRT model suggests that sampling of survey sites in that country may have been biased towards sparsely populated settlements near major roads. Unfortunately, complex interactions in the BRT models hinder clear interpretation of covariate trends from most partial dependence functions, particularly for ecological covariates. This is a recognized limitation of statistical inference using BRT modeling [17].

Using the 5km X5km gridded covariates, relative probability of being sampled for mapping was predicted at the pixel level, and the average is plotted in Figure 4. The differences in socio-ecological characteristics between sampled and simulated locations in Cote d'Ivoire are less pronounced (this is illustrated in Figure 2 by the relatively wider confidence interval across bootstrapped cross-match tests) and the resultant scale of ensemble BRT predictions is narrower than in the other three countries. Pixels with higher predicted values are more socio-ecologically similar to mapping survey locations.

Figure 4:

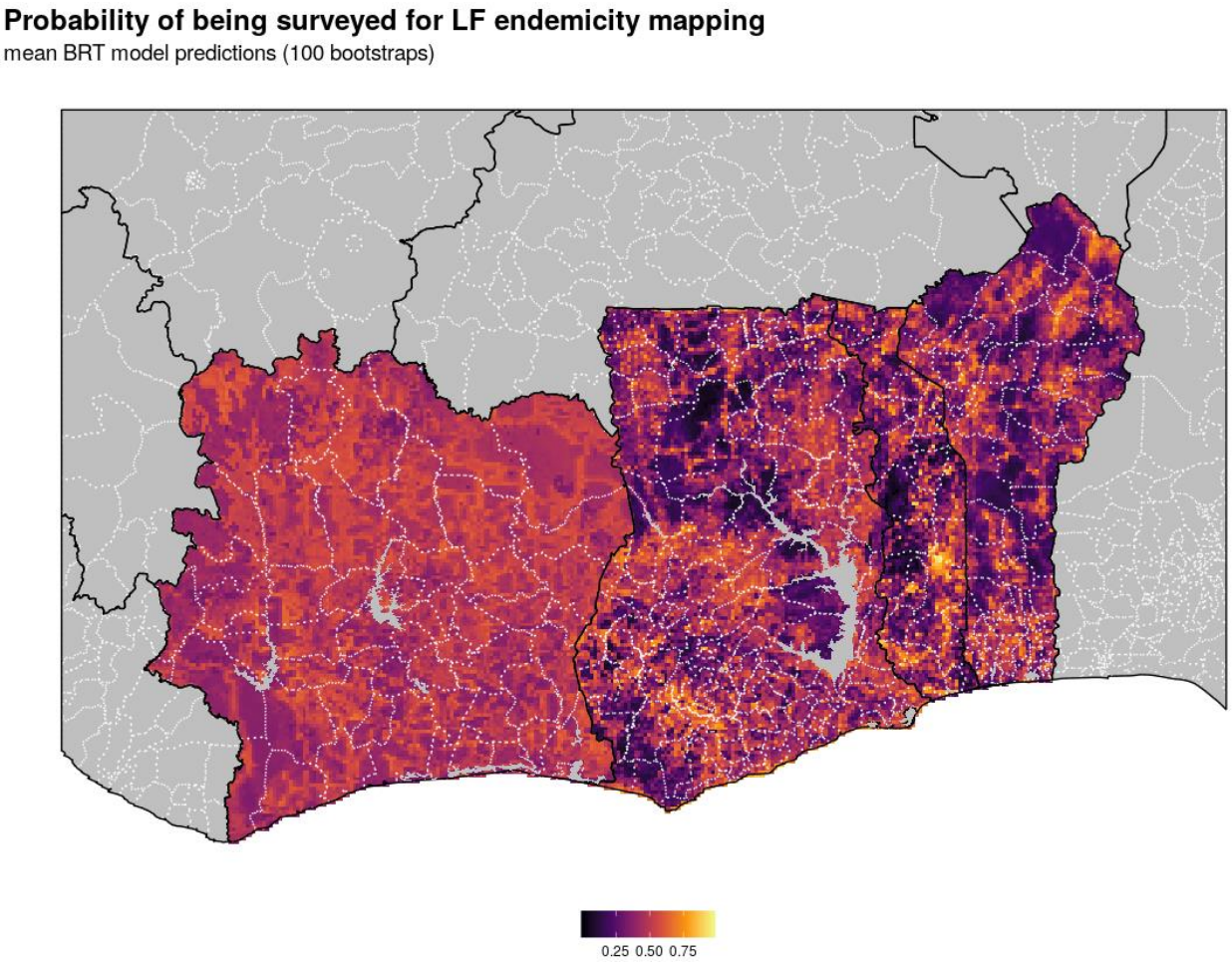

Figure 4: Each individual BRT model is used to predict out the relative probability of being sampled for mapping at every 5 km X 5km pixel and the mean prediction is plotted. IU boundaries are shown in white.

To more clearly differentiate these areas, ensemble model predictions were converted to binary classifications using optimal thresholds determined using ROC curves. Different thresholds are optimal for different country-specific ensemble model. Areas in green in Figure 5 highlight areas with socio-ecological characteristics most predictive of being sampled for mapping surveys. Mapping survey locations are also plotted. These generally overlay green areas, indicating desired predictive behavior. The thresholds used to make this map is presented in Figure 6, overlaid on histograms of model predictions at mapping survey locations (Benin = 0.66; Cote

d'Ivoire = 0.53; Ghana = 0.69; Togo = 0.57). Observations to the right of the red line, above the threshold, are true positives while observations to the left, are false negatives. Sensitivity, the true positive rate, is noticeably higher in the Togo and Benin classifications.

Figure 5:

