Landscape Ecol
https://doi.org/10.1007/s10980-019-00839-2

RESEARCH ARTICLE

An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns

Shi Chen · Ari Whiteman · Ang Li · Tyler Rapp · Eric Delmelle · Gang Chen · Cheryl L. Brown · Patrick Robinson · Maren J. Coffman · Daniel Janies · Michael Dulin

Received: 11 February 2019 / Accepted: 14 May 2019

Abstract

Context Socioeconomic and landscape factors influence mosquito abundance especially in urban areas. Few studies addressed how socioeconomic and landscape factors, especially at micro-scale for mosquito life history, determine mosquito abundance.

Objectives We aim to predict mosquito abundance based on socioeconomic and/or landscape factors using machine learning framework. Additionally, we determine these factors' response to mosquito abundance.

Methods We identified 3985 adult mosquitoes (majority of which were *Aedes* mosquitoes) in 90 sampling sites from Charlotte, NC, USA in 2017. Seven socioeconomic and seven landscape factors were used to predict mosquito abundance. Three supervised learning models, k-nearest neighbor (kNN), artificial neural network (ANN), and support vector machine (SVM) were constructed, tuned, and evaluated using both continuous input factors and binary inputs. Random forest (RF) was used to assess individual input's relative importance and response to mosquito abundance.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10980-019-00839-2) contains supplementary material, which is available to authorized users.

S. Chen (✉) · M. Dulin
Department of Public Health Sciences, University of North Carolina Charlotte, Charlotte, NC 28223, USA
e-mail: schen56@uncc.edu

A. Whiteman · T. Rapp · E. Delmelle · G. Chen · M. Dulin
Department of Geography and Earth Sciences, University of North Carolina Charlotte, Charlotte, NC 28223, USA

A. Li
State Key Laboratory of Vegetation and Environmental Change, Chinese Academy of Sciences, Beijing 100093, China

C. L. Brown
Department of Political Science and Public Administration, University of North Carolina Charlotte, Charlotte, NC 28223, USA

S. Chen
Data Science Initiative, University of North Carolina Charlotte, Charlotte, NC 28223, USA

S. Chen · P. Robinson · M. J. Coffman · M. Dulin
Academy for Population Health Innovation, College of Health and Human Services, University of North Carolina at Charlotte, Charlotte, NC 28211, USA

D. Janies
Department of Bioinformatics, University of North Carolina Charlotte, Charlotte, NC 28223, USA

Published online: 27 May 2019

Results We showed that landscape factors alone yielded equal or better predictability than socioeconomic factors. The inclusion of both types of factors further improved model accuracy using binary inputs. kNN also had robust performance regardless of inputs (accuracy $> 95\%$ for binary and $> 99\%$ for continuous input data). Landscape factors group had higher importance than socioeconomic group (54.4% vs. 45.6%). Landscape heterogeneity (measured by Shannon index) was the single most important input factor for mosquito abundance.

Conclusions Landscape factors were the key for mosquito abundance. Machine learning models were powerful tools to handle complex datasets with multiple socioeconomic and landscape factors to accurately predict mosquito abundance.

Keywords Socioeconomic gradient · Landscape heterogeneity · Mosquito abundance · Machine learning · Urban ecology

Abbreviations

VBD	Vector-borne diseases
POP	Population size factor
INC	Income factor
EMP	Employment rate factor
EDU	Education status factor
DEN	Population density factor
PRI	Home sale price factor
VCR	Violent crime rate factor
TRE	Tree canopy factor
GRS	Grass factor
BLD	Building factor
ROD	Road factor
SHI	Shannon index
SMP	Simpson index
kNN	*k*-Nearest neighbor
ANN	Artificial neural network
SVM	Support vector machine
GLM	Generalized linear model
RF	Random forest
NMDS	Non-metric multidimensional scaling
TP	True positive
TN	True negative
FP	False positive
FN	False negative
RMSE	Root mean squared error

Introduction

Mosquitoes are the most important vector for human health and responsible for many vector-borne diseases (VBD) of human and other animals around the world including malaria (vectored by *Anopheles* spp.), West Nile disease (by *Culex* spp.), dengue fever, zika disease, and yellow fever (these three vectored by *Aedes* spp.,WHO 2018a, b). A fundamental concept underlying the transmission of infectious diseases is the epidemiology triad. Vector mosquitos connect the three components of the triad which are: pathogens, the environment, and hosts (Gordis 2013; WHO 2018a). While the pathology, etiology, and physiology sides of many VBDs have been intensively investigated and are comprehensively understood, the closely coupled and highly heterogeneous human–environment–mosquito system has yet to be fully studied.

Socioeconomic status has been identified and recognized as a critical factor for both individual and population health (Winkleby et al. 1992; Adler and Ostrove 1999; Hawe and Shiell 2000; Rael et al. 2016; Younsi and Chakroun 2016). Socioeconomic development and disparity are closely associated with the human host's nutrition status and risk of infectious diseases (Leigh 1993; Dowd et al. 2009), many of which are VBDs. For example, studies have shown that socioeconomic disparity are strongly associated with Dengue fever risk in India (Khormi and Kumar 2011); Malaria risk in Brazil (Kikuti et al. 2015; Lana et al. 2017) and Uganda (Tusting et al. 2016); and potential risk of VBD in Baltimore/DC (Dowling et al. 2013; LaDeau et al. 2013; Little et al. 2017) as well as in the southeastern US (Obenauer et al. 2017).

Socioeconomic inequality, disparity, and gradient are profound in urban areas, which draws attention to urban health in the complex, human–environment coupled system (Wu et al. 2011; Rydin et al. 2012; Eder et al. 2018; Fournet et al. 2018; Whiteman et al. 2018). Urban ecosystems usually feature a high level of landscape heterogeneity (Wu 2014), which could influence vector mosquito population dispersion and abundance, hence affecting VBD risk such as Malaria and West Nile across different landscapes (Norris 2004; Ruiz et al. 2007; Johnson et al. 2008; Ozdenerol et al. 2008; Foley et al. 2009; Gottdenker et al. 2014; Cushman and Heuttmann 2010, Drew et al. 2011, Li et al. 2014; Roiz et al. 2015, Homan et al. 2016; Eder et al. 2018;). However, there is lack of consensus how

socioeconomic and landscape patterns together determine mosquito abundance and the risk of VBD.

Geospatial techniques are frequently used in studying heterogeneous landscapes in urban ecology (Shao and Wu 2008; Cushman and Heuttmann 2010, Drew et al. 2011). Satellite imaging, remote sensing, and geographic sampling are utilized in spatial modeling of mosquito population management and VBDs such as Malaria and Chikungunya (Keating et al. 2003; Linard et al. 2009; Le Comber et al. 2011; Unlu et al. 2011; Boone et al. 2012; Reiner et al. 2013; Delmelle et al. 2016; Ruiz-Moreno 2016; Whiteman et al. 2018). While these studies provide valuable insights, several drawbacks still exist in current literature. First, the collection sites for mosquitoes usually did not cover a broad socioeconomic gradient. Second, micro-scale landscape heterogeneity was not comprehensively studied especially in the context of VBD. Adult mosquitoes have relatively short flight distance (usually a few 100 m) during its adult lifespan (Hemme et al. 2010), so micro-scale landscape is more relevant to local mosquito population and VBD risk than meso- and macro-scale landscape (Townsend et al. 2001, 2006). While the concept of landscape is generally associated with the environment in the epidemiology triad (Rydin et al. 2012), landscape itself is also influenced by human (host) activies, especially in urban landscape.

Compared with commonly used parametric methods, machine learning does not depend on a specific man-made hypothesis. Machine learning methods can handle complex data structures (e.g., non-normality, heteroscedasticity) in landscape ecology studies and implicit interactions among input factors (Fielding 1999; Baltensperger and Huettmann 2015; Humphries et al. 2018). The complicated nonlinear interactions between socioeconomic, landscape factors, and mosquito population in this study are difficult to measure and model appropriately using more commonly used hypothesis-driven parametric models (e.g., choice of order of the interaction, Jopp et al. 2011). Machine learning methods search in the (usually) high-dimensional data space, identify potential patterns which are generally difficult for parametric method to recognize, and let the data speak for themselves without assigning or relying on specific a priori hypothesis (Fielding 1999; Lantz 2015; Lesmeister 2015; Burger 2018; Humphries et al. 2018). Currently, the power of more recently developed data-driven machine learning techniques (especially supervised classification methods) has yet to be unleashed in landscape ecology and VBD (Lary et al. 2014; Cianci et al. 2015; Humphries et al. 2018; Young et al. 2018).

