Eneanya *et al. Parasites & Vectors* (2018) 11:513
https://doi.org/10.1186/s13071-018-3097-9

Parasites & Vectors

RESEARCH Open Access

Environmental suitability for lymphatic filariasis in Nigeria

Obiora A. Eneanya[1*], Jorge Cano[2], Ilaria Dorigatti[1], Ifeoma Anagbogu[3], Chukwu Okoronkwo[3], Tini Garske[1] and Christl A. Donnelly[1,4]

Abstract

Background: Lymphatic filariasis (LF) is a mosquito-borne parasitic disease and a major cause of disability worldwide. It is one of the neglected tropical diseases identified by the World Health Organization for elimination as a public health problem by 2020. Maps displaying disease distribution are helpful tools to identify high-risk areas and target scarce control resources.

Methods: We used pre-intervention site-level occurrence data from 1192 survey sites collected during extensive mapping surveys by the Nigeria Ministry of Health. Using an ensemble of machine learning modelling algorithms (generalised boosted models and random forest), we mapped the ecological niche of LF at a spatial resolution of 1 km^2. By overlaying gridded estimates of population density, we estimated the human population living in LF risk areas on a 100 × 100 m scale.

Results: Our maps demonstrate that there is a heterogeneous distribution of LF risk areas across Nigeria, with large portions of northern Nigeria having more environmentally suitable conditions for the occurrence of LF. Here we estimated that approximately 110 million individuals live in areas at risk of LF transmission.

Conclusions: Machine learning and ensemble modelling are powerful tools to map disease risk and are known to yield more accurate predictive models with less uncertainty than single models. The resulting map provides a geographical framework to target control efforts and assess its potential impacts.

Keywords: Lymphatic filariasis, Ensemble modelling, Machine learning, Generalised boosted model (GBM), Random forest (RF)

Background

Lymphatic filariasis (LF) is a mosquito-borne disease endemic in tropical regions and caused by the parasitic nematode *Wuchereria bancrofti* in Africa, and by *Brugia malayi* and *B. timori* in Southeast Asia [1]. These parasites are transmitted by various species of mosquitoes, with *Anopheles* spp. being major vectors in Africa [1, 2]. Other mosquito species of the genera *Culex* and *Mansonia* also contribute to transmission to some extent, particularly in urban and peri-urban settings [3, 4]. The majority of infected individuals are asymptomatic, but infections can lead to lymphedema, hydrocele and swellings of the breasts in women [5]. An independent International Task Force for Disease Eradication included LF as one of the nine diseases targeted for elimination [6], and in 1997 the World Health Assembly adopted Resolution WHA50.29 embarking on a global campaign to eliminate LF. Elimination of LF as a public health problem is deemed feasible for a number of reasons: (i) mosquitoes are very inefficient transmitters of filarial parasites [7]; (ii) the small number of animal reservoirs are restricted to particular foci for *B. malayi*, and there are no animal reservoirs for *W. bancrofti* [8]; and (iii) the availability of improved diagnostic tools and the existence of practical interventions for interruption of transmission [9–11].

An understanding of the geographical distribution of LF is required to underpin national elimination programmes. This enables more effective targeting of control efforts on highly endemic areas. Early maps of disease distribution have mostly relied on field surveys at national or

* Correspondence: o.eneanya13@imperial.ac.uk
[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK
Full list of author information is available at the end of the article

sub-national levels [12–15], often with entire administrative units classified based on disease prevalence with no account for within-unit heterogeneity [16]. This might be useful for roughly estimating disease burden [17, 18]; however, such estimates fail to accurately represent disease burden and highly endemic foci may be misclassified [18]. To account for within-region heterogeneity, maps have been created by applying geostatistical modelling on point prevalence data, in combination with potential disease drivers (i.e. climatic, environmental and demographic factors), due to their impacts on mosquito populations and parasite biology [16, 19–24].

In 2003, the Nigerian Lymphatic Filariasis Elimination Programme (NLFEP) commenced LF mapping on a national scale, and to date, 761 out of 774 Local Government Areas (LGAs) have been mapped using immunochromatographic card tests (ICT) [15]. Of these, 574 LGAs are classed as endemic and targeted for mass drug administration (MDA), and 187 LGAs non-endemic for LF [15]. In total, an estimated 128 million people in Nigeria are thought to require preventive chemotherapy, and as of 2016, 54% of this population had been treated [25]. After more than five rounds of MDA in Plateau and Nassarawa states, Transmission Assessment Survey 1 (TAS-1) showed evidence of interruption of LF transmission in these areas [26]. However, for the vast areas of the country in which LF is present, understanding disease distribution on a finer scale is key for more focussed targeting of control measures.

In this work we aim to (i) describe and map the ecological niche of LF in Nigeria and (ii) estimate the human population living in areas that are environmentally suitable for disease transmission. Here we fitted seven different model classes to the same selection of training and evaluation data points, and projected the final ecological niche map using an ensemble of the two best performing models.

Methods

LF occurrence data

Data used for ecological niche modelling (ENM) were pre-intervention site-level data collected during mapping surveys conducted by the Nigeria Ministry of Health from 2000–2013. The sampling geographical level for LF mapping is the implementation unit (IU), which corresponds to a Local Government Area (LGA), the second level administrative unit in Nigeria. In total, we had 1192 data points covering all 36 states, and the Federal Capital Territory in Nigeria. A uniform survey methodology was applied to all survey locations and study participants were tested for the presence of filarial antigenemia using a point-care rapid test (immunochromatographic card test), and following the mapping protocol of the World Health Organization (WHO) Operational Guidelines for Mapping of Bancroftian Filariasis in Africa [27]. Briefly, within each IU (LGA), at least one sample village is randomly selected for survey. Selected villages must be located at least 50 km apart from each other. In each selected village, 50–100 adults (seeking an equal number of males and females when possible, and > 15 years of age) are tested. If 20% or more of the first 50 individuals tested result positive, testing is stopped, and the entire IU is recorded as positive for LF. Otherwise, testing is continued until 100 adults have been examined. In the end, IUs with at least one sampled village yielding a prevalence equal to or more than 1% are considered to be endemic for LF and subsequently targeted with MDA.

For this analysis, occurrence (coded 1) was considered when at least one LF case was recorded during mapping surveys and absence (coded 0) when none of the individuals tested resulted positive to the LF rapid test [28]. In total we had 932 ‘presence’ and 260 ‘absence’ records for this modelling exercise. Figure 1 shows the distribution of survey points and presence-absence records used in this work.

Climatic and environmental data

A suite of environmental variables was considered to describe the ecological niche of LF. Continuous gridded maps of climate, topography, vegetation and land use for Africa were obtained from different sources (Table 1). Climate variables related to precipitation and temperature were downloaded from the WorldClim database [29], which provides a set of global climate layers obtained by interpolation of the data for the period of 1950–2000 collected in weather stations distributed across the world. From the Consortium of Spatial Information (CGIAR-CSI) we obtained raster datasets of potential evapo-transpiration (PET), elevation and aridity index at 1 km^2 resolution [30]. PET is a measure of the ability of the atmosphere to remove water through evapo-transpiration. Our elevation layer resulted from processing and resampling the gridded digital elevation models (DEM) derived from the 30-arcsecond DEM produced by the Shuttle Radar Topography Mission (SRTM) [31]. The elevation layer was used to generate two topography-related datasets: slope of terrain and flow accumulation.

To produce the flow accumulation layer, we initially created a flow direction layer in which the direction of flow was determined by the direction of the steepest descent from each cell in the elevation dataset. This was calculated as follows: change in elevation value/distance × 100. Flow accumulation was then calculated as the accumulated weight of all cells flowing into each downslope cell in the flow direction layer.