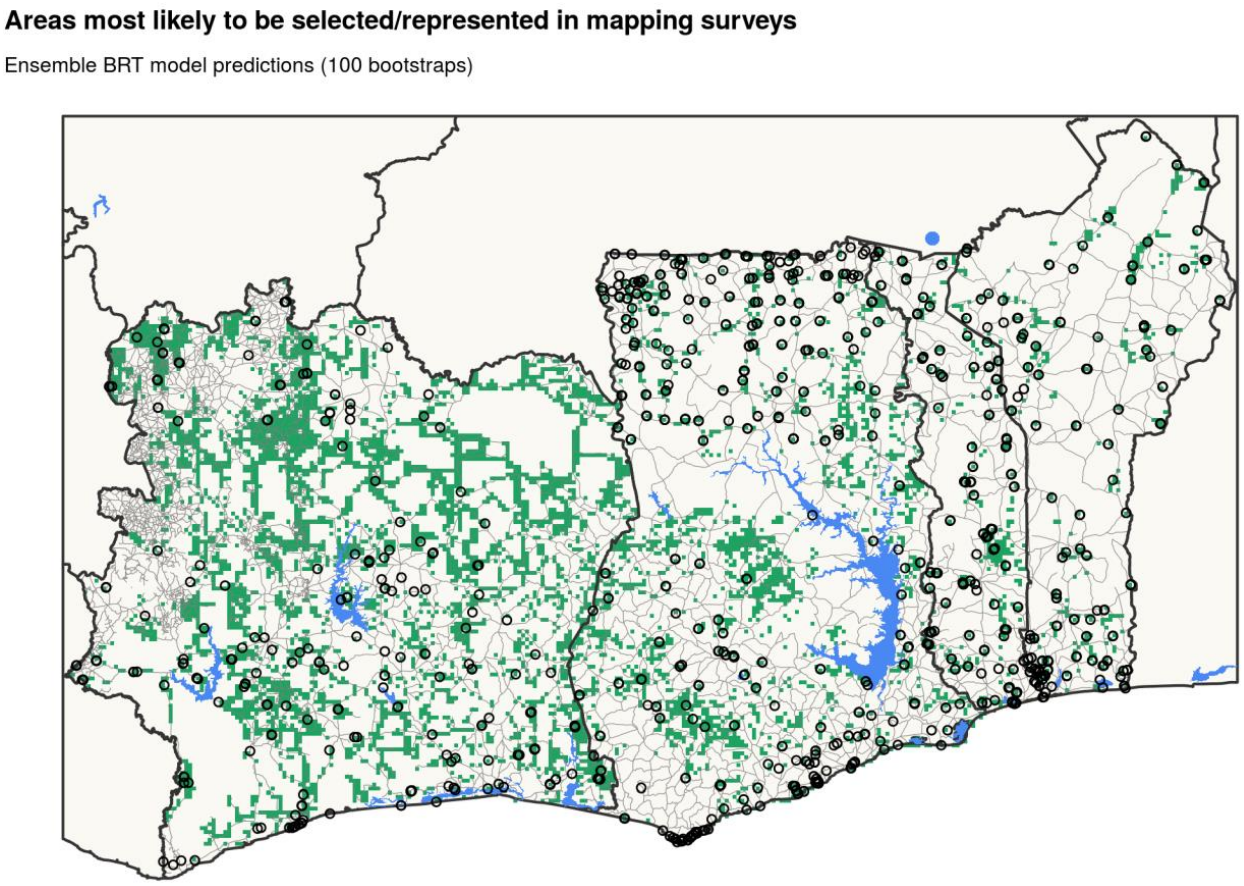

Figure 5: Green areas highlight areas classified as most likely to have been sampled/represented during the country LF endemicity mapping process. This is based on 5km X5km predictions from country ensemble BRT models of location-specific relative probability of being sampled as a function of socio-ecological characteristics. Additional features include locations of roads (light gray), major waterways (blue), and mapping survey sites (black circles).

The modelled sampling bias towards major roads in Cote d'Ivoire mentioned earlier is visible in Figure 5, with 5x5km pixels most characteristic of sampling bias in Côte d'Ivoire mapping coinciding with the location of roads.

Figure 6:

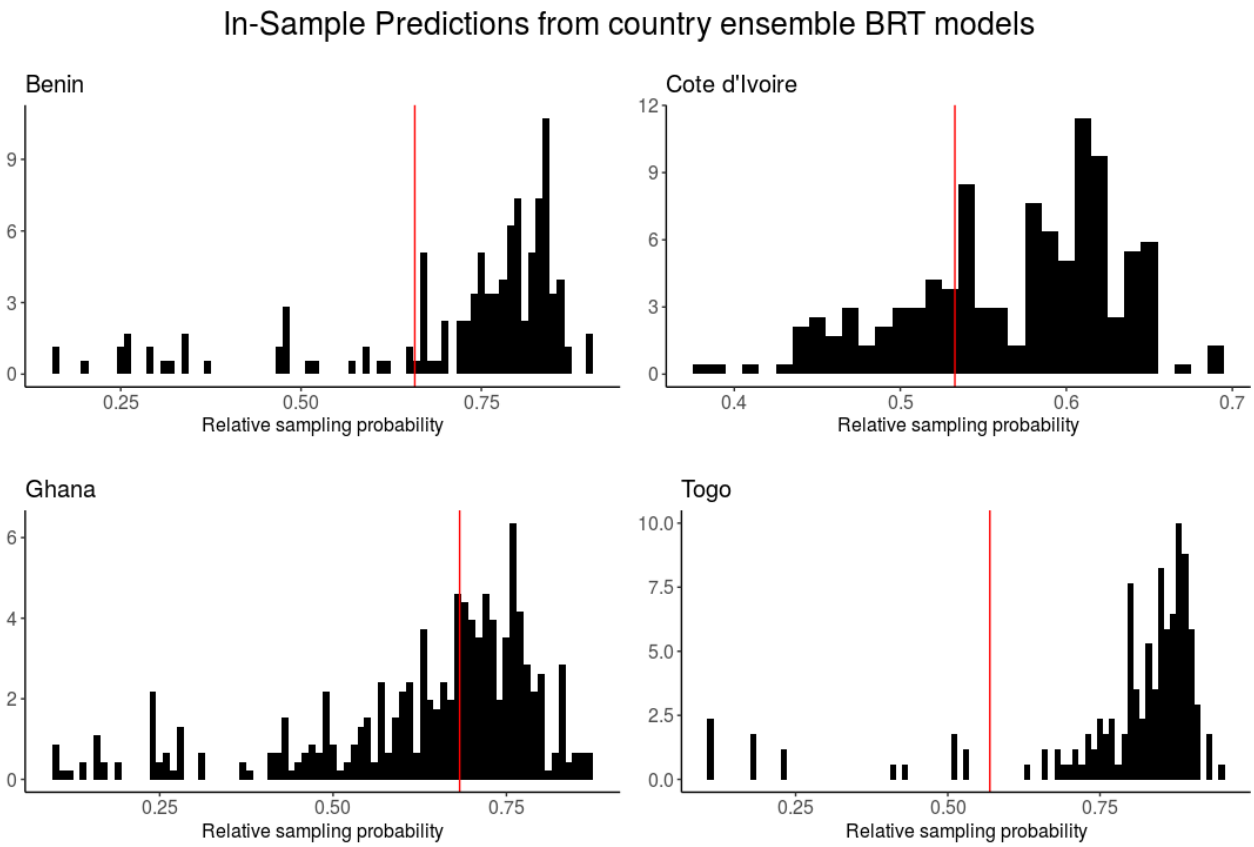

Figure 6: Predicted relative sampling probability at mapping survey locations from country ensemble BRT models. Red line indicates the threshold used to convert predictions to a binary indicator. This threshold was selected using receiver operating characteristics (ROC) curves to maximize the true positive rate and minimize the false positive rate.

Sensitivity Analyses

Iterative reruns of the bootstrapped cross-match test with 10 randomly chosen covariates yielded comparable p-values to when running tests using the full covariate list. All reruns for Benin, Cote d'Ivoire, Ghana, and Togo returned a mean p-value across 1000 bootstraps below 0.05. This indicates that the differences between the socio-ecological characteristics at surveyed and simulated locations are robust and are not contingent on the specific covariate set selected for the comparison.

As hypothesized, the likelihood of BRT convergence failure increases proportionally with the absence of detected geospatial bias - such that countries with higher cross-match statistic have higher rates of failed convergence. This shows agreement with the results of the cross-match statistic.

Discussion

In this study, we tested the national representativeness of locations selected for LF endemicity mapping from 19 West and Central African countries across 17 socio-ecological dimensions by comparing the joint covariate distribution at surveyed locations with distributions from locations sampled at random 1000 times. For 15 countries, this test did not provide sufficient evidence of sampling bias. The socio-ecological profile of mapping survey locations in these countries is not clearly distinguishable from profiles generated through simulated random selection. We found statistically significant difference indicative of potential sampling bias in mapping survey site selection in four West African countries; Benin, Togo, Ghana, and Côte d'Ivoire. BRT modeling of sampling bias in these countries' mapping datasets provided limited insights on socio-ecological correlates of selection probability. Distance from major roads and population density were two of the most influential covariates across country ensemble models. In Côte d'Ivoire and Benin, compared to random selection, mapping survey sites are more likely to be sparsely populated and near major roads. Unfortunately, complex interactions in the BRT models hinder clear interpretation of covariate trends from most partial dependence functions, particularly for ecological covariates.

Sampling bias in mapping data may lead to skewed or spurious results in secondary analyses if sampling probability is associated with LF prevalence. For example, the ensemble BRT model for Cote d'Ivoire suggests that sampling bias may pose an issue if a researcher intended to analyze the relationship between population density and LF prevalence using Cote d'Ivoire

mapping data because sampling probability is associated with population density. Without adjustment, the available data would be insufficient to determine if a detected association is true or induced/skewed by sampling bias.