The objectives of this study are to: (i) identify and quantify the potential association between socioeconomic and/or landscape factors and relative mosquito abundance (encoded as high/low) across Charlotte, the largest city in North Carolina with large socioeconomic disparity and landscape heterogeneity, and (ii) develop an operational and robust machine learning model to assess potential VBD risk across Charlotte's broad socioeconomic and landscape gradient. Mathematically, this question could be written as finding the following mapping (f) from inputs (X_{socio} and/or Xl_{and}) to an output (Y, a 0 or 1 binary value) for a given sampling site i:

$$Y_i = f(X_{socio\_i}, X_{land\_i}), i \in [1, 90], where:$$

$$X_{socio\_i} = [POP, INC, EMP, EDU, DEN, PRI, VCR]_i$$

$$X_{land\_i} = [GRS, TRE, BLD, ROD, DIV, SHA, SMP]_i$$

$$Y_i = 0 \, or \, Y_i = 1$$

We build three commonly-used supervised learning models, including k-nearest neighbor (kNN), artificial neural network (ANN), and support vector machines (SVM), to quantify the association between mosquito abundance and socioeconomic, landscape as well as for both factors combined. We compare the model performance with different input factors. We also compare the machine learning methods' performance with that of a more commonly adopted traditional logistic regression model. Results derived from this study will advance our ability to identify areas with high mosquito abundance and high risk of VBDs within highly heterogeneous urban landscapes. This approach will better inform interventions designed to reduce the transmission of VBD in urban areas through advancement of public health programs, education of the public, and policy changes.

Materials and methods

Mosquito collection and data processing

Charlotte, the largest city in North Carolina with more than 800,000 inhabitants distributed unevenly across a

highly heterogeneous landscape. Charlotte is one of the fastest growing cities in the US with an accumulated growth rate of 59.6% over the past decade. Nevertheless, Charlotte also has one of the highest rates of poverty and lowest rates of social-economic mobility in the US (Chetty et al. 2014). Such fast paced growth combined with high poverty make Charlotte the city with 10th highest income inequality in the entire US (The World Bank 2015). The greater Charlotte area has been consistently infested with various species of mosquitoes (especially during summer month hence when we sampled mosquitoes) with approximately 1000 reported mosquito-borne disease cases between 2004 and 2016 (Hall et al. 2018). Worse still, many cases have been largely underreported so the risk is likely much higher (Hall et al. 2018). Thus, there is public health need to accurately quantify and predict mosquito abundance across the city's socioeconomic gradients and heterogeneous landscapes.

We collected mosquito samples with gravid traps during a 12-week long sampling period (starting early June till late August) in 2017 at 90 pre-identified sites across the city. The 12-week long period covered the most active mosquito season in Charlotte (Whiteman et al. 2018). Gravid traps had higher sensitivity to lure bloodmeal-taken female mosquitoes to oviposit, thus the sample size in gravid trap was more relevant to potential VBD risk, as female mosquitoes cannot oviposit without first biting and taking bloodmeal. The 90 sampling sites were selected to ensure covering a vast majority of the broad socioeconomic gradient in Charlotte with minimal spatial autocorrelation (Delmelle et al. 2016, Whiteman et al. 2018).

Seven socioeconomic factors associated with mosquito presence/abundance and potential VBD risk were included in this study: population size at sampling site (POP), average household income (INC), employment rate (EMP), education status (percent with at least bachelor's degree, EDU), population density (DEN), average home sale price (PRI), and violent crime rate (VCR). These seven factors were selected from a broader collection of more than 15 socioeconomic factors using a variable selection technique described in our previous study (Whiteman et al. 2018).

Orthophotos of Charlotte in 2016 (most recent as of the study) were provided by the Mecklenburg County Public Health Department through publicly accessible GIS portal. Land cover types of the sampling sites (30 m radius from the centroid of each site, representing potential natural flight distance of *Aedes* spp., the dominant mosquitoes in the study region) were analyzed and quantified from the remote-sensing images by Fragstats 4.0 (McGarigal et al. 2012) software. 30 m radius was selected to reflect potential dispersion distance of adult female *Aedes* mosquitoes (Hemme et al. 2010), which are the dominant mosquito in Charlotte and North Carolina in general. Four critical patch types that were important to mosquito presence/abundance and potential human-mosquito interactions were included: tree canopy (TRE), grass (GRS), building (BLG), and road (ROD); all quantified as percentage in the 30 m radius area of the sampling site using Fragstats 4.0 software. Water and bare soil were excluded in this study because the majority of sites did not have these two patch types. Since the main focus of this study is to investigate micro-scale landscape and its influence on mosquito abundance, it is therefore appropriate to disregard waterbodies and focus on areas with human residents in this study. In addition to patch area, Shannon–Wiener diversity index (SHA) and Simpson index (SMP) were calculated for each site to quantify micro-scale landscape diversity (patch types), which could affect mosquito abundance. Additionally, landscape division index (DIV) was calculated to estimate the probability that two mosquitoes might overlap in their geographic locations (Jaeger 2000), hence DIV could influence mosquito reproduction and population dynamics. A detailed list and description of input/output factors can be found in Table S1.

Detailed sampling and mosquito collection procedures were reported in our previous study (Whiteman et al. 2018). Since the factors' numeric ranges varied substantially, factors that were not bounded between 0 and 100 percent were first scaled between 0 and 100 percent using the following feature scaling equation for further analyses. This process could avoid potential overriding from a few factors that had large absolute numbers (e.g., INC, PRI). Besides, feature scaling involved only linear transformation, which preserved the overall shape and property of original data distribution.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}, i \in [1, 90]$$

The final complete dataset had 90 rows (representing 90 sampling sites), 7 columns of socioeconomic factors (coded as POP, INC, EMP, EDU, DEN, PRI, and VCR), 7 columns of landscape factors (TRE, GRS, BLD, ROD, DIV, SHA, and SMP), and 1 column of mosquito abundance (number of mosquitoes in each site, coded as MOS). The complete list of descriptions of these factors are provided in Table S1.

Quantifying association within and between socioeconomic and landscape factors

We calculated Pearson correlation coefficients between each pair of socioeconomic and landscape factors to identify and quantify their associations (independent of mosquito abundance) and then determined potential substantial interaction between socioeconomic and landscape factors. This step was necessary to determine potential interaction terms in the more commonly used hypothesis driven parametric models such as logistic regression, but not required for machine learning, since most machine learning models handled interactions among input factors intrinsically (Humphries et al. 2018).

Modeling association between socioeconomic and/or landscape factors and mosquito abundance

Three common supervised classification machine learning models: *k*-nearest neighbors (kNN), artificial neural network (ANN), and support vector machine (SVM) were built and tuned to quantify the association between socioeconomic and/or landscape factors and mosquito abundance (Lantz 2015; Lesmeister 2015; Burger 2018). Among many other recently developed classification models, these three were chosen because of their popularity, maturity, and practicality (with existing and stable analysis package in *R*). The input factors (socioeconomic and landscape factors) and response variable (mosquito abundance) were both dichotomized into binary values (any value smaller than the median of the factor would be assigned with 0, otherwise 1), and left untreated (raw continuous input data were fed into the model). The two sets of input data reflected availability and quality of the input data: the original continuous input factors (both socioeconomic and landscape) were (in theory) more informative but costly to collect, and places of interest that we wanted to predict mosquito abundance (usually socioeconomically disadvantaged) might not have such detailed data. In comparison, binary input data were much easier to achieve and derive, at a cost of losing information. However, as shown later in result section, we should not assume original continuous input data were naturally better than dichotomized binary in predicting mosquito abundance. Mosquito abundance was expressed as either "high" (coded as 1, $Y_i = 1$) or "low" (coded as 0, $Y_i = 0$) comparing to the median value. This would create an operational criterion for public health officers and general public to efficiently estimate relative mosquito abundance in a given site. For each type of the models, three sets of inputs were used as inputs: seven socioeconomic factors alone, seven landscape factors alone, and all socioeconomic and landscape factors combined. In summary, we had a total of 18 combinations of different models and inputs, factorized by three types of machine learning model (kNN, ANN, or SVM), three sets of input factors (socioeconomic factor alone, landscape factor alone, or combined), and two types of input factors (binary or continuous).

For kNN, the predicted value of mosquito abundance (classifier) was based on its surrounding "neighbors" in the reduced-dimensional space (usually 2-dimensional) whose "identity" (i.e., 0 or 1 value, corresponding to low and high levels of mosquito abundance) were already known. The value of k indicated number of neighbors to determine the classification. For example, $k = 1$ indicated that the predicted outcome value (high/low mosquito abundance) is governed solely by its single nearest neighbor. In general, larger k value (e.g., $k = 15$) could reduce the influence from random noise, at a cost of making the boundary of the neighbors less apparent. In this study, we tested a wide range of values, $k = 1$–15, to bracket the potential optimal k value and identify the most accurate prediction in kNN models.

ANN used hidden layers to map input "neurons" (socioeconomic and/or landscape factors) to output "neuron" (mosquito abundance). Each hidden layer further consisted of several nodes that transform and propagate the input signal. The weighted value associated with the edge between different layers of nodes could be interpreted as relative importance of the current node, conditioned on a previous layer of

nodes and similar to the interpretation of the regression coefficient (Gunther and Fritsch 2010).