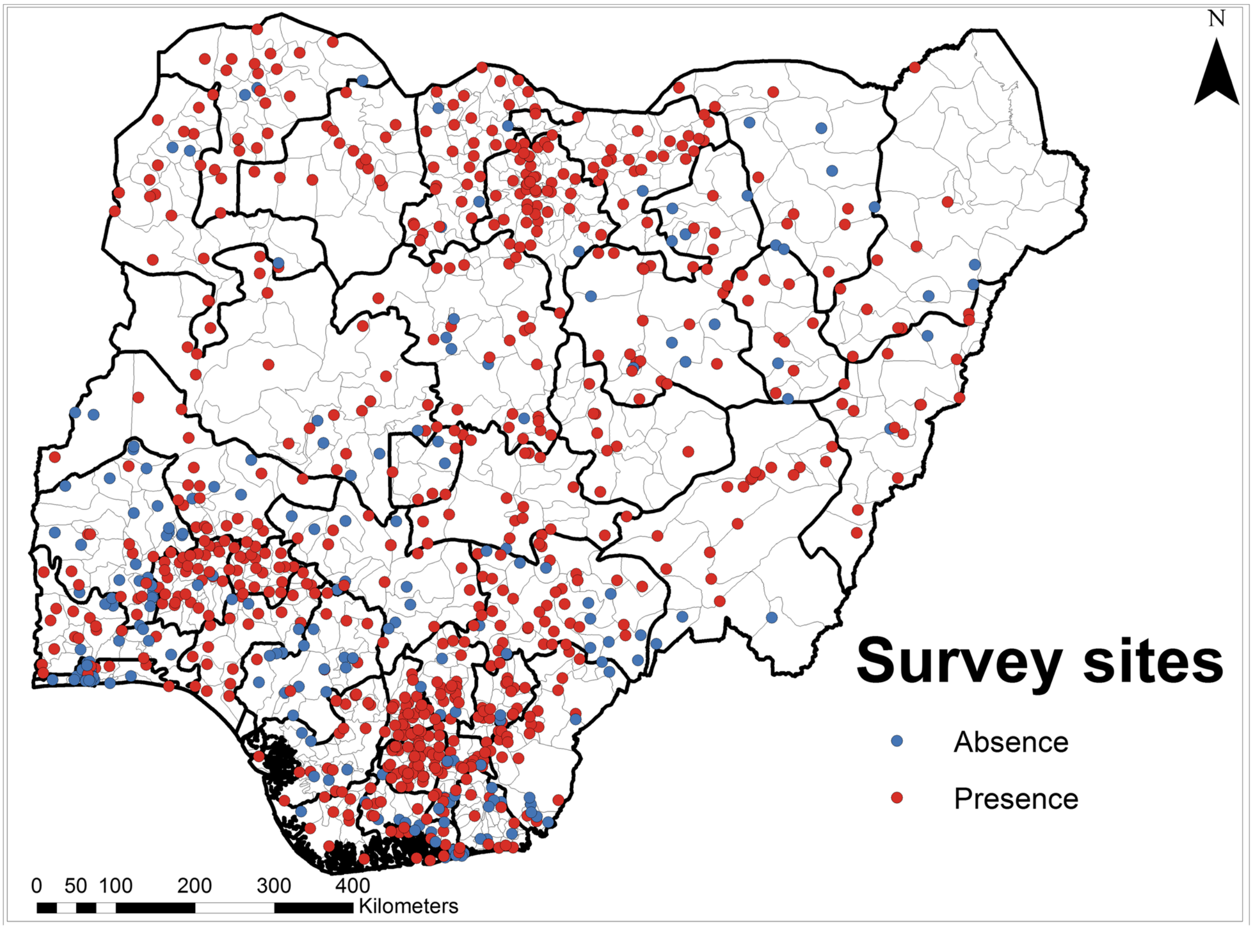

Fig. 1 Location of study sites in Nigeria. Red points show sites with at least one LF case and blue points show sites with no LF case

In addition, we calculated topographic wetness index (TWI) by applying the following algorithm

$$TWI = \ln(a/tan\beta),$$

where a is the upslope contributing area per unit contour length [or specific catchment area (SCA)], which can be approached by using the flow accumulation, and β is the local slope gradient for reflecting the local drainage potential [32].

We also produced continuous surfaces of straight-line distance (Euclidean distance) in kilometres to the nearest water body and permanent rivers. These data were derived from the Global Database of Lakes, Reservoirs and Wetlands [33] and Digital Global Chart [34], respectively. Raster datasets of averaged enhanced vegetation index (EVI) and land surface temperature (LST) for the period 2000–2015 were obtained from the African Soil Information System (AfSIS) project [35]. Here, EVI and LST were generated from remotely sensed data collected from the Moderate Resolution Imaging Spectoradiometer (MODIS) platform. MODIS collects earth data from the same place every 16 days at 250 m spatial resolution. Land cover types (according to the United Nations land cover classification system) were extracted from the GlobCover project at the European Space Agency [36]. Here, maps are derived by an automatic and regionally-tuned classification of a 300 m medium resolution imaging spectrometer (MERIS) sensor on board the ENVISAT satellite mission. Soil data (sand, silt and clay fractions, and soil pH) were downloaded from the International Soil Reference and Information Centre (ISRIC) project [37] at a spatial resolution of 250 m.

Finally, night-light (NL) emissivity for 2006 captured by the Operational Linescan System instrument on board a satellite of the Defence Meteorological Satellite Programme was used as a proxy measure of poverty across Nigeria [38]. This instrument measures visible and infrared radiation emitted at night-time, resulting in remote imagery of lights on the ground. This information has been correlated with gross domestic product in developed countries [39, 40] and, although far from being precise, could provide an indirect measure of poverty

Table 1 Environmental variables used in the ENM and their sources

Variables	Source
Annual cumulative precipitation	WorldClim [29]
Maximum temperature	
Mean temperature	
Minimum temperature	
Mean temperature of the coldest quarter	
Mean temperature of the warmest quarter	
Precipitation of the driest quarter	
Precipitation of the wettest quarter	
Potential evapo-transpiration (PET)	CGIAR-CSI [30]
Aridity index	
Elevation	SRTM [31]
Slope	Derived from elevation
Flow accumulation	Derived from slope
Distance to permanent rivers	Digital Global Chart [34]
Distance to nearest water bodies	Global Database of Lakes, Reservoirs and Wetlands [33]
Land surface temperature (LST)	AfSIS [35]
Enhanced vegetation index (EVI)	
Sand, silt, clay fractions	ISRIC [37]
Soil pH	
Major land cover (forest, agriculture, shrubland-grassland)	Arino et al [36]
Wetness index	Derived from slope and flow accumulation
Distance to stable lights 2006	Elvidge et al. [38]

in developing countries [41]. NL emissivity is provided as gridded maps of 1 km^2 resolution, and values that go from 0 (undetectable NL emissivity) to 60 (maximum NL emissivity). Alternatively, we estimated the Euclidean distance to stable night-lights, considered as NL values > 0.

Input grids were resampled to a common spatial resolution of 1 km^2 using the nearest-neighbour approach [42], clipped to match the geographical extent of a map of mainland Nigeria, and aligned to it. Raster manipulation and processing were undertaken using *raster* package in R (v.3.3.2) [43]. All environmental covariates considered in our models are known to be biologically plausible for LF occurrence [16, 23].

Model implementation

Selection of covariates

Covariate data were extracted corresponding to each of the presence-absence data points. In this work we explored the existence of multicollinearity using the variance inflation factor (VIF). Multicollinearity often arises in statistical models, and occurs when two or more covariates are not statistically independent leading to unstable estimates of variances of regression coefficients [44]. The VIF represents the amount of variability of a covariate which is explained by other covariates. For instance, the VIF for the *i*th covariate can be calculated as: $VIF_i = 1/(1 - R^2{}_i)$, where $R^2{}_i$ is the coefficient of determination obtained by fitting a linear regression model for the *i*th independent covariate. The VIF of the suite of environmental covariates tested here was calculated and correlated variables were excluded in a stepwise procedure at a generally accepted threshold value of 10 [44]. Of the 24 covariates initially tested for multicollinearity, seven (average precipitation, mean temperature, average maximum temperature, average minimum temperature, aridity index, PET and sand soil type) were excluded from further analysis. All remaining covariates were considered to be independent and were included in the analysis. The multicollinearity test was implemented using the *usdm* package in R (v.3.3.2) [43].

The relative importance of the covariates to our presence-absence dataset was identified using the boosted regression trees (BRT) machine-learning algorithm. BRT is a combination of two algorithms, regression trees and boosting. This produces an additive regression model in which simple trees are fitted in a forward, stepwise fashion. This method has been widely used in disease prediction and considered a powerful tool for ecological studies [24, 45, 46]. Relative importance is defined as the frequency of selection of covariates for splitting, weighted by the squared improvements to the model, and averaged over all trees [45]. Higher relative importance scores, which are computed as percentages (and scaled to a maximum sum of 100%), indicate greater contribution to the model. Variables that showed no substantial contribution to the model (we set this at a threshold of 10%) [47] were excluded in fitting the final ensemble of models. Variables dropped at this stage were soil pH, night-light emissivity, EVI, distance to the nearest water body, distance to rivers, flow accumulation, mean temperature of the coldest quarter, mean temperature of the warmest quarter, and clay and silt soil fraction. The remaining predictors: precipitation in the driest quarter, precipitation in the wettest quarter, wetness index, land surface temperature, elevation, distance to stable lights and terrain slope, were included in the final analysis for ensemble modelling.

Building the ensemble model

We fitted our data using seven model algorithms, namely generalised linear models (GLM), surface range envelopes (SRE), multivariate additive regression splines (MARS), artificial neural networks (ANN), BRT, also known as GBM (generalised boosted regression modelling), random forest (RF) and maximum entropy ecological niche models (MaxEnt). These algorithms are included within Biodiversity Modelling (BIOMOD) [48], a computational framework intended for modelling species distribution. BIOMOD was used to build the ecological niche model, and implemented with the package *biomod2* in R (v.3.3.2) [43]. The MaxEnt algorithm is not in-built in BIOMOD; however, there is provision to include this as an add-on. The software was developed by S. Phillips and colleagues and is freely available from https://biodiversityinformatics.amnh.org/open_source/maxent/ (v.3.4.1).

Ideally, model accuracy should be evaluated with data that are independent of the training data. As we did not have an independent dataset, the original data was partitioned into two, with a random sample of 30% of the original data retained for testing/evaluation and was considered 'quasi-independent', while the remaining 70% was used to train/calibrate the model. To evaluate the validity of model performance and accuracy on the quasi-independent data, BIOMOD offers an alternative ability to perform internal cross-validation whereby a set number of data splitting runs are computed. In each model run, the model is fitted to one part of the dataset and tested on the other part. This internal cross-validation does not provide a measure of predictive performance *per se*, but provides a measure of internal consistency of models [49]. We performed an iteration of 100 model runs for each of the seven algorithms, and the evaluation values of each run was stored and then averaged, to make the final result more robust. Model evaluation was performed based on the area under the receiver operating characteristic (ROC) curve and the Hanssen-Kuipers discriminant (also known as true skill statistic, TSS). TSS compares the number of correct predictions, minus predictions attributable to random guessing [50], taking into account both sensitivity and specificity. Its value ranges from -1 to +1, where +1 indicates perfect score, 0 indicates random performance and values of 0.5 or higher are generally considered to indicate acceptable model performance [49, 50]. The TSS value is not affected by the size of the validation dataset. Evaluation values for the cross-validation runs were then compared to the values from the runs from the quasi-independent data, checking for consistency in predictive accuracy scores.

The two best-performing model algorithms, based on ROC and TSS scores, were then selected for ensemble projection. The evaluation summaries of ensemble predictions are presented as mean ROC and TSS, median ROC and TSS, and lower and upper confidence bounds of ROC and TSS. Sensitivity and specificity were calculated and a threshold value that maximizes their sum (optimal threshold value for each condition) was considered to generate binary maps that display areas where LF transmission is more likely to occur based on environmental suitability.

A gridded map of estimated population density for Nigeria was obtained from the WorldPop Africa dataset [51]. Population density data available for Nigeria from this resource were for the years 2006, 2010, 2015 and 2020. As our data spanned from 2000–2013, we estimated the population based on population density estimates for the year 2010. We calculated population for each state by summing estimated numbers of people per pixel falling within predicted LF suitable areas and aggregating this to represent population by state. This analysis was performed using the Zonal Statistics function available within the Spatial Analyst Tool in ArcGIS 10.3 [52].

Results

Analysis was performed using 1192 survey sites reporting presence-absence of LF covering all 36 states and the Federal Capital Territory in Nigeria. Survey participants were tested for the presence of filarial antigenemia using ICT. In total, 142,881 individuals were surveyed and 11,479 tested positive for LF infection.

Figure 2 shows the performance of the seven model algorithms implemented in the BIOMOD package. Here, the RF and GBM models outperform the others with area under the ROC and TSS values of > 0.95 and > 0.75, respectively. The two hundred models generated by these two algorithms where therefore chosen for constructing the final ensemble model.

Figures 3 and 4 show the relative contribution (RC, as percentage) and marginal effect plots of each covariate on the predicted suitability of occurrence for LF to the final GBM and RF models, respectively. For both model algorithms, precipitation of the driest quarter, precipitation of the wettest quarter, and elevation, were the major contributors to the ensemble of models. In total these three covariates contributed 60.91 and 59.38% to the fitted ensemble of GBM and RF models, respectively. The probability of LF occurrence appeared to steadily decrease with increasing precipitation of the driest quarter. High suitability for LF was positively associated with elevation up to 500 metres above sea level (masl), and then appeared to flatten up to 1500 masl. Land surface temperature and distance to stable lights also showed a negative correlation with LF occurrence.

A continuous risk map of environmental suitability of LF was projected on a geographical space based on the pre-selected environmental predictors (those shown in Figs. 3 and 4). The mean area under the ROC values on the evaluation dataset (30% of the full dataset) for the final ensemble model was 0.991 and the median was 0.993 (95% CI: 0.879–0.995). These areas under the ROC values measure the performance of the final ensemble model in fitting the presence-absence data and predicting across unsampled locations.

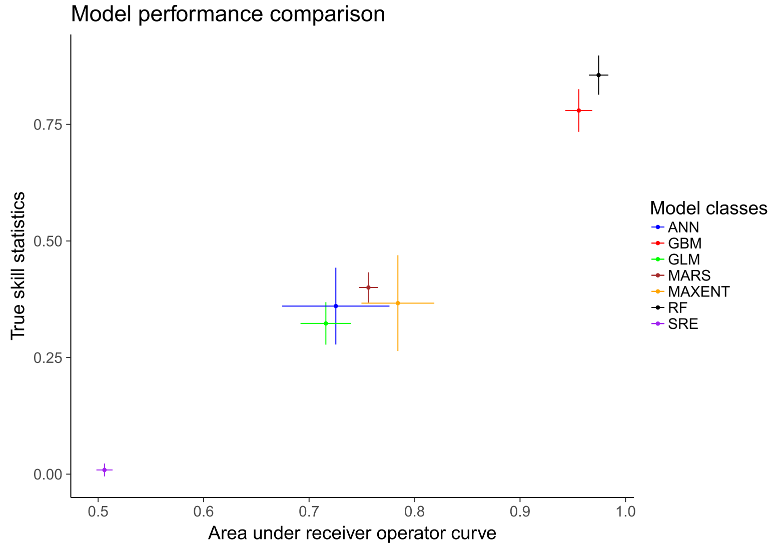

Fig. 2 Model performance comparison by area under the receiver operating characteristic curve (ROC) and true skill statistic (TSS) values of all model classes. The points represent the mean estimates and the solid lines represent the 95% confidence intervals. *Abbreviations*: ANN, artificial neural networks; GBM, generalised boosted models; GLM, generalised linear models; MARS, multivariate additive regression splines; MAXENT, maximum entropy ecological niche models; RF, random forest; SRE, surface range envelope