Socio-ecological covariates used for this analysis were chosen due to perceived or measured association with LF prevalence. Since our results suggest sampling probability is associated with some of these factors in Benin, Cote d'Ivoire, Togo, and Ghana, researchers using mapping data from these countries for secondary analyses should be vigilant of the effects of sampling bias.

Limitations

Our analytical framework may be limited in detecting sampling bias due to a variety of factors. Covariates used in this analysis may be insufficient to capture all factors relevant to sampling bias in LF endemicity mapping. Most importantly, our analysis is dependent on the quality of underlying data. In order to do this analysis, we assumed the ESPEN mapping survey dataset and the UN OCHA settlements database to be complete and its geo-positioning accurate. In reality, data quality of both data sources likely varies across countries. ESPEN, launched in 2016 by WHO AFRO, is dependent on national NTD programs for accurate and complete data. Geospatial coordinates in some country mapping datasets may have been georeferenced post-hoc rather than collected in the field. Inaccurate geo-positioning is difficult to assess and may bias our analysis if inaccuracy is non-random. For example, if the GPS location of the nearest large settlement is used in place of the true sampled location, our analysis would erroneously detect sampling bias, with sampling probability correlated with population density and proximity to large settlements and roads.

Similarly, if a non-random pattern of missingness exist in the UN OCHA settlements database, such missingness would percolate through our analysis and may lead to erroneous results. For

example, if less assessable settlements are under-represented in the database, our simulation of random sampling would consistently generate a socio-ecological profile biased away from characteristics of less assessable settlements. We attempted to quantify country-specific completeness in the database with a simple metric by calculating the proportion of mapping survey locations within 5km of a UN OCHA settlement, and found these to be high, ranging from 82-100%.

Conclusions

In conclusion, this analysis is the first quantitative assessment of sampling bias in LF mapping surveys across multiple countries. Systematic differences in the socio-ecological profile of mapping locations from random selection suggests sampling bias in the mapping of four West African countries, Benin, Togo, Ghana, and Côte d'Ivoire. Use of mapping data from these contexts for secondary analyses should consider the implications of this bias.

If unaccounted for, sampling bias may lead to (1) mischaracterization of socio-ecological predictors of LF transmission risk in models of LF prevalence or environmental suitability and (2) diminished generalizability. Areas with under-sampled socio-ecological characteristics, pixels with low predicted relative sampling probability in Figure 4, would likely be disproportionately affected by sampling bias in secondary analyses. True covariate trends in these areas may significantly differ from those found using a biased dataset, leading to inaccurate predictions of LF transmission risk. From this analysis alone, it is not possible to conclude whether this inaccuracy tend to result in over or under prediction, or if the bias could lead to in-optimal programmatic decision-making. Interpretation of secondary analyses using potentially biased data sources should ideally involve local partners with the context-specific knowledge necessary to vet the face validity of results.

New data collection may help clarify the representativeness of original mapping surveys. However, it is unclear whether such efforts are appropriate or necessary in Benin, Cote d'Ivoire, Ghana, or Togo. These countries were amongst the first to initiate MDA in Africa, and, in 2017, Togo became the first African country to achieve elimination. While there is precedent for further data collection to clarify mapping data, this has so far only been done in pre-intervention settings. In parts of Ethiopia and Tanzania, uncertainty in the interpretation of mapping survey results led to the development of a confirmatory mapping tool [3, 20]. This geographic representative school-based large cluster survey methodology clarified the status of LF transmission in IUs with previously uncertain endemicity.

Researchers in Ghana have expressed concern about some districts previously classified as non-endemic that have since reported new LF cases [38]. This, along with inconsistencies in diagnostics used during mapping [38] and the possibility of selection bias described in this survey, suggest further data collection may be warranted.