SVM tried to separate the two classes (high and low mosquito abundance in this study) as far as possible (i.e., maximizing margin) in the high dimensional space (or reduced 2-dimensional space). This was done by computing a specific SVM classifier. Hyperparameter tuning of these three models were shown in supplementary material S1. Note that none of these three models were consistently better than the other two, and the model performance depended on the specific dataset, model specifications, and model tuning.

Model performance evaluation and prediction

A total of 18 machine learning models were constructed, tuned, and predicted against the observations. Cross-validation was done using a 10-fold cross-validation approach because of relatively small dataset size. We randomly split the data into 10 equal subsamples (each subsample contained nine sampling sites), used 9/10 of subsamples (comprising randomly selected 81 sampling sites) for model construction and the remaining 1/10 (9 sampling sites) for model validation. This process was repeated until each subsample was used exactly once for validation against the observation. Then root mean squared errors (RMSE) were computed to check the validity of the models and in general, the smaller RMSE value the better model predicted against unseen samples.

A 2-by-2 confusion matrix (also known as the contingency table in other disciplines such as ecology and epidemiology) was then constructed with the four cells (elements) representing number of true positives (TP, corresponding to model correctly predicted high risk site); false positives (FP, model incorrectly predicted high risk site where it should be low risk); false negatives (FN, model incorrectly predicted low risk site where it should be high risk), and true negatives (TN, model correctly predicted low risk site). Model accuracy and its 95% confidence interval was calculated from the confusion matrix to evaluate model performance over all samples. Note that model performance described accuracy across all observations and cross-validation measured accuracy on unseen observations (i.e., predictability). This two-step process evaluated the potential risk of model overfitting (i.e., a model worked well on observed data but failed to make accurate predictions on new dataset).

Some other commonly used machine learning performance evaluation metrics such as sensitivity and specificity were also calculated. Intuitively, sensitivity indicated how well the model could detect true positives (correctly identified high mosquito abundance sites in this study) whereas specificity quantified how well the model did with true negatives (correctly identified low mosquito abundance sites in this study). Another important factor, the F1 score, balancing precision and recall, was also computed. Ideally, a desirable classification model (or the classifier) should have high values (e.g., over 80%) across all these metrics. The equations of these metrics were provided in Supplementary Material S1.

We the used non-metric multidimensional scaling (NMDS) to demonstrate the performance of the optimal model (which had high accuracy as well as low RMSE) on a 2-dimensional space, showing TP, FP, TN, and FN of the 90 sampling sites based on their socioeconomic and/or landscape factors.

Relative importance of socioeconomic and landscape factors on mosquito abundance

We applied random forest (RF), another type of commonly used machine learning model, to explicitly assess the relative importance of continuous input factors on mosquito abundance because RF was designed to differentiate the relative contributions of the inputs (i.e., an input factor was a node in the "forest", Liaw and Wiener 2002). Input factor's relative importance was quantified as the mean decrease of Gini coefficient. The larger the coefficient, the more contribution the factor was to mosquito abundance, and all contributions should sum up to unity. Then each input factor's response curve to mosquito abundance was constructed, conditioned on all other input factors, using RF as well. Here, only continuous inputs were used, and binary inputs were not considered because we wanted to explore the response curve and see whether there existed an "optimal" value of the input factor that was associated with lowest mosquito abundance. From vector control and disease prevention perspective, it was imperative to know how each input factor was related to mosquito abundance. All models were analyzed in *R* 3.5.0 with additional necessary packages (R core team 2019).

Both the original dataset and analysis codes were open to public and freely available upon request.

Results

Association within and between socioeconomic and landscape factors

Distributions of socioeconomic factors and landscape factors (as well as mosquito abundance) are shown in Figs. 1 and 2, respectively. Among socioeconomic factors, PRI and VCR distributions were highly skewed and asymmetric (Fig. 1; note the trapezoids were usually asymmetric). For landscape factors, TRE always had the highest patch coverage, followed by GRS, BLD, and ROD within 30 m radius of sampling sites. The distributions were not as highly skewed as DIV, SHA, and SMP. The distribution of MOS was also highly skewed: the majority of sites had relatively low mosquito abundance while a few had high abundance, in concordance with the 20–80 rule (also known as the Pareto principle, i.e., a few sites having high abundance of mosquitoes). These results showed non-normality of many of the socioeconomic, landscape factors and mosquito abundance which can be problematic for hypothesis-driven models.

Pearson correlations between socioeconomic and landscape factors were calculated and shown in Fig. 3. Within the socioeconomic factor group, POP and DEN, INC and EDU/PRI, EDU and PRI were highly positively associated (using 0.5 as reference value). Within the landscape factor group, TRE was highly negatively associated with BLD/SHA/SMP, whereas both BLD and ROD were highly positively associated with SHA/SMP metrics. Furthermore, DIV, SHA, and SMP were also highly correlated. These high correlations indicated potential collinearity of factors, which impeded the traditional methods (such as GLM) to establish the accurate relations between MOS and the two kinds of independent factors. In addition, no socioeconomic factors had any significant correlation with landscape factors. Such complexity in input data (high collinearity within group of factors but low correlation across groups) posed challenge to GLM but could be solved by data-driven machine learning methods.

Association between socioeconomic and/or landscape factors and mosquito abundance

During the field sampling season in 2017, we collected and identified a total of 3645 *Aedes albopictus* (the Asian tiger mosquito), the dominant species of mosquito in Charlotte. Others were *A. trisariatus* (n = 203), *A. vexans* (n = 41), *A. japonicus* (n = 39), *Culex restuans* (n = 41), and *C. pipiens* (n = 16). Highly degraded samples were sent to the

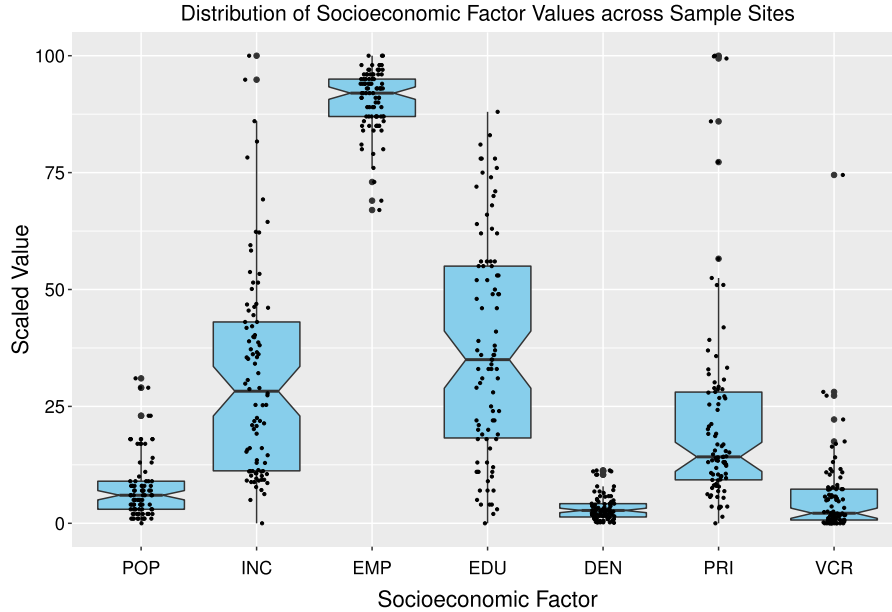

Fig. 1 Scaled distribution of socioeconomic factors across 90 sample sites

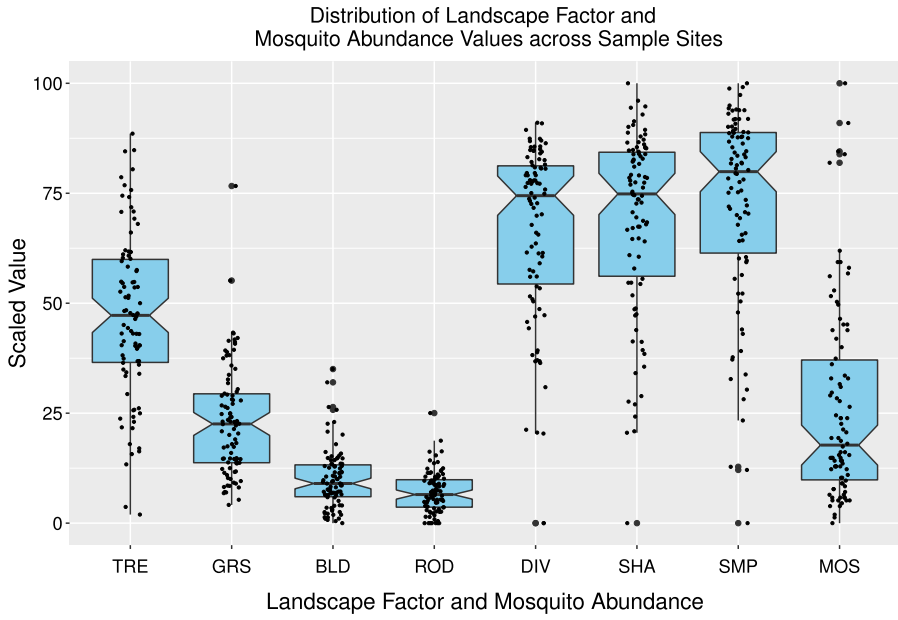

Fig. 2 Scaled distribution of landscape factors and mosquito abundance across 90 sample sites. The horizontal line between two trapezoids indicates mean value for the factor. The more asymmetric the line lie between the two trapezoids, the more skewed the factor distributes (i.e., non-normality). Non-normality is a common challenge to hypothesis-driven model (GLM)