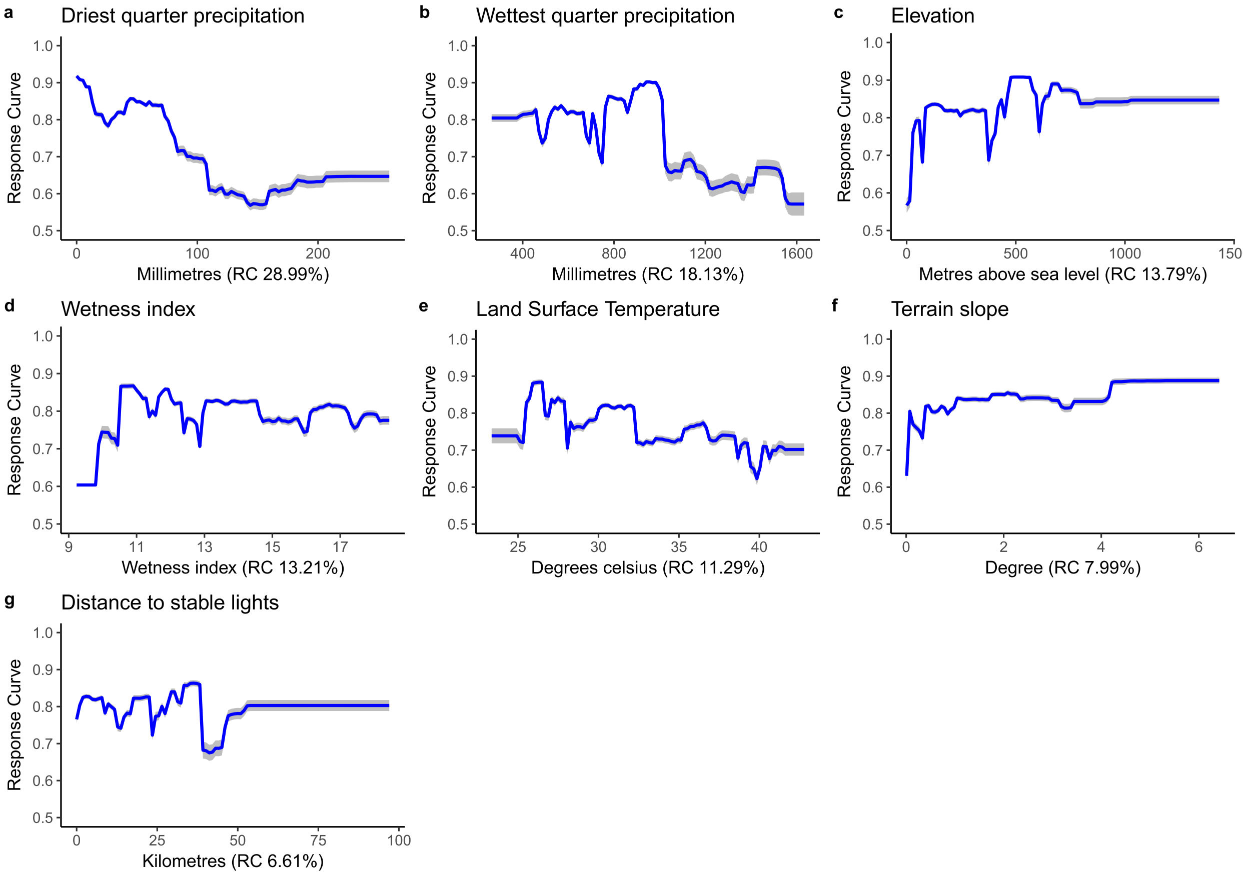

Fig. 3 Marginal effects curves for covariates included in 100 ensembles of generalised boosted models. Blue lines represent the mean marginal effect and grey shading indicates the 95% bootstrap confidence intervals. The figures in parentheses indicate the relative contribution (RC) of each covariate, which add up to 100%. The Y-axis is the response (probability of LF occurrence) and the X-axis is the full range of covariate values

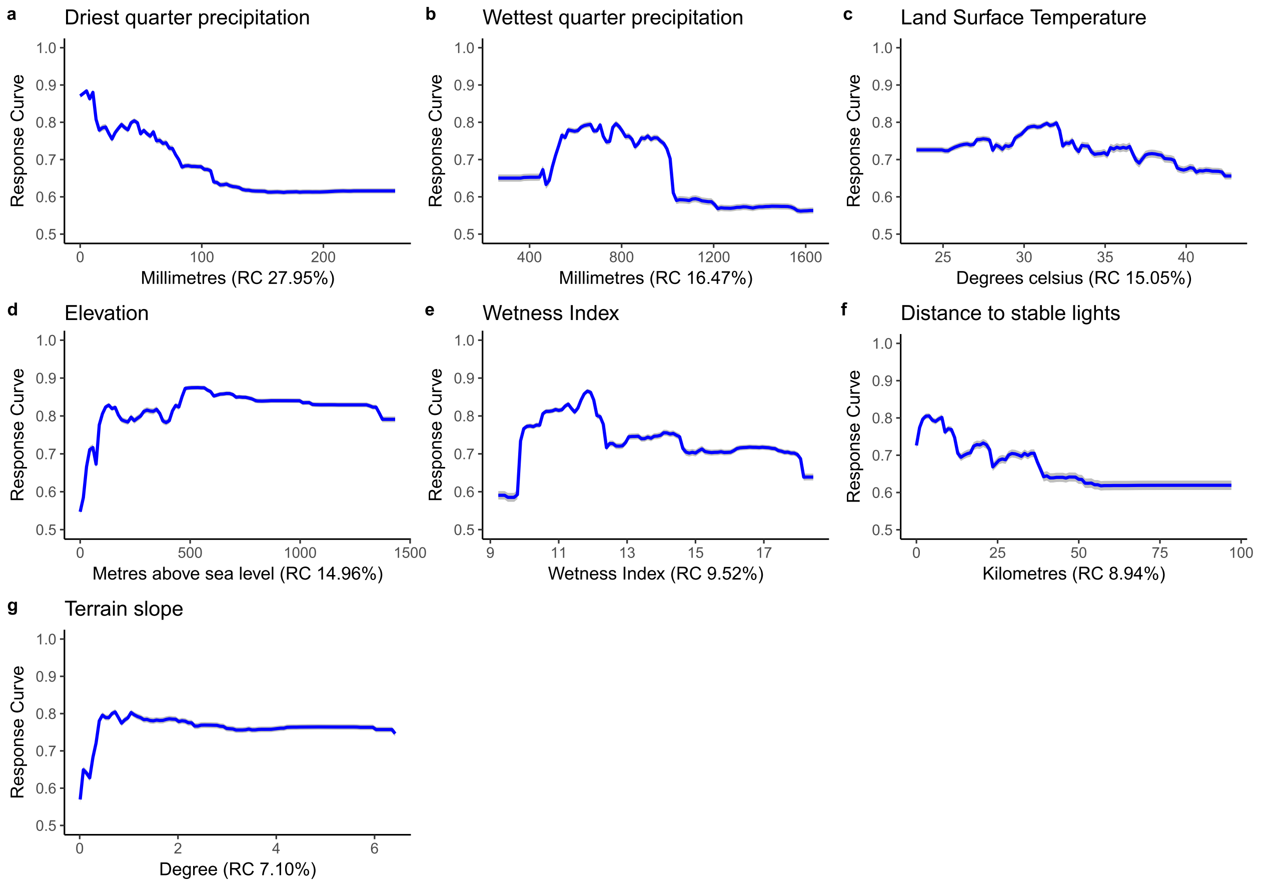

Fig. 4 Marginal effects curves for covariates included in 100 ensembles of random forest models. Blue lines represent the mean marginal effect and grey shading indicates the 95% bootstrap confidence intervals. The figures in parentheses indicate the relative contribution (RC) of each covariate, which add up to 100%. The Y-axis is the response (probability of LF occurrence) and the X-axis is the full range of covariate values

The map presented in Fig. 5 suggests that large proportions of northern Nigeria are environmentally suitable and better able to drive LF transmission, although low suitability was predicted in the north-central state of Kogi. Low LF suitability was however predicted in most southern states in Nigeria.

A suitability threshold of 0.711 with a sensitivity of 95% and a specificity of 96.2% provided the best discrimination between presence and absence values according to the evaluation dataset, and therefore was used to reclassify the continuous predictive maps into binary maps, delineating land areas into either suitable or unsuitable for LF transmission (Fig. 6).

Estimating population at risk

The total national population living in areas that are environmentally suitable for LF was estimated to be 110 (95% CI: 106–124) million, which corresponds to about 67% of Nigeria's population in 2010. The largest portion of the population living in areas environmentally suited to LF transmission were found in Kano, Kaduna, and Katsina states with predicted populations of 9.6, 5.9 and 5.6 million, respectively (Table 2). All other states had population in at-risk areas of less than 5 million.

Discussion

In this study, we have produced maps at a resolution of 1 km^2 to inform ongoing interventions by delineating areas of highest transmission risk which are prone to resurgence, thus to help in efficiently targeting control measures at the lowest administrative level. Our occurrence map (Fig. 6) indicates that suitability to LF transmission is widely distributed in Nigeria. However, parts of the north-east state of Borno, southern states of Cross River, Rivers, Akwa Ibom, Delta and Edo, and south-west states of Lagos, Oyo, Ogun and Ondo, are not environmentally suited to LF transmission. According to our estimates, about 67% of the Nigerian population live in areas environmentally suitable for LF transmission.

The benefits of machine learning algorithms compared to logistic regression models for niche and species distribution modelling have been thoroughly reviewed [45, 53–58]. Machine learning algorithms allow to account for complex non-linear associations between the response (e.g. disease

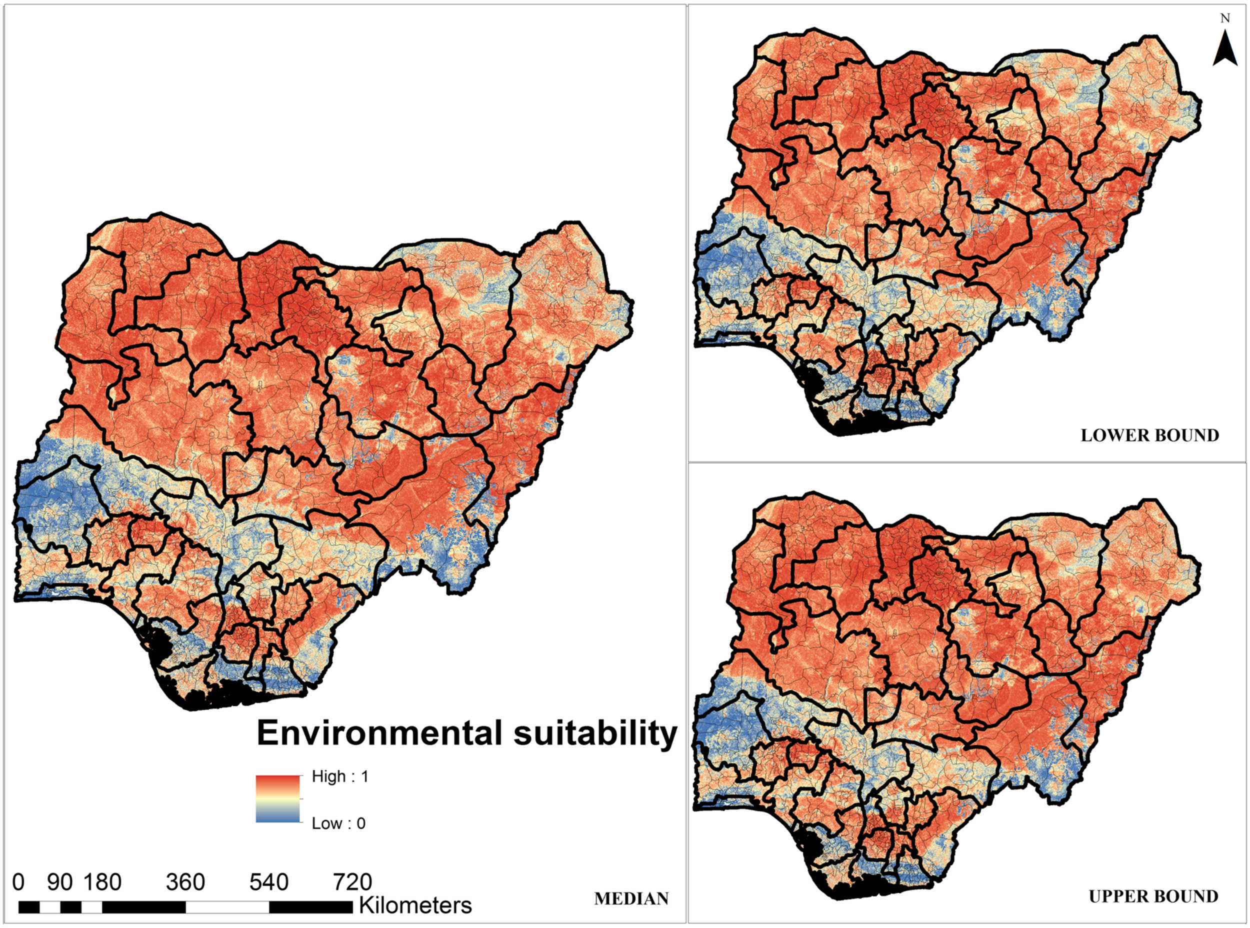

Fig. 5 Median predicted environmental suitability of LF with the lower and upper bounds of the occurrence limits

occurrence) and explanatory variables, and control for interactions among explanatory variables [19]. Furthermore, combining predictions from more than one modelling algorithm to form an ensemble has been found to produce more precise estimates when used for ecological niche and species distribution models [59]. In our work, we have mapped the ecological niche of LF in Nigeria using an ensemble of two algorithms (GBM and RF), which are widely considered to produce accurate estimates of disease distribution [46, 58]. GBM models combine fitting regression trees with boosting, by recursively partitioning data into smaller binary splits, and splits repeatedly applied to their own until the best split is chosen [46]. 'Boosting' allows fine tuning of the overall model and is performed in a forward stepwise manner minimising residual variation in the response [46]. The combination of regression trees and boosting have been demonstrated to avoid over-fitting [45]. For RF models, a large number of trees are grown with the root node of each new tree containing a different random bootstrap prediction of the original data [58, 60]. For final predictions, the average values of all bootstrapped predictions are taken. It has been demonstrated that RF models are efficient in tuning and improving the accuracy of models [58].

Our work provides an insight into the regional distribution of LF in Nigeria, and we find that the areas less suitable for LF transmission correspond to mangrove ecosystems and freshwater swamps in the southern parts of the country, and also to short grass savanna in the north-east [61]. We have identified environmental factors associated with the occurrence of LF in Nigeria, with precipitation during the driest quarter contributing the most in driving the probability of LF occurrence. This finding shows that availability of temporal breeding sites during the driest period is critical for the major LF vectors, *Anopheles* spp. mosquitoes, to sustain the transmission [1]. However, marginal effect plots also showed that the probability of LF occurrence would decline when precipitation exceeds 800 mm during the wettest quarter of the year, which may suggest that although rainfall is required for vector survival and breeding, excessive rainfall may cause flooding and destruction of breeding sites [16, 19]. Similarly, the probability of LF occurrence started to decline at high land surface

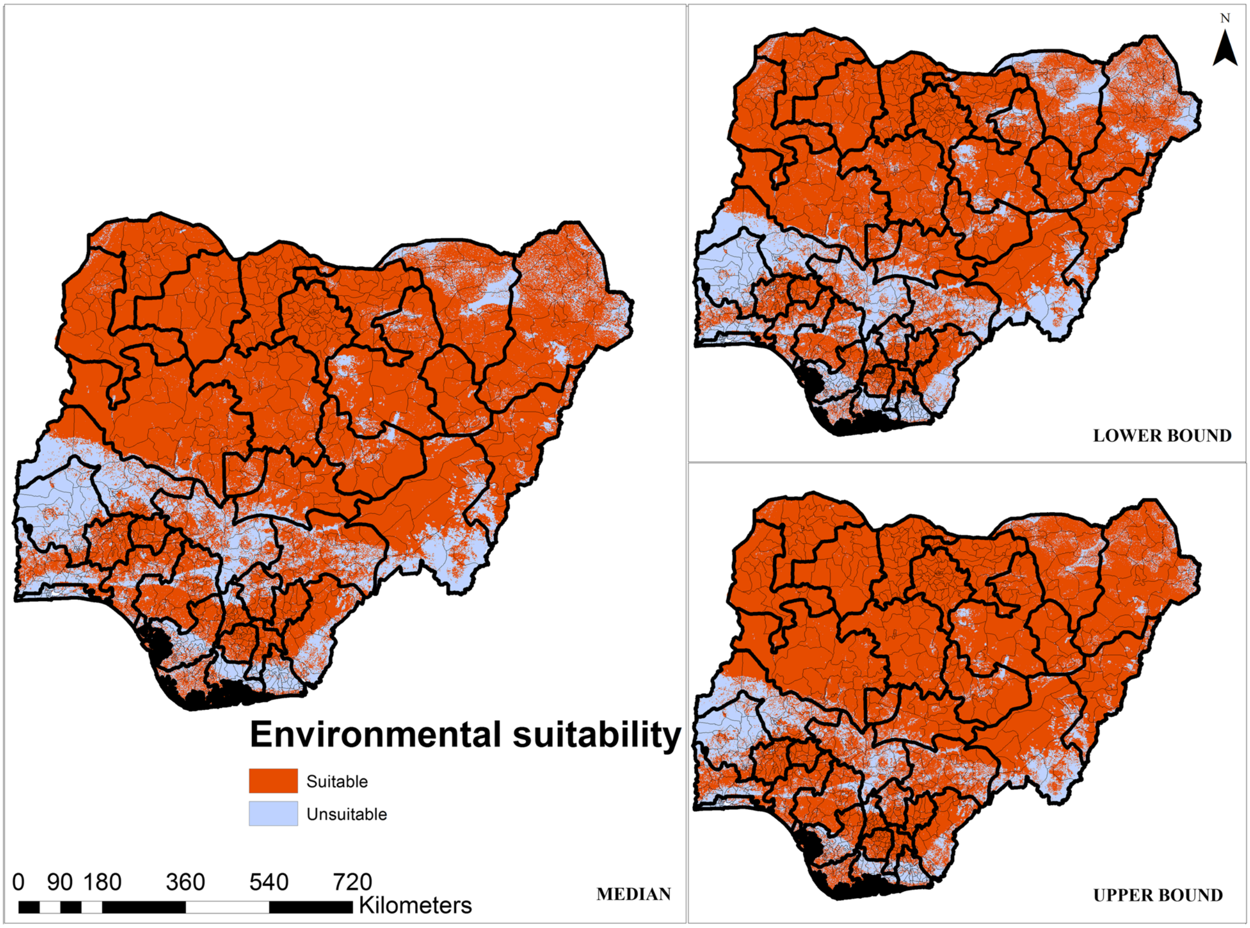

Fig. 6 Median predicted binary map of environmental suitability for LF with the lower and upper bounds of the occurrence limits

temperatures. This is consistent with experimental findings on adult mosquito survival and larval development [62–64], which suggest that both adults and larvae are unable to thrive at high temperatures.

The probability of LF occurrence appeared to increase with increasing elevation, and levels off at around 500 metres above sea level. This phenomenon has been previously recorded [16, 19] and is thought to reflect the negative effect of decreasing temperature with increasing altitude on mosquito survival and the rate of parasite development within the vector [65]. We found a negative correlation between higher terrain slope and suitability of environment to LF, perhaps because steeper inclinations of terrain cause more rapid surface water runoff, thus reducing the collection of water pockets which may serve as breeding sites for mosquito vectors.

Finally, it seems that environmental suitability for LF steadily declined with increasing distance from stable lights. Stable light is considered for any pixels at a value of > 0; however, values < 0 do not mean total pixel darkness but rather that the intensity of light emitted does not reach the threshold to be captured by the sensor. Here we took stable lights as proxy for rural-urban divide and economic activity, as urban areas are more likely to emit night-light (and thus stable lights) than rural areas. As the distance from stable night-light increased, the probability for LF occurrence decreased. This drop may be explained by the absence of stable lights in uninhabited areas where the mosquito population is more likely to be of low abundance, or in more rural settings where stable lights are less likely to be present, as electricity is in short supply in large parts of rural Nigeria. Although LF has always been associated with more rural areas [66–68], a recent study in Tanzania has highlighted the burden of LF in urban settings [69], and corroborated in a study conducted in an urban Nigerian setting [68]. Studies have also illustrated that mosquitoes are more likely to aggregate around human populations [70, 71]. Smith et al. [70] reported that the distribution of the human population influenced the aggregation of adult mosquitoes because mosquitoes are more likely to gravitate towards the human host. These authors demonstrated that mosquito density was lowest in rural settings but higher in peri-urban and urban settings.

Table 2 Estimated human population living in areas environmentally suited to LF transmission by state in Nigeria in 2010

Zones in Nigeria		Population in areas environmentally suited to LF transmission	Total population
North-central States	Benue	2,997,209	4,853,000
	Kogi	1,299,057	3,838,000
	Kwara	841,730	2,852,000
	Nassarawa	2,007,317	2,151,000
	Niger	4,342,252	4,538,000
	Plateau	3,568,619	3,659,000
	FCT	1,438,127	1,537,000
Subtotal		16,494,311	23,428,000
North-east States	Adamawa	3,087,599	3,272,000
	Bauchi	4,893,787	5,257,000
	Borno	4,115,294	4,752,000
	Gombe	2,755,106	2,773,000
	Taraba	2,061,291	2,657,000
	Yobe	2,228,136	2,652,000
Subtotal		19,141,213	21,363,000
North-west States	Jigawa	4,701,572	5,054,000
	Kaduna	5,903,960	6,927,000
	Kano	9,625,825	10,765,000
	Katsina	5,640,911	6,550,000
	Kebbi	3,700,485	3,758,000
	Sokoto	3,973,503	4,137,000
	Zamfara	3,462,653	3,689,000
Subtotal		37,008,909	40,880,000
South-east States	Abia	2,034,246	3,269,000
	Anambra	4,314,081	4,819,000
	Ebonyi	2,251,489	2,345,000
	Enugu	2,778,413	3,717,000
	Imo	4,190,754	4,402,000
Subtotal		15,568,983	18,552,000
South-south States	Akwa Ibom	1,073,592	4,461,000
	Cross River	2,351,796	3,472,000
	Bayelsa	1,206,577	2,087,000
	Rivers	1,630,531	5,759,000
	Delta	2,025,928	4,747,000
	Edo	2,910,697	3,804,000
Subtotal		11,199,121	24,330,000
South-west States	Ekiti	2,357,067	2,516,000
	Lagos	538,364	14,480,000
	Ogun	1,281,100	3,953,000
	Ondo	2,608,852	3,679,000
	Osun	3,020,890	4,105,000
	Oyo	1,496,952	6,532,000
Subtotal		11,303,129	35,265,000
Sum total		110,715,856	163,818,000

In Nigeria, *Anopheles* spp. are the principal vectors for LF [1, 72, 73], and our maps of environmental suitability correspond well with known historical distribution patterns of these mosquitoes in Nigeria [1]. This may be due to the relatively stable nature of the peri-domestic environmental factors driving the abundance and distribution of *Anopheles* mosquitoes in the past 20 years [16]. Previous studies have demonstrated that environmental factors may affect mosquito species differently. For instance, precipitation, which was a major driving factor in our model, is thought to have a greater effect on *Anopheles* spp. than it does on *Culex* spp. [16], perhaps due to their different breeding habitats. *Culex* mosquitoes are known to breed in areas with poor sanitary and housing conditions [3], thus human factors may play a more important role than precipitation. It will therefore be interesting to test our model in geographical areas where *Culex* spp. are the predominant vectors for LF. Furthermore, although LF transmission has been interrupted in the north-central states of Plateau and Nasarawa [26], the mosquito vectors remain; thus there is a risk of recrudescence due to within-country human migration [51] and a possible re-introduction of the LF parasite.

Estimates of the human population living in areas environmentally suited to LF transmission have steadily increased over the years and varied significantly in previous studies [74–76]. This may be due to improved diagnosis and surveillance as well as population growth. In 1992, an estimated 113 million people lived in LF at-risk areas in Africa [74] and by 2009, this figure was estimated to be 212 million [76]. Nigeria was reported to have the third highest national LF burden with an estimated 114 million individuals living in at-risk areas in 2016 [15]. These estimates are usually calculated by summing up the populations of each of district where infection is detected, and may thus overestimate the actual LF burden since this approach does not account for within-district spatial variations and is highly dependent on the existence of field survey data in endemic areas. In addition, field surveys are more likely to be carried out where LF infection is suspected or in areas where universities are located [77] and thus surveillance may in some locations fail to capture the true infectious status of vast areas of the country. Since 2000, Nigeria has seen an increase in violence and militancy in the southern Niger Delta regions [78], and terrorist activities in the north-eastern parts of the country [78, 79]. These conditions make it difficult to carry out field surveys and data from these areas are patchy. In contrast, estimates of the human population at risk derived by geostatistical and machine learning-based models estimate the complete distribution of infection and thus may produce more accurate estimates of the true extent of infection and the human populations at risk [19]. The additional prospect of our model producing a threshold value that maximises the sum of specificity and sensitivity improves the accuracy of our binary map (Fig. 6) [48, 80].

The human population living in areas classified as environmentally suitable for LF transmission in Nigeria using a machine learning approach was previously estimated to be 143.97 million [19]. This figure was derived from modelling only 27 presence-only data points from surveys conducted between 1977 and 1990 and of varying diagnostic methods and study designs, whereas our model was implemented using 1192 presence-absence data points derived from standardised field surveys. Furthermore, niche models derived by using presence-only data points are often spatially biased as surveys are usually conducted in areas that are more easily accessible, and are usually positive (presence) counts [81]. This bias is usually remedied by selecting background or pseudo-absence data points [82], that is assumed absence data drawn at random for the region of interest, to balance the presence-absence data point ratio. Since presence points are usually concentrated in particular geographical regions due to convenience of sampling, by randomly selecting pseudo-absence data points analysis may be biased as true presence points might be treated as 'absence', and thus leading to inaccurate model predictions [81]. The geographical spread and the amount of presence-absence data used in the present study, together with the standardised methods for field surveys and data collection, implies that our models provide accurate population estimates of the number of individuals living in areas environmentally suited to LF transmission.

Conclusions

The data used in this analysis represent a unique resource and provide the most comprehensive database for LF distribution in Nigeria. As the national LF control programme moves towards elimination, the methods and results presented in this study will inform surveillance activities and help optimise resource allocation for disease control.

Abbreviations

AfSIS: African Soil Information System; ANN: Artificial neural networks; BIOMOD: Biodiversity modelling; BRT: Boosted regression trees; CI: Confidence interval; DEM: Digital elevation model; ENM: Ecological niche model; EVI: Enhanced vegetation index; GBM: Generalised boosted models; GLM: Generalised linear models; ICT: Immunochromatographic card tests; ISRIC: Soil Reference and Information Centre; IU: Implementation unit; LF: Lymphatic filariasis; LGA: Local government area; LST: Land surface temperature; MARS: Multivariate additive regression splines; MaxEnt: Maximum entropy ecological niche models; MDA: Mass drug administration; MERIS: Medium resolution imaging spectrometer; MODIS: Moderate resolution imaging spectoradiometer; NL: Night-light; NLFEP: Nigerian Lymphatic Filariasis Elimination Programme; PET: Potential evapo-transpiration; RC: Relative contribution; RF: Random Forest; ROC: Receiver operating characteristic curve; SCA: Specific catchment area; SRE: Surface range envelopes; SRTM: Shuttle Radar Topography Mission; TAS: Transmission assessment survey; TSS: True skill statistic; TWI: Topographic wetness index; VIF: Variance inflation factor; WHO: World Health Organization

Acknowledgements

We are grateful to Ministry of Health staff in Nigeria who conducted data collection and to all participants who gave their time generously during the surveys.

Funding

The Commonwealth Scholarship Commission provided funding for the doctoral studentship of OAE. ID thanks the Imperial College Junior Research Fellowship scheme. JC is funded by grants from the Bill and Melinda Gates Foundation (OPP 1033751). All authors thank the Medical Research Council, UK, for Centre funding. Funders played no role in formal analysis, interpretation of results or decision to publish this study.

Availability of data and materials

Survey datasets analysed during this current study are available in the Expanded Special Project for Elimination of Neglected Tropical Diseases (ESPEN) repository http://espen.afro.who.int/countries/nigeria. Predictor datasets and R code are available upon reasonable request to the corresponding author.

Authors' contributions

OAE, CAD, TG and JC conceived and designed the study. OAE carried out formal analysis, visualisation and wrote the first draft of manuscript. JC and ID advised on spatial and machine learning models. IA and CO contributed survey data. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The process of obtaining ethical approvals, informed consent, and arranging logistical procedures for the field surveys were handled in-country by the Nigeria Ministry of Health, with technical support by the WHO.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK. [2]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [3]Federal Ministry of Health, Abuja, Nigeria. [4]Department of Statistics, University of Oxford, Oxford, UK.

Received: 16 May 2018 Accepted: 4 September 2018
Published online: 17 September 2018

References

1. Okorie PN, McKenzie FE, Ademowo OG, Bockarie M, Kelly-Hope L. Nigeria *Anopheles* vector database: an overview of 100 years' research. PLoS One. 2011;6:e28347.
2. Lindsay SW, Parson L, Thomas CJ. Mapping the ranges and relative abundance of the two principal African malaria vectors, *Anopheles gambiae sensu stricto* and *An. arabiensis*, using climate data. Proc Biol Sci. 1998;265:847–54.
3. Knudsen AB, Slooff R. Vector-borne disease problems in rapid urbanization: new approaches to vector control. Bull World Health Organ. 1992;70:1–6.
4. Simonsen PE, Mwakitalu ME. Urban lymphatic filariasis. Parasitol Res. 2013;112:35–44.
5. Addiss DG, Dimock KA, Eberhard ML, Lammie PJ. Clinical, parasitologic, and immunologic observations of patients with hydrocele and elephantiasis in an area with endemic lymphatic filariasis. J Infect Dis. 1995;171:755–8.
6. CDC. Recommendations of the International Task Force for Disease Eradication. MMWR weekly morbidity and mortalilty report. Atlanta, Georgia: Centers for Disease Control; 1993.

7. Hairston NG, de Meillon B. On the inefficiency of transmission of *Wuchereria bancrofti* from mosquito to human host. Bull World Health Organ. 1968;38:935–941.
8. Lek-Uthai U, Tomoen W. Susceptibility of *Mansonia uniformis* to *Brugia malayi* microfilariae from infected domestic cat. Southeast Asian J Trop Med Publ Health. 2005;36:434–41.
9. Rebollo MP, Bockarie MJ. Toward the elimination of lymphatic filariasis by 2020: treatment update and impact assessment for the endgame. Expert Rev Anti Infect Ther. 2013;11:723–31.
10. Bockarie MJ, Kelly-Hope LA, Rebollo M, Molyneux DH. Preventive chemotherapy as a strategy for elimination of neglected tropical parasitic diseases: endgame challenges. Philos Trans Roy Soc London Series B. Biol Sci. 2013;368:20120144.
11. Ottesen EA, Duke BO, Karam M, Behbehani K. Strategies and tools for the control/elimination of lymphatic filariasis. Bull World Health Organ. 1997;75:491–503.
12. Hawking F. The distribution of bancroftian filariasis in Africa. Bull World Health Organ. 1957;16:581–92.
13. Hawking F. The distribution of human filariasis throughout the world. Part III. Africa. Trop Dis Bull. 1977;74:649–79.
14. Federal Ministry of Health Nigeria. Nigeria Master Plan for Neglected Tropical Diseases (NTDs) 2013–2017. Abuja, Nigeria: FMoH, Nigeria; 2012.
15. Federal Ministry of Health Nigeria. Neglected Tropical Diseases Nigeria Multi-Year Master Plan 2015-2020. Abuja, Nigeria: FMoH, Nigeria; 2016.
16. Cano J, Rebollo MP, Golding N, Pullan RL, Crellen T, Soler A, et al. The global distribution and transmission limits of lymphatic filariasis: past and present. Parasit Vectors. 2014;7:466.
17. Michael E, Bundy DAP, Grenfell BT. Re-assessing the global prevalence and distribution of lymphatic filariasis. Parasitology. 1996;112:409–28.
18. 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;96:695–705.
19. Slater H, Michael E. Predicting the current and future potential distributions of lymphatic filariasis in Africa using maximum entropy ecological niche modelling. PLoS One. 2012;7:e32202.
20. O'Hanlon SJ, Slater HC, Cheke RA, Boatin4 BA, Coffeng LE, Pion SDS, et al. Model-based geostatistical mapping of the prevalence of *Onchocerca volvulus* in West Africa. PLoS Negl Trop Dis. 2016;10:e0004328.
21. Slater H, Michael E. Mapping, bayesian geostatistical analysis and spatial prediction of lymphatic filariasis prevalence in Africa. PLoS One. 2013;8:e71574.
22. Rebollo MP, Sime H, Assefa A, Cano J, Deribe K, Gonzalez-Escalada A, et al. Shrinking the lymphatic filariasis map of Ethiopia: reassessing the population at risk through nationwide mapping. PLoS Negl Trop Dis. 2015;9:15.
23. 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;8:560.
24. Deribe K, Cano J, Newport MJ, Golding N, Pullan RL, Sime H, et al. Mapping and modelling the geographical distribution and environmental limits of podoconiosis in Ethiopia. PLoS Negl Trop Dis. 2015;9:7.
25. Expanded Special Projects for Elimination of Neglected Tropical Diseases (ESPEN). Status of lymphatic filariasis MDA (2005–2016) - Nigeria. Geneva: WHO; 2018. http://espen.afro.who.int/system/files/content/maps/WHO_LF_IU_MDA_TC_trend_NG.pdf.
26. The Cater Centre. Two states in Nigeria eliminate disfiguring parasitic disease lymphatic filariasis as public health problem. 2017. https://www.cartercenter.org/news/pr/nigeria-101317.html. Accessed 27 Feb 2018.
27. WHO. Operational guidelines for rapid mapping of Bancroftian filariasis in Africa. Geneva: World Health Organization; 2000.
28. 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. 1996;29:129–51.
29. WorldClim. Free climate data for ecological modeling and GIS. 2017. http://worldclim.com/. Accessed 1 May 2017.
30. Consortium for Spatial Information. 2017. www.cgiar-csi.org. Accessed 1 May 2017.
31. NASA. Enhanced shuttle land elevation data. 2017. https://www2.jpl.nasa.gov/srtm/. Accessed 1 May 2017.
32. Qin C-Z, Zhu A-X, Pei T, Li B-L, Scholten T, Behrens T, et al. An approach to computing topographic wetness index based on maximum downslope gradient. Prec Agric. 2011;12:32–43.
33. Lehner B, Döll P. Development and validation of a global database of lakes, reservoirs and wetlands. J Hydrol. 2004;296:1–22.
34. Digital Global Chart: inland waters. 2017. http://www.diva-gis.org/gdata. Accessed 1 May 2017.
35. Africa Soil Information Service. The collection of Africa continient-wide grids include data from MODIS, TRMM, WorldClim and ESA. 2017. http://africasoils.net. Accessed 3 May 2017.
36. Arino O, Gross D, Ranera F, Leroy M, Bicheron P, Brockman C, et al., editors. GlobCover: ESA service for global land cover from MERIS. 2007 IEEE International Geoscience and Remote Sensing Symposium, 23–28 July 2007, Barcelona, Spain; 2007.
37. ISRIC - World Soil Information. Soil property maps of Africa at 1 km. 2013. http://www.isric.org/. Accessed 3 May 2017.
38. Elvidge CD, Baugh KE, Kihn EA, Kroehl HW, Davis ER. Mapping city lights with nighttime data from the DMSP operational linescan system. Photogramm Eng Remote Sens. 1997;63:727–34.
39. Doll CNH, Muller JP, Morley JG. Mapping regional economic activity from night-time light satellite imagery. Ecol Econ. 2006;57:75–92.
40. Ebener S, Murray C, Tandon A, Elvidge CC. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. Int J Health Geogr. 2005;4:5.
41. Noor AM, Alegana VA, Gething PW, Tatem AJ, Snow RW. Using remotely sensed night-time light as a proxy for poverty in Africa. Popul Health Metr. 2008;6:5.
42. Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour Res. 2003;39:1199–214.
43. R Developement Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
44. Craney TA, Surles JG. Model-dependent variance inflation factor cutoff values. Qual Eng. 2002;14:391–403.
45. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL. The global distribution and burden of dengue. Nature. 2013;496:504–7.
46. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77:802–13.
47. Rogers DJ. Models for vectors and vector-borne diseases. Adv Parasitol. 2006;62:1–399.
48. Thuiller W, Lafourcade B, Engler R, Araújo MB. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography. 2009;32:369–73.
49. Araújo MB, Alagador D, Cabeza M, Nogués-Bravo D, Thuiller W. Climate change threatens European conservation areas. Ecol Lett. 2011;14:484–92.
50. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006;43:1223–32.
51. Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One. 2012;7:e31743.
52. Esri - ArcGIS 10.3. Geographic information system software. 2017. https://www.esri.com/en-us/home.
53. Pearce JL, Boyce MS. Modelling distribution and abundance with presence-only data. J Appl Ecol. 2006;43:405–12.
54. Julian DO, Joshua JL, LeRoy NP. Machine learning methods without tears: a primer for ecologists. Q Rev Biol. 2008;83:171–93.
55. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006;190:231–59.
56. Thuiller W, Araújo MB, Lavorel S. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. J Veg Sci. 2003;14:669–80.
57. Machado-Machado EA. Empirical mapping of suitability to dengue fever in Mexico using species distribution modeling. Appl Geogr. 2012;33:82–93.
58. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
59. Araújo MB, New M. Ensemble forecasting of species distributions. Trends Ecol Evol. 2007;22:42–7.
60. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40.
61. The University of Texas at Austin. Nigeria Maps: Map Collection 2018. 2018. http://legacy.lib.utexas.edu/maps/nigeria.html. Accessed 14 Feb 2018.
62. Martens WJM, Jetten TH, Focks DA. Sensitivity of malaria, schistosomiasis and dengue to global warming. Clim Change. 1997;35:145–56.
63. Lardeux F, Cheffort J. Temperature thresholds and statistical modelling of larval *Wuchereria bancrofti* (Filariidea: Onchocercidae) developmental rates. Parasitology. 1997;114:123–34.

64. Lardeux F, Cheffort J. Ambient temperature effects on the extrinsic incubation period of *Wuchereria bancrofti* in *Aedes polynesiensis*: implications for filariasis transmission dynamics and distribution in French Polynesia. Med Vet Entomol. 2001;15:167–76.
65. Ngwira BM, Tambala P, Perez AM, Bowie C, Molyneux DH. The geographical distribution of lymphatic filariasis infection in Malawi. Filaria J. 2007;6:12.
66. Durrheim DN, Wynd S, Liese B, Gyapong JO. Editorial: Lymphatic filariasis endemicity - an indicator of poverty? Trop Med Int Health. 2004;9:843–5.
67. Hotez P. Forgotten people, forgotten diseases: the neglected tropical diseases and their impact on global health and development. Washington DC: ASM Press; 2008.
68. Terranella A, Eigiege A, Gontor I, Dagwa P, Damishi S, Miri E, et al. Urban lymphatic filariasis in central Nigeria. Ann Trop Med Parasitol. 2006;100:163–72.
69. Mwingira U, Chikawe M, Mandara WL, Mableson HE, Uisso C, Mremi I, et al. Lymphatic filariasis patient identification in a large urban area of Tanzania: an application of a community-led Mhealth system. PLoS Negl Trop Dis. 2017;11:e0005748.
70. Smith DL, Dushoff J, McKenzie FE. The risk of a mosquito-borne infection in a heterogeneous environment. PLoS Biol. 2004;2:e368.
71. Martens P, Hall L. Malaria on the move: human population movement and malaria transmission. Emerg Infect Dis. 2000;6:103–9.
72. Okorie PN, Ademowo GO, Saka Y, Davies E, Okoronkwo C, Bockarie MJ, et al. Lymphatic filariasis in Nigeria; micro-stratification overlap mapping (MOM) as a prerequisite for cost-effective resource utilization in control and surveillance. PLoS Negl Trop Dis. 2013;7:e2416.
73. Brant TA, Okorie PN, Ogunmola O, Ojeyode NB, Fatunade SB, Davies E, et al. Integrated risk mapping and landscape characterisation of lymphatic filariasis and loiasis in South West Nigeria. Parasite Epidemiol Control. 2018;3:21–35.
74. WHO Expert Committee on Filariasis & World Health Organisation. Lymphatic filariasis: the disease and its control, fifth report of the WHO Expert Committee on Filariasis [meeting held in Geneva from 1 to 8 October 1991]. Geneva: World Health Organisation; 1992.
75. WHO/Department of Communicable Disease Prevention, Control and Eradication. Global Programme to Eliminate Lymphatic Filariasis: Progress report 2004. Wkly Epidemiol Record. 2005;80:201–12.
76. WHO/Department of Control of Neglected Tropical Diseases Global Programme to Eliminate Lymphatic Filariasis: Progress report on mass drug administration in 2008. Wkly Epidemiol Rec. 2009;42:437–44.
77. Okorie PN, Bockarie MJ, Molyneux DH, Kelly-Hope LA. Neglected tropical diseases: a systematic evaluation of research capacity in Nigeria. PLoS Negl Trop Dis. 2014;8:e3078.
78. Abimbola S, Malik AU, Mansoor GF. The final push for polio eradication: addressing the challenge of violence in Afghanistan, Pakistan, and Nigeria. PLoS Med. 2013;10:e1001529.
79. Adebayo AA. Implications of 'Boko Haram' terrorism on national development in Nigeria: a critical review. Mediterr J Soc Sci. 2014;5:480–9.
80. Jiménez-Valverde A, Lobo JM. Threshold criteria for conversion of probability of species presence to either-or presence-absence. Acta Oecol. 2007;31:361–9.
81. Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, et al. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl. 2009;19:181–97.
82. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol. 2012;3:327–38.