References:

1. WHO | Global Programme to Eliminate Lymphatic Filariasis [Internet]. WHO. [cited 2019 Mar 11]. Available from: http://www.who.int/lymphatic_filariasis/elimination-programme/en/
2. WHO/ Department of Communicable Disease Prevention, Control and Eradication, Rio, Francecso. Operational guidelines for rapid mapping of Bancroftian filariasis in Africa (Revised during an inter-country workshop held in Ouagadougou, 8−12 March 2000). 2000 Mar.
3. Gass KM, Sime H, Mwingira UJ, Nshala A, Chikawe M, Pelletreau S, et al. The rationale and cost-effectiveness of a confirmatory mapping tool for lymphatic filariasis: Examples from Ethiopia and Tanzania. PLOS Neglected Tropical Diseases. 2017 Oct 4;11(10):e0005944.
4. Lymphatic filariasis | ESPEN [Internet]. [cited 2019 Mar 11]. Available from: http://espen.afro.who.int/diseases/lymphatic-filariasis
5. Stanton MC, Molyneux DH, Kyelem D, Bougma RW, Koudou BG, Kelly-Hope LA. Baseline drivers of lymphatic filariasis in Burkina Faso. Geospat Health. 2013 Nov;8(1):159–73.
6. Eneanya OA, Cano J, Dorigatti I, Anagbogu I, Okoronkwo C, Garske T, et al. Environmental suitability for lymphatic filariasis in Nigeria. Parasit Vectors. 2018 Sep 17;11(1):513.
7. Gyapong JO, Kyelem D, Kleinschmidt I, Agbo K, Ahouandogbo F, Gaba J, et al. The use of spatial analysis in mapping the distribution of bancroftian filariasis in four West African countries. Ann Trop Med Parasitol. 2002 Oct;96(7):695–705.
8. Moraga P, Cano J, Baggaley RF, Gyapong JO, Njenga SM, Nikolay B, et al. Modelling the distribution and transmission intensity of lymphatic filariasis in sub-Saharan Africa prior to scaling up interventions: integrated use of geostatistical and mathematical modelling. Parasit Vectors. 2015 Oct 24;8:560.
9. Eneanya OA, Fronterre C, Anagbogu I, Okoronkwo C, Garske T, Cano J, et al. Mapping the baseline prevalence of lymphatic filariasis across Nigeria. Parasites and Vectors [Internet]. 2019 [cited 2019 Aug 14]; Available from: https://ora.ox.ac.uk/objects/uuid:bfbb307d-145b-49e2-9e86-449adf583ea1
10. Michael E, Singh BK, Mayala BK, Smith ME, Hampton S, Nabrzyski J. Continental-scale, data-driven predictive assessment of eliminating the vector-borne disease, lymphatic filariasis, in sub-Saharan Africa by 2020. BMC Med. 2017 27;15(1):176.
11. Kastner RJ, Stone CM, Steinmann P, Tanner M, Tediosi F. What Is Needed to Eradicate Lymphatic Filariasis? A Model-Based Assessment on the Impact of Scaling Up Mass Drug Administration Programs. PLoS Negl Trop Dis [Internet]. 2015 Oct 9;9(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4599939/
12. West and Central Africa - Administrative boundaries levels 0 - 2 / settlements - Humanitarian Data Exchange [Internet]. [cited 2019 Mar 11]. Available from: https://data.humdata.org/dataset/west-and-central-africa-administrative-boundaries-levels
13. Rosenbaum P. An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004 Jun;515–30.