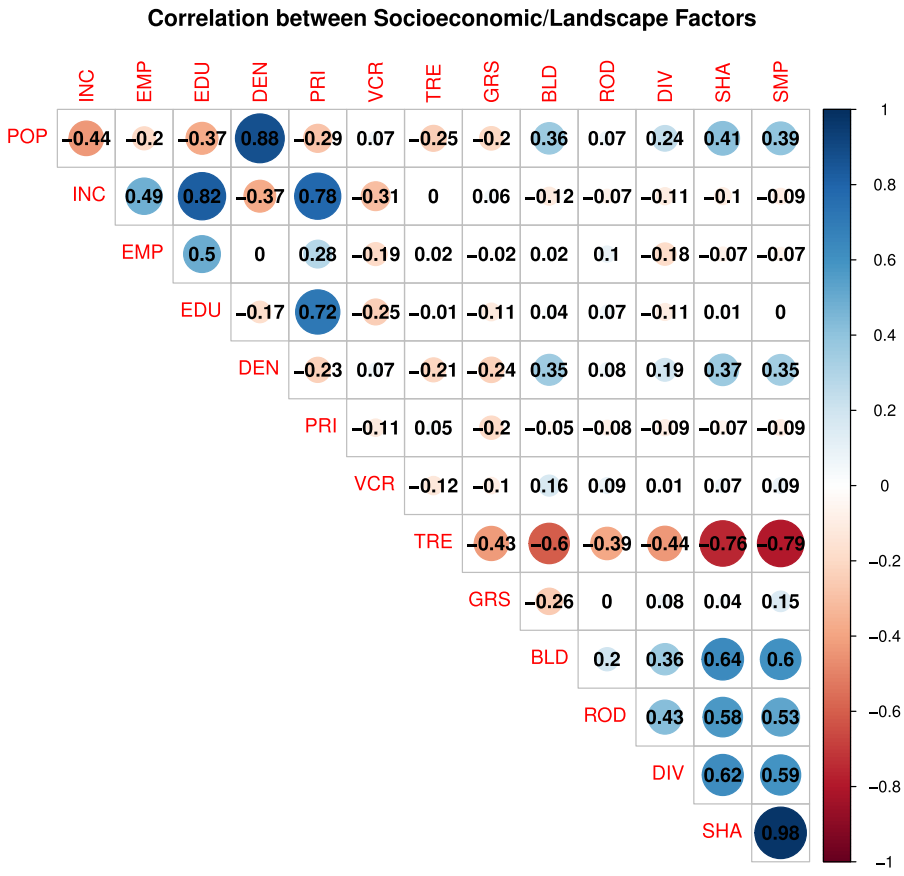

Fig. 3 Correlation between socioeconomic and landscape factors. Using ± 0.50 as reference value, there is no significant correlation between any pair of socioeconomic and landscape factor. However, there are a lot of significant correlations within socioeconomic or landscape factor group, indicating high collinearity within these two groups, which is considered as another potential problem in hypothesis-driven parametric models such as logistic regression

biosystematics unit of the Walter Reed Army Institute of Research for further identification using PCR. *A. aegypti* (the native yellow fever mosquito) was not present in Charlotte, indicating that it had been completely replaced by its invasive counterpart, *A. albopictus* which had first appeared in the city in the 1980s. These *Aedes* vector mosquitoes are responsible for several endemic vector-borne diseases in this region, such as Dengue fever, Eastern equine encephalitis, and La Crosse encephalitis. In addition, they could be potential vectors for traveling-related disease such as Zika to spread locally in this region.

Model performance (measured by model accuracy and 95% confidence interval from each of the confusion matrix associated with specific model/input) was quantified and summarized in Tables 1 and 2, for binary and continuous input, respectively.

For binary input, the highest model accuracy was obtained in kNN (k = 1, accuracy > 95%) using both socioeconomic and landscape factors as inputs, followed by SVM using both factors as inputs (accuracy > 90%), and ANN using both factors as inputs (accuracy > 80%). In general, including landscape factors alone performed equally well or slightly better than socioeconomic factors alone across all models (i.e., green bars were almost always higher than red bars in Fig. 4 in all four panels). The only exception was in model specificity in SVM model (lower right panel in Fig. 4) where landscape factors alone resulted in slightly worse performance than socioeconomic factors alone. Nevertheless, including both groups of factors resulted in substantial increase in model performance in all four metrics across all three machine learning models (i.e., blue bars always higher than either red or green bars in Fig. 4 in all four panels). This finding highlighted the importance of landscape heterogeneity on mosquito abundance. However, it was unexpected using parametric model such as logistic regression, as Pearson correlation didn't detect any significant correlation between socioeconomic and landscape factors groups, hence interaction terms would not be included. There existed some implicit interaction that were critical to mosquito distribution, and such interaction could be identified by machine learning methods, which did not rely on a priori hypotheses.

Comparing across the three machine learning methods, kNN performed the best with regard to model accuracy, F1 value, and sensitivity, followed by SVM and ANN (Fig. 4). This implied that if we wanted to build a machine learning model with high predictability on high mosquito abundance alone (measured by sensitivity) or overall abundance (both high and low mosquito abundance, measured by accuracy and F1 value), kNN (with k = 1) would be the optimal one to start with. For model specificity, kNN was not much ahead of SVM. Nevertheless, from the mosquito population and VBD risk estimation perspective, high mosquito abundance sites should be given higher priority and attention, thus model sensitivity was a more important measurement than specificity. Consequently, we reckoned that kNN was the most appropriate supervised machine learning method to predict mosquito abundance based on both socioeconomic and landscape factors/inputs in Charlotte, with 0–1 binary inputs. Thus, kNN model with socioeconomic and landscape factors as inputs reached optimal predictability (97.7% accuracy) for use in an urban landscape without a need for further fine-tuning.

When continuous input values (only scaled but not ditchotomized) were used, kNN remained to be highly accurate. Using landscape factors alone yielded 100% accurate prediction of mosquito abundance over 90 sampling sites, shadowing already very good 94% accuracy using socioeconomic factors alone.

Table 1 Model accuracy and 95% confidence interval, binary input

Model/input	Socioeconomic	Landscape	Socioeconomic + landscape
kNN	0.78 (0.68–0.86)**	0.79 (0.69–0.87)***	0.98 (0.92–1.00)***
ANN	0.69 (0.58–0.78)***	0.69 (0.58–0.78)***	0.81 (0.71–0.89)***
SVM	0.70 (0.59–0.79)***	0.76 (0.65–0.84)***	0.91 (0.83–0.96)***

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; these p-values were derived from the confusion matrices and not from the actual models directly

Table 2 Model accuracy and 95% confidence interval, continuous input

Model/input	Socioeconomic	Landscape	Socioeconomic + landscape
kNN	0.94 (0.88–0.98)***	1.00 (0.96–1.00)***	1.00 (0.96–1.00)***
ANN	0.66 (0.55–0.75)***	0.66 (0.56–0.75)***	0.67 (0.56–0.76)***
SVM	0.65 (0.54–0.75)***	0.77 (0.67–0.85)***	0.73 (0.63–0.82)***

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; these p-values were derived from the confusion matrices and not from the actual models directly

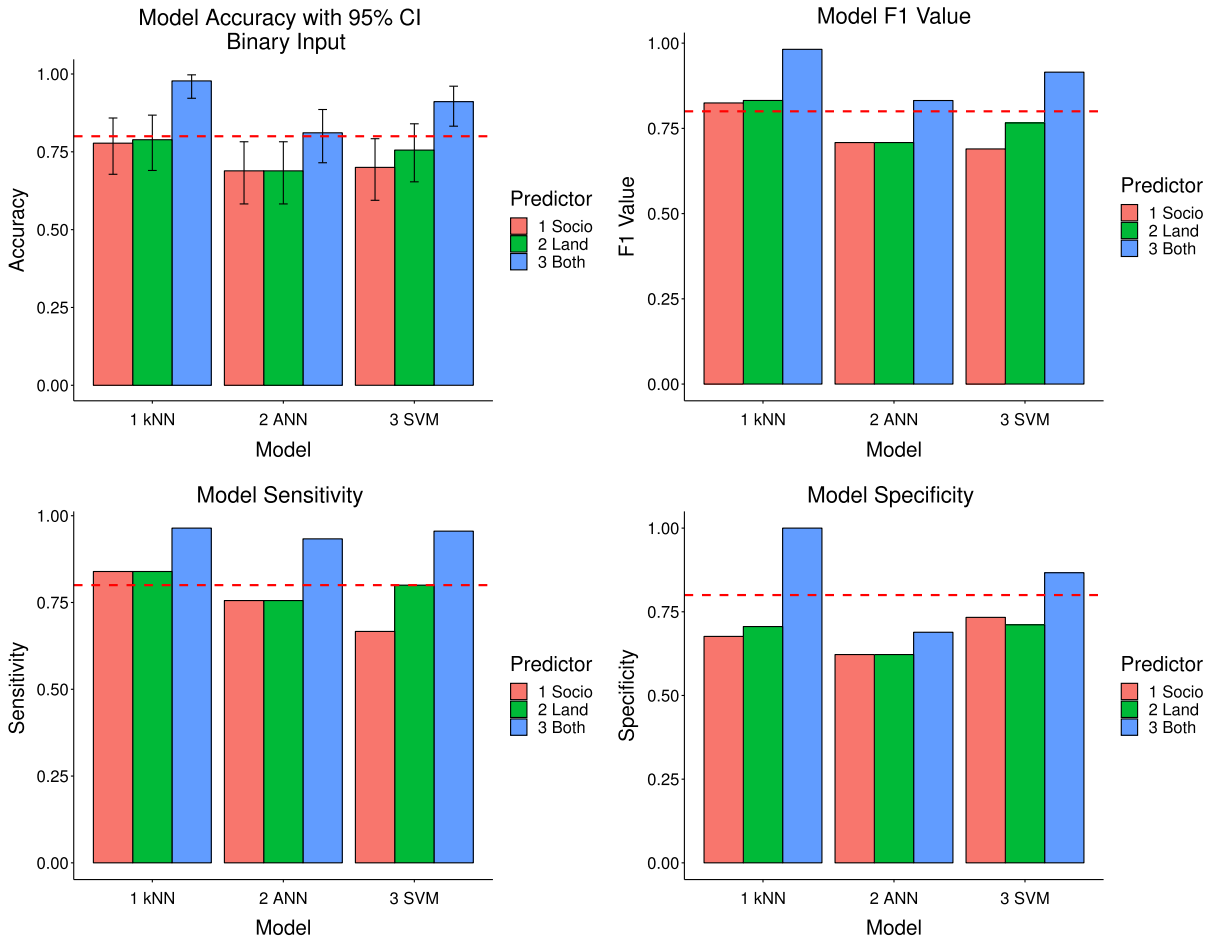

Fig. 4 Comparison of model performance of three machine learning models with different sets of binary input variables. Dashed red line indicates 80% metric value and is chosen to reflect excellent model performance in this study. Having both socioeconomic and landscape factors as input usually yield excellent model performance (accuracy, F1, sensitivity, and specificity) for three machine learning models. (Color figure online)

Landscape factors continued to be an equally well or better group of predictors (input factors) than socioeconomic alone for kNN (100% vs. 94%), ANN (66% vs. 66%), and SVM (77% vs. 65%). It was also worth noting that while kNN performed better using continuous inputs, ANN and SVM actually had slightly less optimal performance when fed with continuous inputs.

Since socioeconomic and landscape factors combined showed the highest model accuracy, we

evaluated four models' predictability based on the two factors combined as input with 10-fold cross-validation. RMSE of these models with binary inputs were 0.57 ± 0.13, 0.69 ± 0.07, and 0.67 ± 0.14 (mean ± S.D.) for kNN, ANN, and SVM, respectively, using binary inputs; and 0.67 ± 0.10, 0.72 ± 0.11, and 0.67 ± 0.13 for kNN, ANN, and SVM, respectively, using continuous inputs.

Based on these results, kNN not only had the highest model accuracy using either continuous or binary input type, but also performed the best in cross-validation with the smallest RMSE value among all three machine learning methods. Consequently, kNN's predictability on new dataset was the optimal among the three models as well. Practically, when tested on another year of data captured in the future or at another location, kNN should perform consistently well on the new dataset (with fine-tuning of model parameters). Predictability of the other three types of model did not differ much based on their RMSE values.

Individual input factor's contribution and relationship to mosquito abundance

We showed individual input factor's contribution (relative importance) to mosquito abundance in Fig. 6, derived from the mean decrease in Gini coefficient using RF model. In general, landscape factors (labeled in red) were of more importance than socioeconomic factors. Seven landscape factors had a total of 54.4% contribution on mosquito abundance, while seven socioeconomic factors had only 45.6%. This finding was consistent from the previous section that landscape factors alone yielded equally well or better performance in predicting mosquito abundance than socioeconomic factors alone (Figs. 4 and 5). We suggested that the reason was landscape factors were actually an interplay between host socioeconomic development and ambient environment, hence landscape factors should include some information regarding socioeconomic factors. Besides, micro-scale landscape heterogeneity (the Shannon–Wiener diversity index, SHA), was the single most influential factor (9.56%) for mosquito abundance and was far ahead of the second most influential factor (violent crime rate, 7.77%). Thus, we confirmed again that landscape factors, especially landscape heterogeneity, substantially influenced mosquito abundance at their micro-scale landscape.

Individual input factor's response curves (conditioned on all other input factors) were provided in Fig. 7. In general, negative contribution meant that mosquito abundance was negatively impacted. For instance, SHA, the most influential factor, had an optimal value (around 65–70%) that minimized the contribution (i.e., the "saddle" portion on SHA plot panel of Fig. 7). This indicated that at around 65–70% SHA, mosquito population size should be minimal. However, larger ($> 70\%$ SHA) or smaller (especially $< 40\%$ SHA) were both associated with higher mosquito abundance. Therefore, larger landscape diversity did not necessarily result in lower mosquito abundance. Similarly, GRS (grass coverage) also had a saddle point (Fig. 7) around 40% GRS, indicating that an optimum of 40% GRS coverage was associated with the least mosquito abundance in Charlotte-Mecklenburg County. All these results shed light to more appropriate landscape design and urban planning with regard to potential mosquito infestation.

In conclusion, we showed that micro-scale landscape factors, especially landscape heterogeneity, was the key to accurately predict mosquito abundance. The non-normality, high within-group collinearity, and low between-group correlation in observed data all pose technical challenges to derive accurate model using conventional hypothesis-driven parametric methods. Supervised machine learning models, especially kNN, were able to provide robust and accurate predictions of mosquito abundance across socioeconomic gradients in heterogeneous urban landscapes such as in Charlotte. These insights, such as the nonlinear responses of mosquito abundance to landscape and socioeconomic factors, can be used to guide landscape design and urban planning in large metropolitan areas.

Discussion

In this study, we have demonstrated that landscape factors were important to mosquito abundance and was further intermingled with socioeconomic disparity. These associations could be detected and quantified by machine learning, especially supervised learning models. These models provide powerful tools that can identify potential hidden connections inside

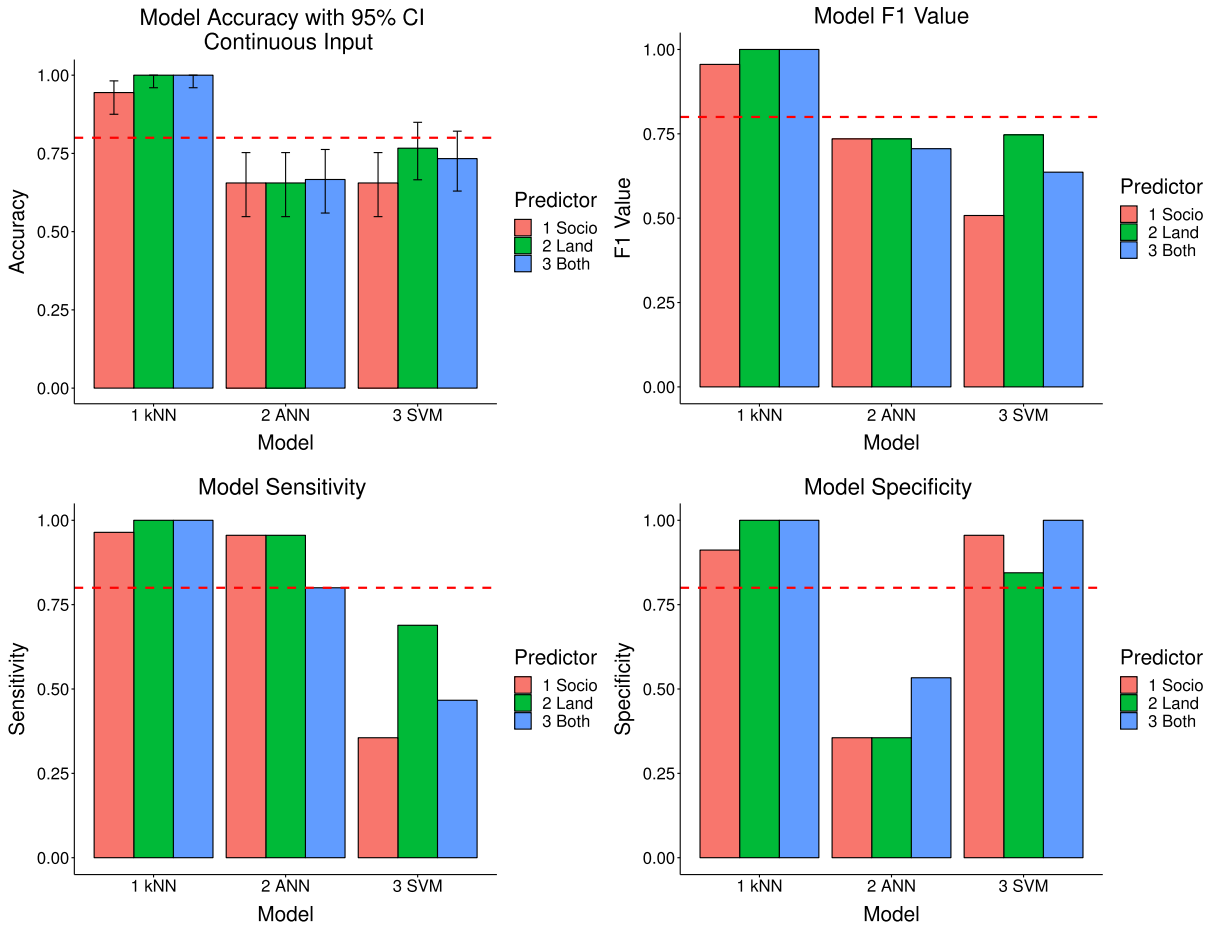

Fig. 5 Comparison of model performance of three machine learning models with different sets of continuous input variables. Dashed red line indicates 80% metric value and is chosen to reflect excellent model performance in this study. (Color figure online)

complicated datasets that are usually undetectable by traditional methods. For instance, there is no significant correlation between socioeconomic and landscape factors (i.e., Pearson correlation coefficient > 0.5), and our previous study using logistic regression (continuous response, Whiteman et al. 2018) showed no model improvement by adding landscape factors atop socioeconomic factors. However, these two groups of factors are synergistic with regard to mosquito abundance, as revealed by all three types of machine learning models used in this study (kNN, ANN, and SVM, comparing model accuracy among three sets of input factors).

Furthermore, these supervised learning models not only help identify hidden interactions between different factors, but also provide high accuracy especially using kNN. In addition, our proposed models use socioeconomic and landscape factors as inputs, which are not as dynamic as other commonly-used environmental factors such as temperature and humidity reducing the costs of monitoring and data collection. Thus, machine learning models are able to accurately estimate and predict mosquito abundance across broad socioeconomic gradients and heterogeneous landscapes, especially in an urban area, and they can provide integrated evidence for public awareness of VBD risk and decision-making around VBD control efforts. In this study we have evaluated feasibility of machine learning models, and they can be readily applied to predict mosquito abundance in other locations, too. While the model has used 90 sampling sites in Charlotte, we demonstrate its ability to predict mosquito abundance in other locations as long as their socioeconomic and landscape factors are quantified.

Further insights could be gained on which factor combinations would yield high mosquito abundance, thus helping design more effective mosquito/VBD surveillance and control programs across Charlotte.

Among all three supervised learning models that we have built, ANN is the most explicit and straightforward. It has an explicit model structure and resembles a structure equation model. However, a structure equation model usually does not involve a hidden node or a hidden layer and it is still considered a parametric method. The actual biological interpretation of these hidden layers and nodes are generally unclear and hidden layer/nodes are assigned ad hoc. Machine learning models are data-driven, often much more complicated to build and test, and they rely on specific codes with fine-tuned parameters (also known as hyperparameters in machine learning field, e.g., k value in kNN model, number of layers and nodes in ANN, and kernel specification in SVM) that require experience to operate.

From a practical mosquito population/VBD surveillance perspective, it is more important for public health officers and the general public to know where high mosquito abundance will be (i.e., "hot-spot") in a heterogeneous landscape (Fournet et al. 2018), rather than knowing the exact population size, which might be less accurate secondary to over-fitting. Thus, our modeling framework was able to deal with different needs and input data quality (continuous or binary, representing more accurate but hard-to-collect surveillance or easier to retrieve relative information).

We suggest that machine learning (data-driven method) is not meant to completely replace traditional hypothesis-driven methods, just like frequentist and Bayesian methods should be synergistic and not mutually exclusive. For example, while this study has focused on machine learning methods, model accuracy is actually evaluated with the hypothesis-driven method (e.g., calculations from confusion matrix). While machine learning methods can provide additional insights (and sometimes more intensive scrutiny) of the data, they rely on researchers to make the judgement regarding their validity and interpretability. We should not treat machine learning methods as a self-contained autonomous and automatic data processor, especially in the field of landscape ecology where many aspects of model implementation need careful calibrations. Instead, we recommend using machine learning in conjunction with hypothesis-driven method. For instance, in this study we have computed individual factor's response to mosquito abundance based on machine learning method (Fig. 6), and specific hypotheses can be proposed and tested to further investigate such non-linear responses.

Future direction for this study will be applying the proposed machine learning models across multiple years and in different locations. We have

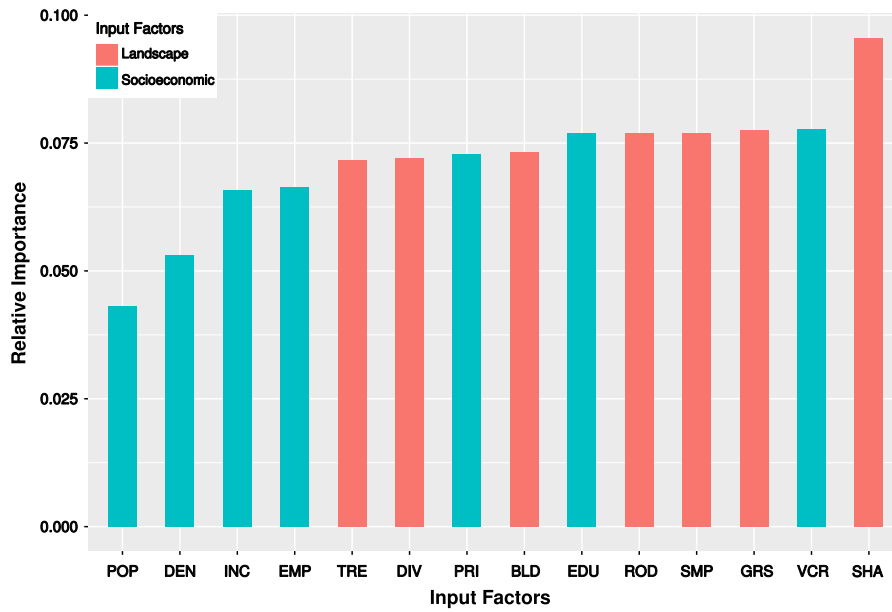

Fig. 6 Individual input factor's contribution to mosquito abundance. SHA (Shannon diversity index) indicates the landscape (landcover) type heterogeneity at micro-scale landscape (in 30 m radius of sampling site), it does not mean diversity of mosquitoes

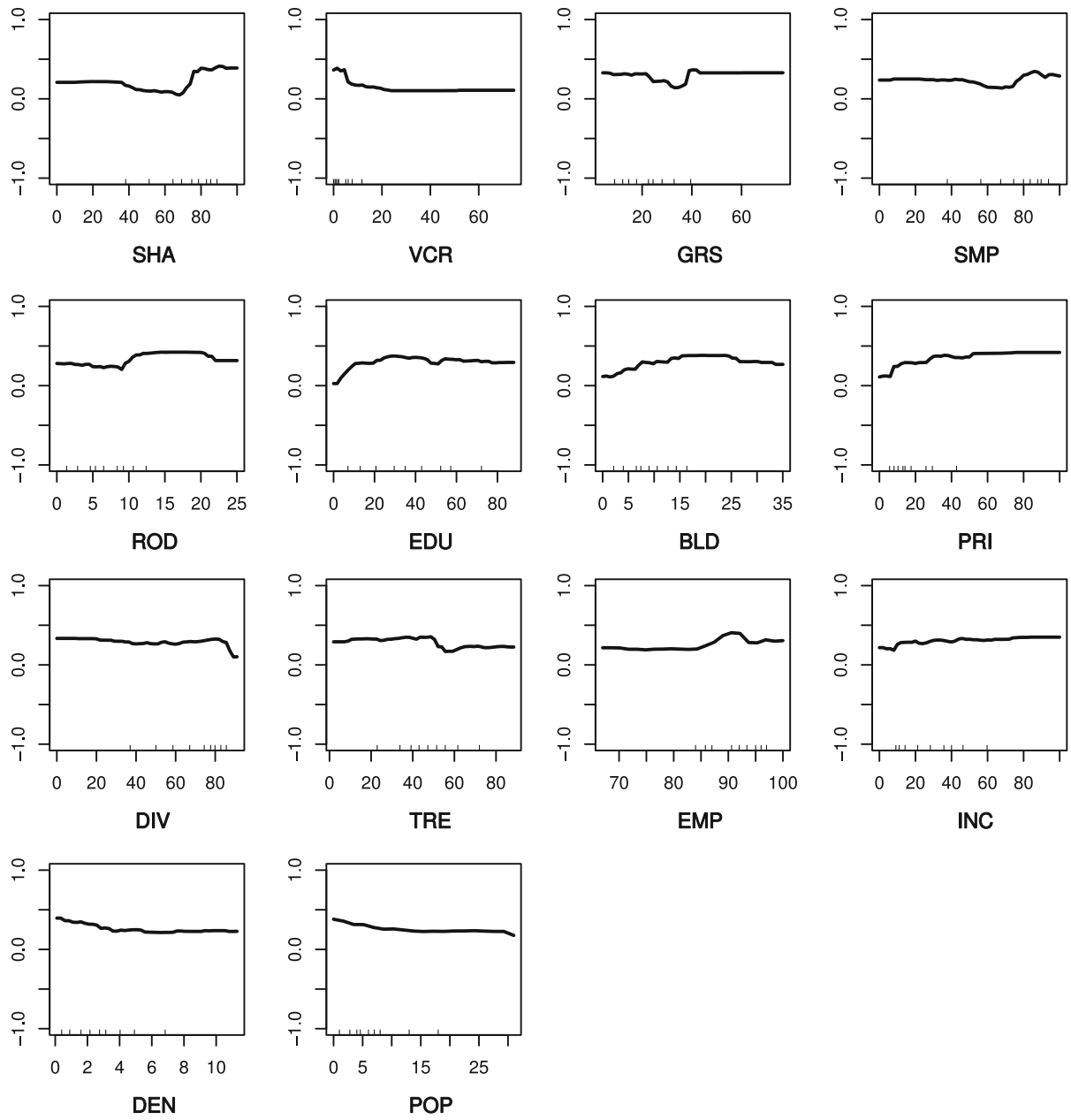

Fig. 7 Individual input factor's response curve to mosquito abundance. The order of the input factor reflects its relative importance as shown in Fig. 5. Positive *Y*-axis value means positively associated with mosquito abundance (i.e., increasing mosquito abundance) while negative *Y*-axis value indicates negative association (more relevant to mosquito control)

demonstrated the effectiveness of machine learning models on predicting mosquito abundance, but their robustness is yet to be validated. Although kNN achieves the optimal model performance in this study with both socioeconomic and landscape factors inputs, other models with different set of input might outperform kNN based on new observed dataset. Besides, environmental factors, especially temperature fluctuation within micro-climate (at spatial resolution relevant to mosquito activity, Chen et al. 2013, 2015) and host-level interventions such as hygiene (Degroote et al. 2018), could also be incorporated to complete the epidemiology triad and model potential VBD risk. The ultimate goal is to work

toward a more effective surveillance system for VBD risk especially in urban areas (Eder et al. 2018; Fournet et al. 2018). We will also investigate and understand the detailed mechanism of individual factor's response to mosquito abundance, and propose corresponding vector-control strategies.

Acknowledgement We thank Mecklenburg County Health Department for providing funding for the mosquito sampling work in 2017 and access to the orthophotos. We are also grateful for the field work of ten volunteering undergraduate students from UNC Charlotte.

References

Adjei PO-W, Kyei PO (2013) Linkages between income, housing quality and disease occurrence in rural Ghana. J Hous Built Environ 28(1):35–49

Adler NE, Ostrove Joan M (1999) Socioeconomic status and health: what we know and what we don't. Ann N Y Acad Sci 896:3–15

Baltensperger AP, Huettmann F (2015) Predictive spatial niche and biodiversity hotspot models for small mammal communities in Alaska: applying machine-learning to conservation planning. Landscape Ecol 30:681–697

Benedict MQ, Levine RS, Hawley WA, Lounibos LP (2007) Spread of the tiger: global risk of invasion by the mosquito *Aedes albopictus*. Vector Borne Zoonotic Dis 7(1):76–85

Boone CG, Cook E, Hall SJ et al (2012) A comparative gradient approach as a tool for understanding and managing urban ecosystems. Urban Ecosyst 15(4):795–807

Brown ME, Grace K, Shively G, Johnson KB, Carroll M (2014) Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. Popul Environ 36(1):48–72

Buckner EA, Blackmore MS, Golladay SW, Covich AP (2011) Weather and landscape factors associated with adult mosquito abundance in southwestern Georgia, U.S.A. J Vector Ecol 36(2):269–278

Burger SV (2018) Introduction to machine learning with R. O'Reilly, Sebastopol

Chaves LF, Hamer GL, Walker ED, Brown WM, Ruiz MO, Kitron UD (2011) Climatic variability and landscape heterogeneity impact urban mosquito diversity and vector abundance and infection. Ecosphere 2(6):70

Chen S, Blanford JI, Fleischer SJ, Hutchinson M, Saunders MC, Thomas MB (2013) Estimating West Nile virus transmission period in pennsylvania using an optimized degree-day model. Vector Borne Zoonotic Dis 13(7):489–497

Chen S, Fleischer SJ, Saunders MC, Thomas MB (2015) The influence of diurnal temperature variation on degree-day accumulation and insect life history. PLoS ONE 10(3):1–15

Chetty R, Hendren N, Kline P, Saez E (2014) Where is the land of Opportunity? The geography of intergenerational mobility in the united states. Q J Econ 129(4):1553–1623

Cianci D, Hartemink N, Ibáñez-Justicia A (2015) Modelling the potential spatial distribution of mosquito species using three different techniques. Int J Health Geogr 14(1):10

Cushman SA, Heuttmann F (2010) Spatial complexity, informatics, and wildlife conservation. Springer, Tokyo

Degroote S, Bermudez-Tamayo C, Ridde V (2018) Approach to identifying research gaps on vector-borne and other infectious diseases of poverty in urban settings: scoping review protocol from the VERDAS consortium and reflections on the project's implementation. Infect Dis Poverty 7(1):98

Delmelle E, Hagenlocher M, Kienberger S, Casas I (2016) A spatial model of socioeconomic and environmental determinants of dengue fever in Cali, Colombia. Acta Trop 164:169–176

Dowd JB, Zajacova A, Aiello A (2009) Early origins of health disparities: burden of infection, health, and socioeconomic status in U.S. children. Soc Sci Med 68(4):699–707

Dowling Z, Armbruster P, LaDeau SL, DeCotiis M, Mottley J, Leisnham PT (2013) Linking mosquito infestation to resident socioeconomic status, knowledge, and source reduction practices in suburban Washington, DC. EcoHealth 10(1):36–47

Drew CA, Wiersma Y, Heuttmann F (2011) Predictive species and habitat modeling in landscape ecology. Springer, New York

Eder M, Cortes F, de Siqueira Teixeira, Filha N et al (2018) Scoping review on vector-borne diseases in urban areas: transmission dynamics, vectorial capacity and co-infection. Infect Dis Poverty 7(1):90

Fielding AH (1999) Machine learning methods for ecological applications. Springer, Berlin

Foley DH, Wilkerson RC, Rueda LM (2009) Importance of the "what", "when", and "where" of mosquito collection events. J Med Entomol 46(4):717–722

Fournet F, Jourdain F, Bonnet E, Degroote S, Ridde V (2018) Effective surveillance systems for vector-borne diseases in urban settings and translation of the data into action: a scoping review. Infect Dis Poverty 7(1):99

Gordis L (2013) Epidemiology, 5th edn. Elsevier, Canada

Gottdenker NL, Streicker DG, Faust CL, Carroll CR (2014) Anthropogenic land use change and infectious diseases: a review of the evidence. EcoHealth 11(4):619–632

Gunther F, Fritsch S (2010) Neuralnet: training of neural networks. R Journal 2(1):30–38

Hall V, Walker WL, Lindsey NP, Lehman JA, Kolsin J, Dandry K, Rabe IB, Hills SL, Fischer M, Staples JE, Gould CV, Martin SW (2018) Update: noncongential Zika virus disease cases — 50 U.S. States and the District of Columbia, 2016. Morb Moral Wkly Rep 67(9):265–269

Hawe P, Shiell A (2000) Social capital and health promotion: a review. Soc Sci Med 51(6):871–885

Hayden MH, Uejio CK, Walker K et al (2010) Microclimate and human factors in the divergent ecology of Aedes aegypti along the Arizona, U.S./Sonora, MX Border. EcoHealth 7(1):64–77

Hemme RR, Thomas CL, Chadee DD, Severson DW (2010) Influence of urban landscapes on population dynamics in a short-distance migrant mosquito: evidence for the dengue vector *Aedes aegypti*. PLOS Negl Trop Dis 4(3):1–9

Homan T, Maire N, Hiscox A et al (2016) Spatially variable risk factors for malaria in a geographically heterogeneous landscape, western Kenya: an explorative study. Malar J 15(1):1

Humphries G, Magness DR, Huettmann F (2018) Machine learning for ecology and sustainable natural resource management. Springer, Switzerland

Jaeger JAG (2000) Landscape division, splitting index, and effective mesh size: new measures of landscape fragmentation. Landscape Ecol 15(2):115–130

Johnson MF, Gómez A, Pinedo-Vasquez M (2008) Land use and mosquito diversity in the Peruvian amazon. J Med Entomol 45(6):1023–1030

Jopp F, Reuter H, Brecklings B (2011) Modelling complex ecological dynamics. Springer, Berlin

Kabaria CW, Molteni F, Mandike R et al (2016) Mapping intra-urban malaria risk using high resolution satellite imagery: a case study of Dares Salaam. Int J Health Geogr 15(1):26

Kalluri S, Gilruth P, Rogers D, Szczur M (2007) Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. PLoS Pathog 3(10):1–11

Karatzoglou A, Meyer D, Hornik K (2006) Support vector machines in R. J Stat Softw 15:1–28

Keating J, Macintyre K, Mbogo C et al (2003) A geographic sampling strategy for studying relationships between human activity and malaria vectors in Urban Africa. Am J Trop Med Hyg 68(3):357–365

Khormi HM, Kumar L (2011) Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study. Sci Total Environ 409(22):4713–4719

Kikuti M, Cunha GM, Paploski IAD et al (2015) Spatial distribution of dengue in a brazilian urban slum setting: role of socioeconomic gradient in disease risk. PLOS Negl Trop Dis 9(7):1–18

LaDeau SL, Leisnham PT, Biehler D, Bodner D (2013) Higher mosquito production in low-income neighborhoods of Baltimore and Washington, DC: understanding ecological drivers and mosquito-borne disease risk in temperate cities. Int J Environ Res Public Health 10(4):1505–1526

Lana RM, Riback TIS, Lima TFM et al (2017) Socioeconomic and demographic characterization of an endemic malaria region in Brazil by multiple correspondence analysis. Malar J 16(1):397

Lantz B (2015) Machine learning with R, 2nd edn. Packt Publishing, Birmingham

Lary DJ, Woof S, Faruque F et al (2014) Holistics 3.0 for health. Int J Geoinf 3:1023–1038

Le Comber SC, Rossmo D, Hassan AN, Fuller DO, Beier JC (2011) Geographic profiling as a novel spatial tool for targeting infectious disease control. Int J Health Geogr 10(1):35

Leigh JP (1993) Multidisciplinary findings on socioeconomic status and health. Am J Public Health 83:289–290

Leisnham PT, Juliano SA (2012) Impacts of climate, land use, and biological invasion on the ecology of immature *Aedes* mosquitoes: implications for La Crosse emergence. EcoHealth 9(2):217–228

Lesmeister C (2015) Mastering machine learning with R. Packt Publishing, Birmingham

Li Y, Kamara F, Zhou G et al (2014) Urbanization increases *Aedes albopictus* larval habitats and accelerates mosquito development and survivorship. PLOS Negl Trop Dis 8(11):1–12

Liaw A, Wiener M (2002) Classification and regression by randomForest. RNews 2(3):18–22

Linard C, Ponçon N, Fontenille D, Lambin EF (2009) Risk of malaria reemergence in southern France: testing scenarios with a multiagent simulation model. EcoHealth 6(1):135

Little E, Biehler D, Leisnham PT, Jordan R, Wilson S, LaDeau SL (2017) Socio-ecological mechanisms supporting high densities of *Aedes albopictus* (Diptera: Culicidae) in Baltimore, MD. J Med Entomol 54(5):1183–1192

McGarigal K, Cushman, SA, Ene E (2012) Fragstats v4: spatial pattern analysis program for categorical and continous maps. http://www.umass.edu/landeco/research/fragstats/fragstats.html

Monaghan AJ, Sampson KM, Steinhoff DF et al (2018) The potential impacts of 21st century climatic and population changes on human exposure to the virus vector mosquito Aedes aegypti. Clim Change 146(3):487–500

Norris DE (2004) Mosquito-borne diseases as a consequence of land use change. EcoHealth 1(1):19–24

Obenauer JF, Andrew JT, Harris JB (2017) The importance of human population characteristics in modeling *Aedes aegypti* distributions and assessing risk of mosquito-borne infectious diseases. Trop Med Health 45(1):38

Osorio L, Garcia JA, Parra LG et al (2018) A scoping review on the field validation and implementation of rapid diagnostic tests for vector-borne and other infectious diseases of poverty in urban areas. Infect Dis Poverty 7(1):87

Ozdenerol E, Bialkowska-Jelinska E, Taff GN (2008) Locating suitable habitats for West Nile Virus-infected mosquitoes through association of environmental characteristics with infected mosquito locations: a case study in Shelby County, Tennessee. Int J Health Geogr 7(1):12

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rael RC, Peterson AC, Ghersi BM, Childs J, Blum MJ (2016) Disturbance, reassembly, and disease risk in socioecological systems. EcoHealth 13(3):450–455

Reiner RC, Perkins TA, Barker CM et al (2013) A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J R Soc Interface 10(81):20120921

Robertson C (2017) Towards a geocomputational landscape epidemiology: surveillance, modelling, and interventions. GeoJournal 82(2):397–414

Roiz D, Ruiz S, Soriguer R, Figuerola J (2015) Landscape effects on the presence, abundance and diversity of mosquitoes in mediterranean wetlands. PLoS ONE 10(6):1–17

Rosenberg R, Lindsey NP, Fischer M et al (2018) Vital signs: trends in reported vectorborne disease cases -United States and territories, 2004-2016. Morb Mortal Wkly Rep 67(17):496–501

Ruiz MO, Walker ED, Foster ES, Haramis LD, Kitron UD (2007) Association of West Nile virus illness and urban landscapes in Chicago and Detroit. Int J Health Geogr 6(1):10

Ruiz-Moreno D (2016) Assessing Chikungunya risk in a metropolitan area of Argentina through satellite images and mathematical models. BMC Infect Dis 16(1):49

Rydin Y, Bleahu A, Davies M et al (2012) Shaping cities for health: complexity and the planning of urban environments in the 21st century. Lancet 379:2079–2108

Shao G, Wu J (2008) On the accuracy of landscape pattern analysis using remote sensing data. Landscape Ecol 23:505–511

The World Bank (2015) GINI index (world bank estimate). The World Bank: 1–16. http://data.worldbank.org/indicator/SI.POV.GINI

Townsend AT (2006) Ecological niche modeling and spatial patterns of disease transmission. Emerg Infect Dis 12(12):1822–1826

Townsend AT, Vieglais DA (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem: a new approach to ecological nich modeling, based on new tolls drawn from biodiversity informatics, is applied to the challenge of predicting potential species' invasions. Bioscience 51(5):363–371

Tusting LS, Rek J, Arinaitwe E et al (2016) Why is malaria associated with poverty? Findings from a cohort study in rural Uganda. Infect Dis Poverty 5(1):78

Unlu I, Farajollahi A, Healy SP et al (2011) Area-wide management of *Aedes albopictus*: choice of study sites based on geospatial characteristics, socioeconomic factors and mosquito populations. Pest Manag Sci 67(8):965–974

Whiteman A, Delmelle E, Rapp T, Chen S, Chen G, Dulin M (2018) A novel sampling method to measure socio-ecological drivers of *Aedes albopictus* distribution in Charlotte, NC. Int J Environ Res Public Health 15(10):2179

Winkleby MA, Jatulis DE, Frank E, Fortmann SP (1992) Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. Am J Public Health 82:816–820

World Health Organization (2018) Vector-borne disease. http://www.who.int/heli/risks/vectors/vector/en/. Accessed 20 Oct 2018

World Health Organization (2018) Handbook for integrated vector management. http://apps.who.int/iris/bitstream/handle/10665/44768/9789241502801_eng.pdf. Accessed 20 Oct 2018

Wu J (2014) Urban ecology and sustainability: the state-of-the-science and future directions. Landsc Urban Plan 125:209–221

Wu J, Jenerette GD, Buyantuyev A, Redman CL (2011) Quantifying spatiotemporal patterns of urbanization: the case of the two fastest growing metropolitan regions in the United States. Ecol Complex 8:1–8

Young BD, Yarie J, Verbyla D, Huettmann F, Chapin FS (2018) Machine learning for ecology and sustainable natural resource management. Springer Nature, Switzerland

Young BD, Yarie J, Verbyla D, Huttmann F, Herrick K, Chapin FS (2017) Modeling and mapping forest diversity in the boreal forest of interior Alaska. Landscape Ecol 32:397

Younsi M, Chakroun M (2016) Does social capital determine health? Empirical evidence from MENA countries. Soc Sci J 53(3):371–379