14. Kadmon R, Farber O, Danin A. Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models. Ecological Applications. 2004;14(2):401–13.
15. Reddy S, Dávalos LM. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography. 2003;30(11):1719–27.
16. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nature Microbiology. 2019 Jun 10;1.
17. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology. 2008;77(4):802–13.
18. Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006;29(2):129–51.
19. Leathwick JR, Elith J, Francis M. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. ResearchGate [Internet]. [cited 2019 Aug 15]
20. Mwingira U, Sime H, Pelletreau S, Deming M, Rebollo MP, Gass K. Shrinking the lymphatic filariasis map: A new tool to assess active transmission of LF during mapping. American Society of Tropical Medicine & Hygiene (ASTMH) 64th Annual Meeting. 2015 Oct 25;Abstr 1063.
21. Weiss, D. J. et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 533, 333-336 (2018).
22. Harris, I., Jones, P. d., Osborn, T. j. & Lister, D. h. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 dataset. Int. J. Climatol. 34, 623–642
23. University of East Anglia. Climatic Research Unit TS v. 3.24 dataset. Available at: https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_3.24.0 1/. (Accessed: 24th July 2017). (2014)
24. Andreadis KM, Schumann GJ-P, Pavelsky T. A simple global river bankfull width and depth database. Water Resources Research. 2013;49(10):7164–8.
25. Young, A. H., K. R. Knapp, A. Inamdar, W. B. Rossow, and W. Hankins, 2017: "The International Satellite Cloud Climatology Project, H-Series Climate Data Record Product", Earth System Science Data, in preparation.
26. Huete, A., Justice, C. & van Leeuwen, W. MODIS vegetation index (MOD 13) algorithm theoretical basis document. (1999).
27. USGS & NASA. Vegetation indices 16-Day L3 global 500m MOD13A1 dataset. Available at: https://lpdaac.usgs.gov/dataset_discovery/modis/m odis_products_table/mod13a1. (Accessed: 25th July 2017)
28. Weiss, D. J. et al. An effective approach for gapfilling continental scale remotely sensed timeseries. Isprs J. Photogramm. Remote Sens. 98, 106–118 (2014).
29. C. Schaaf, Z. Wang. (2015). MCD43A1 MODIS/Terra+Aqua BRDF/Albedo Model Parameters Daily L3 Global - 500m V006. NASA EOSDIS Land Processes DAAC. http://doi.org/10.5067/MODIS/MCD43A1.006
30. Lloyd, C. T., Sorichetta, A. & Tatem, A. J. High resolution global gridded data for use in population studies. Sci. Data 4, sdata20171 (2017).
31. World Pop. Get data. Available at: http://www.worldpop.org.uk/data/get_data/. (Accessed: 25th July 2017)
32. Beck, H.E., A.I.J.M. van Dijk, V. Levizzani, J. Schellekens, D.G. Miralles, B. Martens, A. de Roo: MSWEP: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging

gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21(1), 589-615, 2017.
33. Center for International Earth Science Information Network - CIESIN - Columbia University, and Information Technology Outreach Services - ITOS - University of Georgia. 2013. Global Roads Open Access Data Set, Version 1 (gROADSv1). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4VD6WCT.
34. FAO. GAEZ - Global Agro-Ecological Zones data portal. Available at: http://www.fao.org/nr/gaez/about-data-portal/en/. (Accessed: 25th July 2017)
35. S. Siebert, P. Doll, S. Feick, J. Hoogeveen, and K. Frenken. 2007. Global Map of Irrigation Areas, Version 4.0.1, Johann Wolfgang Goethe University, Frankfurt am Main, Germany / Food and Agriculture Organization of the United Nations, Rome, Italy http://www.fao.org/nr/water/aquastat/irrigationmap/index10.stm
36. NASA & NOAA. Advanced Very High Resolution Radiometer (AVHRR) Normalized Difference Vegetation Index (NDVI) dataset. Available at: https://nex.nasa.gov/nex/projects/1349/. (Accessed: 25th July 2017)
37. Pesaresi, M. et al. Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. (Publications Office of the European Union, 2016).
38. Biritwum N-K, Souza DK de, Marfo B, Odoom S, Alomatu B, Asiedu O, et al. Fifteen years of programme implementation for the elimination of Lymphatic Filariasis in Ghana: Impact of MDA on immunoparasitological indicators. PLOS Neglected Tropical Diseases. 2017 Mar 23;11(3):e0005280.
39. Robert J. Hijmans, Steven Phillips, John Leathwick and Jane Elith (2017). dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo