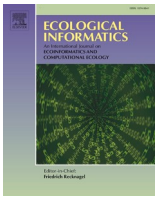
Ecological Informatics 81 (2024) 102610

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Towards accurate spatial prediction of *Glossina pallidipes* relative densities at country-scale in Kenya

Stella Gachoki [a,b,*], Thomas A. Groen [b], Anton Vrieling [b], Andrew Skidmore [b], Daniel Masiga [a]

[a] International Centre of Insect Physiology and Ecology (icipe), P.O. Box 30772-00100, Nairobi, Kenya
[b] University of Twente, Faculty of Geo-information Science and Earth Observation (ITC), P.O. Box 217, 7500 AE Enschede, the Netherlands

ARTICLE INFO

Keywords:
Tsetse abundance
Machine learning
Vector borne diseases
Spatial extrapolations
Satellite data
Random forest

ABSTRACT

Vector-borne diseases, like those transmitted by tsetse flies, pose a significant global public health threat. Reducing vector populations is a promising strategy for disease control, especially in the case of tsetse-transmitted African trypanosomiasis. However, the cost-effective implementation of large-scale vector surveillance and control measures face challenges due to the lack of spatially explicit and reliable maps identifying vector hotspots. In this study, we assessed the accuracy of predicting *Glossina pallidipes* relative densities across Kenya by linking constrained in-situ tsetse catch data from 660 traps across three Kenyan regions with readily available gridded satellite information (human population, land cover, soil properties, elevation, precipitation, and land surface temperature) using a classical random forest algorithm. To enhance predictive performance, we employed two feature elimination techniques specifically designed for machine learning algorithms, i.e., Recursive Feature Elimination (RFE) and Variable Selection Using Random Forests (VSURF). For each set of retained variables, we trained a Random Forest model using a spatial cross-validation technique. Our findings showed that tsetse fly relative densities decreased with mean annual precipitation, and soil moisture, and conversely increased with higher tree cover. Based on the cross-validated R^2, 41% of the spatial variability in relative densities of tsetse flies could be explained. For spatial extrapolation, only the set of predictors retained by VSURF closely matched known tsetse fly distributions in Kenya. This more accurate performance of VSURF may be attributed to its approach of assessing variables for both importance and their contribution to reducing prediction error. Our study demonstrates the potential of using a random forest method to upscale tsetse relative abundance predictions to the national level. However, the reliability of the current extrapolated map remains uncertain. We recommend: 1) increasing tsetse fly sampling efforts, particularly in the data-limited northern and eastern regions of Kenya, and 2) developing a more precise and accurate land cover map with classes that directly associate with known habitat characteristics of the target tsetse species.

1. Introduction

Vector-borne diseases (VBDs) pose a significant and pervasive threat to public health worldwide, affecting both humans and livestock. These diseases are transmitted through the bites of various vectors, predominantly insects. According to World Health Organization (WHO), the impact of most VBDs can be mitigated through the proper implementation of various vector control strategies (World Health Assembly, 2017). Quantifying the distribution and relative abundance of disease vectors plays a pivotal role in guiding decision-making processes and executing timely and efficient control measures. While research has made reasonable progress in understanding the environmental factors that shape the geographic distribution of various disease vectors and has extensively advanced in producing spatial and temporal maps of their suitable habitats, only a few studies attempted to assess the spatial relative abundance of disease vectors (Waldock et al., 2022). Understanding the spatial variation in disease vector populations is crucial, as these numbers serve as indicators of disease risk and persistence.

Conducting large-scale, in-situ monitoring of disease vectors to understand their spatial and temporal relative abundance and identify areas of potential risk of disease and consequently identify priority areas for control is often impractical. Remote sensing technology enables the mapping of environmental and weather information over extensive regions, and remote sensing derived datasets have become readily

* Corresponding author at: International Centre of Insect Physiology and Ecology (*icipe*), P.O. Box 30772-00100, Nairobi, Kenya.
E-mail address: s.m.gachoki@utwente.nl (S. Gachoki).

https://doi.org/10.1016/j.ecoinf.2024.102610
Received 16 November 2023; Received in revised form 18 April 2024; Accepted 20 April 2024
Available online 22 April 2024

available over the past years. Despite challenges related to integrating in-situ data with satellite-based estimates, such as data quality and model complexities, the improved integration of geospatial and remote sensing expertise into VBD control programs has notably advanced global research on the environmental and weather factors that affect disease vector abundance (Carrasco-Escobar et al., 2022; Dlamini et al., 2019; Kalluri et al., 2007; Mechan et al., 2023; Palaniyandi et al., 2021). In addition, the rapid evolution of technology has led to the development of increasingly sophisticated predictive modeling methodologies, particularly through the utilization of machine learning (Kaur et al., 2021, 2022; Keshavamurthy et al., 2022; Yu et al., 2022), offering an advantage over traditional geostatistical methods by facilitating the capturing of intricate non-linear interactions among variables (Taconet et al., 2021).

Among the disease vectors, studies aiming to map and understand their spatial relative densities have focused mostly on mosquitoes and ticks. In addition to the commonly employed classical geostatistical models (Mudele et al., 2021; Rosà et al., 2019; Shutt et al., 2022; Talbot et al., 2019), these studies have increasingly turned to machine-learning technologies (González Jiménez et al., 2019; Joshi and Miller, 2021; Makridou et al., 2023; Rahman et al., 2021; Schneider et al., 2022). A significant advantage of machine learning over traditional geostatistical methods is its innate capacity to capture complex associations among variables, often resulting in higher predictive accuracies. For instance, Ibañez-Justicia and Cianci (2015) showed more accurate performance of random forest over other models in predicting mosquito abundance, a finding that was reiterated by Rahman et al. (2021) when predicting the abundance of *Aedes aegypti* female mosquitoes in Thailand. For tick abundance in Southern Scandinavia, Jung Kjær et al. (2019) employed Boosted Regression Trees (BRT) with gridded environmental and weather variables. Their approach yielded higher accuracy for tick larvae and nymphs (R^2 of 0.69) but less accurate results for adult ticks (R^2 of 0.1). Ceia-Hasse et al. (2023) demonstrated an improved performance of deep learning over classical machine learning (area under curve values 0.83 and 0.75 respectively) in predicting the abundance of yellow fever mosquitoes in Madeira, Portugal. These findings collectively highlight the promising potential of various machine learning techniques in developing high-performing models for assessing disease vector abundance.

This study focuses on tsetse flies. While other insects can mechanically transmit trypanosome pathogens (Desquesnes and Dia, 2003; Mihok et al., 1995), tsetse flies are the sole biological vectors of trypanosome pathogens causing African Trypanosomiasis in humans and livestock across Sub-Saharan Africa and by far the most significant vectors. While environmental factors influencing tsetse fly distribution are well-documented (Bishop et al., 2021; De Beer et al., 2021; Gachoki et al., 2021) based on the pioneering work by Rogers and Randolph (1986), only a few studies assessed spatial variations in tsetse numbers and these are limited to the use of standard geostatistical models. For instance, Lord et al. (2018) used generalized linear models (GLMs) and satellite data to forecast tsetse numbers inside and outside Serengeti National Park in Tanzania. However, their model revealed varying prediction accuracy due to a temporal mismatch between tsetse data collection (2010 and 2015) and satellite predictors used (2015). Mugenyi et al. (2021) employed standard Poisson and zero-inflated Poisson models to predict tsetse numbers per trap per day in Uganda. This study found that during dry seasons high tsetse numbers were concentrated in low-lying areas, animal reserves, wooded landscapes, and shrub-covered regions. However, the models failed to capture tsetse abundance patterns in the wet season, which was associated with an increased dispersal rate of tsetse flies during this period.

Among the many machine learning techniques, here we selected the classical random forest method due to its well-documented success in predicting the abundance of other disease vectors, such as mosquitoes. To enhance the precision of our method, we implemented two feature elimination techniques namely, Recursive Feature Elimination (RFE; Darst et al., 2018; Khun, 2022) and Variable Selection Using Random Forests (VSURF; Genuer et al., 2022; Speiser, 2021; Speiser et al., 2019). RFE retains variables based on their importance in explaining the response variable (Khun, 2022), while the VSURF method goes a step further by assessing how these important variables contribute to predicting the response variable and retains only those that lead to a reduction in prediction error (Genuer et al., 2022). To the best of our knowledge, there have been no previous attempts to develop random forest models for predicting relative tsetse fly numbers by integrating in-situ tsetse catches with satellite-based variables. Furthermore, the impact of different feature elimination techniques on tsetse predictions has not been assessed previously.

We employed freely and readily available satellite-based estimates of human population, land cover, soil properties, elevation, precipitation, and land surface temperature. In previous studies, these environmental variables have been effectively associated with tsetse relative abundance (Gachoki et al., 2023a; Lord et al., 2018; Mugenyi et al., 2021; Ngonyoka et al., 2017a, 2017b) and distribution (Gachoki et al., 2021). Our main aim was to assess how well a random forest model, utilizing different sets of predictor variables, could predict the relative spatial abundance of *Glossina pallidipes* tsetse species across a broader region of Kenya, while utilizing geographically constrained tsetse data. The tsetse trap locations used in this study have previously been employed to predict tsetse habitats at different life stages (Gachoki et al., 2021), evaluate the accuracy of transferring these habitats between regions (Gachoki et al., 2023b), and comprehend the environmental factors driving the observed temporal dynamics in tsetse numbers (Gachoki et al., 2023a). Here, our specific objectives were to: a) leverage readily available satellite-based estimates of environmental factors and train a random forest model for predicting the relative spatial abundance of *G. pallidipes*, b) apply the trained model beyond the spatial domain of the data used for training and evaluate the reliability of the spatial predictions, and c) identify limitations and opportunities for accurately mapping the relative spatial densities of tsetse flies beyond the monitored localities.

2. Data and methods

2.1. Tsetse density data

The tsetse fly data used to train the random forest model were collected from three geographically separated areas in Kenya (**Fig. 1a**, **b**, and **c**) during different months in 2021 (Table 1). Tsetse monitoring involved a mix of biconical (Brightwell et al., 1987), and NGU (Dransfield et al., 1991) traps baited with cow urine and acetone, except in the Nguruman conservancy (Fig. 1b), where only NGU traps were deployed. In all regions, the traps were emptied every two days, with two or four repetitions per month (see Table 1). Although in previous studies (Gachoki et al., 2021, 2023a, 2023b) data from traps at the same, or similar, locations were used, those studies did not use the 2021 data used here. The *G. pallidipes* tsetse species was the sole species that occurred in all sampling sites and accounted for the majority of trapped tsetse species (Table 1). While we have highlighted the percentages of other tsetse fly species captured per region, all analyses in this study utilize only the *G. pallidipes* data. The datasets for the three areas were combined into a single database. To standardize and ensure comparability across the diverse monitoring periods of our traps, we implemented a systematic approach to the data analysis. First, we calculated the number of flies per trap per day (FTD) by dividing the total *G. pallidipes* count with the number of days the trap was monitored. Given the trapping period ranged between 8 and 32 days in our data (Table 1), we wanted to avoid the effect of monitoring effort on our tsetse fly abundance observations entering the model. To achieve this, we assumed a baseline expectation of at least one tsetse fly catch every 8 days. In instances where traps were monitored for periods exceeding 8 days with only one tsetse fly trapped, these were treated as if no flies were caught. While it is known that the number of tsetse flies entering a

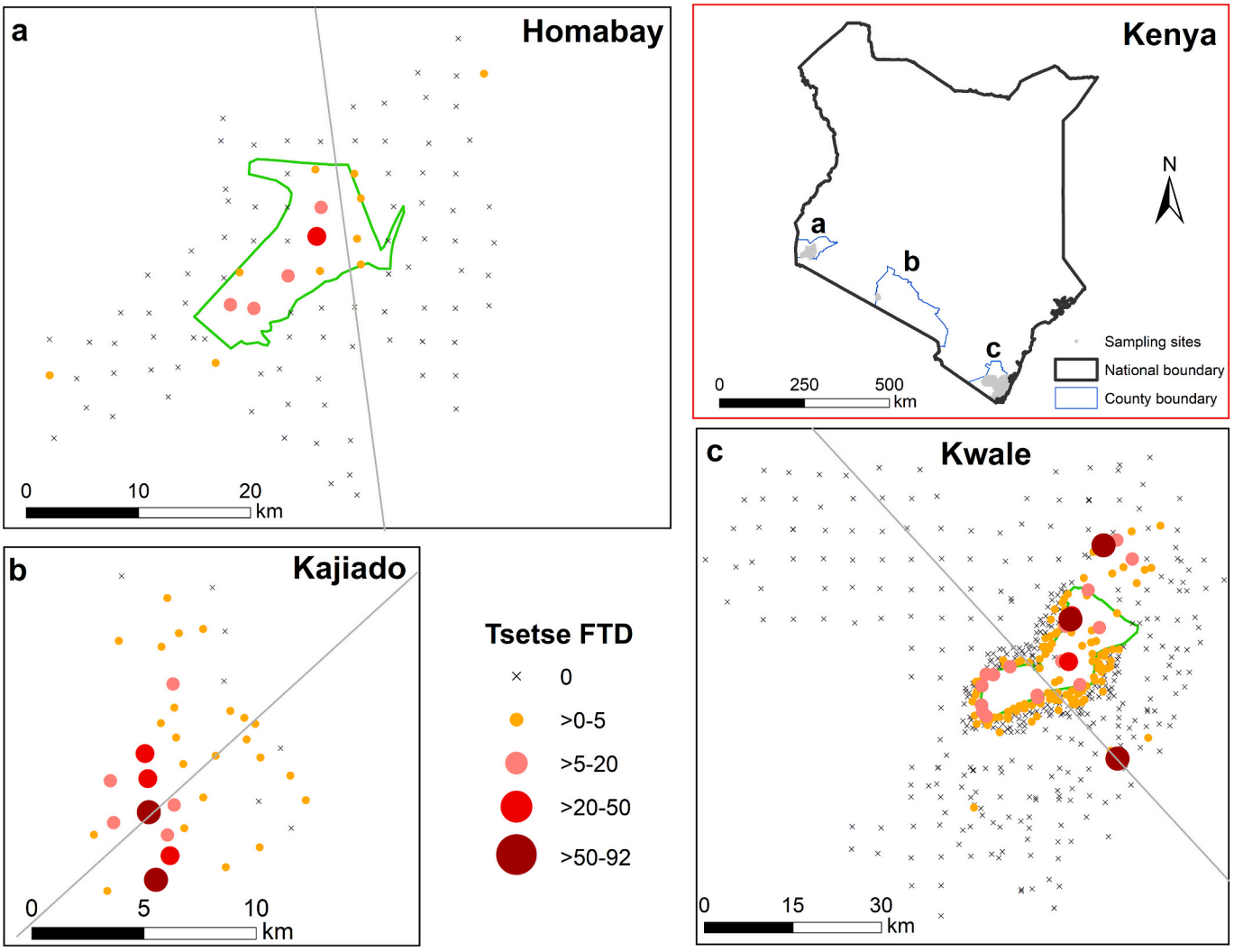


Fig. 1. Sampling sites for tsetse fly in Homabay (a), Kajiado (b) and Kwale (c) counties in Kenya. The flies per trap per day (FTD) belong to trapped Glossina pallidipes species. The light green boundaries in a) and c) are the Ruma National Park and Shimba Hills National Reserve boundaries respectively. The grey diagonal lines indicate the split of each cluster for spatial cross-validation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

trap is usually low (Lindh et al., 2009) and that bi-conical traps are less efficient in trapping *G. pallidipes* compared to the NGU trap (Asfaw et al., 2022; Dransfield and Brightwell, 2001), this adjustment allowed us to align the data and facilitate meaningful comparisons across all monitoring periods, excluding the effect of monitoring effort. For modeling purposes, we converted the FTD-values into $\log_{10}(\text{FTD} + 1)$ to address data skewness and variance stabilization (Feng et al., 2014); adding +1 to avoid undefined numbers in case of no observations).

2.2. Environmental and weather predictor variables

To minimize model complexity, we utilized predictor variables that have been demonstrated to correlate with tsetse occurrence and population dynamics in previous studies (Gachoki et al., 2021, 2023a; Lord et al., 2018; Mugenyi et al., 2021). These included land cover fractions for multiple classes, elevation, average daily annual rainfall, soil moisture, land surface temperature, human population density, sand, and silt content (Table 2). We created two databases; one for 2021 (henceforth referred to as "2021") and another covering multi-annual averages from 2011 to 2021(referred to as "2011–2021"). For land cover fractions we calculated the percentage 10 m-by-10 m of pixels of the European Space Agency (ESA) global land cover (Zanaga et al., 2021) data within a 1 km-by-1 km grid for each land cover class. The aggregation of land cover classes into a 1 km-by-1 km grid size was chosen based on findings that the distribution of *G. pallidipes* is significantly influenced by the abundance of vegetation cover within these distances (Gachoki et al., 2021). Additionally, this resolution aligns with previous research suggesting the potential for tsetse flies to travel distances of up to 1 km within their geographic range in a day (Vale et al., 1984; Williams et al., 1992). All the other predictor variables were sourced from openly available resources (Table 2) and resampled to a 1 km-by-1 km spatial resolution by taking the average value.

We extracted the relevant values of each set of environmental and weather variable for both 2021 and 2011–2021 at the trap locations ($n = 660$) and a random sample of the same size and plotted histograms to compare each predictor variable at the trap level against the random sample. The two databases differed solely in terms of temporary varying variables.

2.3. Predictive modeling

2.3.1. Variable elimination

While machine learning methods can partially address multicollinearity (Fig. A.1), removing redundant and irrelevant variables can

Table 1
Details on tsetse traps for the three geographic regions in Kenya. n = number of tsetse flies observed.

	Homabay County *n = 696*	Kajiado County *n = 7606*	Kwale County *n = 11,396*
Number of traps	111	40	22 – *inside Shimba Hills National Reserve (SHNR)* 231 – *within 5 km outside SHNR* 256 – *larger Kwale*
Months monitored (#days for each month)	November (8)	February (8), June (8), and November (8)	*Inside SHNR* – February (8) *Within 5 km of SHNR* - February (8), March (8), April (8), June (8) *Larger Kwale* – April (4), June (4)
Tsetse species identified (%)	*G. pallidipes (86)* *G. fuscipes (14)*	*G. pallidipes (99)* *G. longipennis (1)*	*G. pallidipes (92)* *G. brevipalpis (3)* *G. austeni (5)*
Elevation range (m.a.s.l.)	1100–1585	640–800	7–460

prevent potential performance degradation of the model and avoids overfitting, which can limit the ability to make good extrapolations to unseen regions (Duque-Lazo et al., 2016). In this study, we employed two common variable elimination techniques for random forests: 1) RFE (Recursive Feature Elimination, Khun, 2022) and 2)VSURF (Variable Selection Using Random Forests; Genuer et al., 2022). In RFE, the user defines a termination condition for model performance, and the algorithm iteratively removes one variable at a time while evaluating its impact on the model's performance. This process continues until the algorithm reaches the best predefined level of model performance. In this study, Root Mean Square Error (RMSE) was used as the termination condition. As a result, RFE retained all variables whose removal led to a deterioration in the best RMSE value. VSURF follows a three-step process. First, it utilizes random forests to evaluate variable importance and systematically removes those with very low importance. Second, the retained variables are ranked based on their importance scores, and the top-ranked variables that contribute most to the predictive power of the model are retained. Finally, these selected variables are employed for making predictions of the response variables, retaining only those that reduce the prediction error. Both the RFE and VSURF variable elimination methods were independently applied to the two sets of data, resulting in four retained databases: 1) RFE_{2021}, 2) $RFE_{2011–2021}$, 3) $VSURF_{2021}$, and 4) $VSURF_{2011–2021}$. All analyses were performed in R programming using the Caret (Khun, 2022) and VSURF (Genuer et al., 2022) packages.

2.3.2. Spatial cross-validation, model training, and spatial mapping

To create spatial clusters for the spatial cross-validation, we divided the tsetse monitoring traps in each area into two distinct sets by placing a diagonal line across each area, ensuring that both presence and absence samples were present in each set (Fig. 1; grey diagonal lines). We partitioned the data into training and testing datasets with the "CreateSpacetimeFolds" function from the CAST package (Meyer et al., 2019) resulting in a single dataset with six distinct training ("index") and corresponding testing ("indexOut") subsets, which were used directly as part of the tuning parameters during the model training process. We used the log transformed FTD as our response variable and applied the ranger method (a faster implementation of random forest; Wright and Ziegler, 2017) within the caret package to fit the random forest model with each set of predictor variables.

The model performance was assessed based on the final trained model (average of the six models) "out-of-bag" R^2 and RMSE values, which is an indication of how well the model generalizes the unseen data. Variable importance was determined using the permutation method, where variables are ranked according to how much the model's performance degrades when the values of specific variables are randomly shuffled (Breiman, 2001). We used the 'pdp' package (Greenwell, 2022) to generate partial dependence plots, which show the impact of each predictor on the response variable while maintaining all other predictors constant.

To create spatial maps of tsetse relative abundance, we applied the trained models to the respective predictor variables used during their training, covering the entire land area of Kenya. Because of the geographical constraints within our training dataset, we utilized the "aoa" (Area of Application) function from the CAST package to demarcate areas for which the environmental conditions are sufficiently represented by our training data (Meyer and Pebesma, 2021). The delineation of these areas depends on a threshold value that is internally calculated by measuring the dissimilarity between the predictor variables in the training data and those used in the model extrapolation. Predictor variables are assigned weights in this dissimilarity measurement according to their significance in explaining the response variable.

Table 2
Predictor variables used and their sources. The ranges indicate the minimum and maximum values for the whole of Kenya. For the variable column, the italicized letters in brackets represent the acronym used to refer to these variables in this study.

Variable	Spatial resolution (m)	Units	Range	Source	Year
Tree cover (*Tlc*) Cropland cover (*Clc*) Sparse vegetation cover (*SVlc*) Shrub cover (*Slc*) Grassland cover (*Glc*) Built-up cover (*Blc*)	10	%	0–100	ESA (Zanaga et al., 2021)	2020
Elevation (*El*)	30	m.a.s.l.	−0.9–4679.8	NASA Shuttle Radar Topography Mission (Farr et al., 2007)	2000
Slope (*Sl*)		degrees	0–26.8		
Silt content (*Si*)	30	%	0–42.36	Innovative solutions for Decision Agriculture (Hengl et al., 2021)	2017
Sand content (*Sa*)	30	%	0–75.67		
Population density (*Pd*)		people/ $100m^2$	0–772	https://www.worldpop.org/	2020
Soil moisture (*Sm*)	4863	mm	0.1–195.93	TerraClimate (Abatzoglou et al., 2018)	2021
Annual daily average precipitation (*P*)	5566	mm/day	0.084–9.37	Climate Hazards Group InfraRed Precipitation with Station Data (Funk et al., 2015)	2021
LST day (*lstD*)	1000	°C	16.56–64.34	Moderate Resolution Imaging Spectroradiometer (MODIS)	2021
LST night (*lstN*)			0.02–34.1		

3. Results

3.1. Distribution of predictor variable values

Thirty-two percent of the 660 trapping locations had at least one fly trapped every eight days. The distribution of predictor variables at the trap level represented only a fraction of the broader range found within Kenya for most of the variables (Fig. 2). For example, temperature is a key factor influencing tsetse population dynamics (Are and Hargrove, 2020) and the range of daytime land surface temperature within our dataset is ~30 °C–45 °C while in larger Kenya it is ~20 °C–50 °C. This implies that extrapolating the trained model to areas where predictor variables have values outside the training data range, may result in underestimation or overestimation of tsetse relative abundance, depending on how those variables influence the relative abundance of tsetse.

3.2. Feature elimination and model performance

The two elimination techniques retained different sets of predictor variables, but with some overlaps. Notably, the RFE method retained a higher number of variables compared to the VSURF elimination method. Precipitation (P), population density (Pd), soil moisture (Sm), and tree cover fraction (Tlc) were retained for both elimination techniques across the two datasets (Table 3). Models trained using the predictor variables from 2021 measured an "Out-Of-Bag" R^2 and RMSE value of 0.41 and 0.52, respectively. In contrast, models trained using multi-year averages from 2011 to 2021 exhibited a lower average R^2 (0.38) and higher average RMSE value (0.55). This suggests that the spatial variability of tsetse fly numbers is better explained by environmental and weather conditions near the time of sampling.

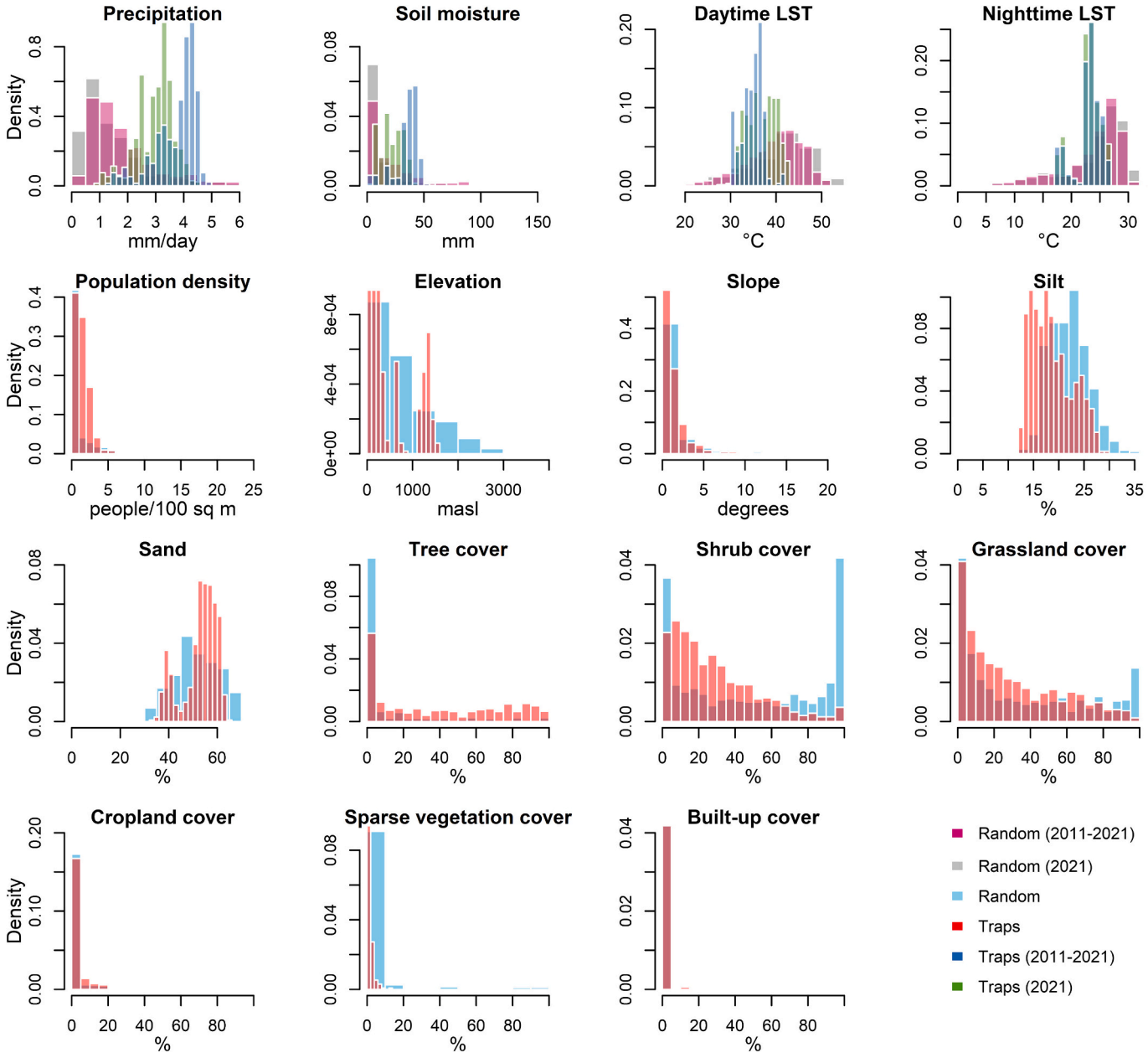


Fig. 2. Histograms showing the distribution of the various predictor variables at trap level and at random samples level within Kenya. The sky-blue (random) and red (traps) bars represent the static predictor variables among the two data combinations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Retained variables based on the various feature elimination techniques.

	P	Pd	El	Sl	Si	Sa	Sm	lstD	lstN	Tlc	Glc	Clc
RFE_{2021}	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
$RFE_{2011-2021}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$VSURF_{2021}$	✓	✓					✓			✓		
$VSURF_{2011-2021}$	✓	✓					✓			✓		

3.3. Variable importance and partial dependence plots

Discrepancies were observed in the ranking of key predictors elucidating the spatial variability in relative tsetse numbers across the four sets of predictors utilized (Fig. 3**).** Within the subset of predictors determined through the VSURF feature elimination method, tree cover percentages and precipitation emerged as the foremost variables explaining *G. pallidipes*, both for yearly (2021) and multi-annual (2011−2021) average predictors. In the RFE retained predictors soil moisture and human density were among the top predictors for 2021 while for 2011–2021, significant predictors included cropland fractions and precipitation.

We observed that tsetse numbers exhibited an increase with higher tree cover fractions and a decrease as population density, croplands, soil moisture and rainfall increased (Fig. 4). The presence of abundant tree cover creates favorable conditions for tsetse populations, as it provides shaded areas for their resting and breeding. Conversely, in densely populated regions, humans may alter environments that were originally conducive to tsetse flies through activities such as the removal of woodlands and wild hosts, resulting in a reduced population of flies. Additionally, we observed a unimodal response with soil moisture, suggesting that both very dry and very wet soils are unsuitable for tsetse flies (Fig. 4). The plausible explanation is that excessively dry soils are too hard, preventing larvae from burrowing and pupating, while overly wet conditions can lead to drowning. Furthermore, excessive rainfall may cause flooding or water accumulation, posing risks to both adult tsetse flies and their larvae, or washing away burrowed pupae. Ultimately, these factors contribute to a decline in tsetse populations.

3.4. Tsetse relative density maps

The extrapolated predictions for Kenya reveal significant disparities among the four models (Fig. 5). The hatched black lines show areas that fall outside the range of environmental conditions observed in our training data based on the area of application analysis and predictions in these regions should be regarded as less reliable. Note that the hatched areas also show differences between maps because different predictor variables are used. Models generated using variables retained through VSURF for year 2021 (Fig. 5b) exhibit more prominent tsetse hotspots outside of monitored regions when compared to the predictions based on other sets of variables (**Fig. 5a, c** and **d**). The majority of these hotspots (>6 FTD; **Fig. 5b** and **c**) are within known tsetse fly belts In Kenya (DeVisser and Messina, 2009). However, without tsetse ground data, it is impossible to definitively conclude that this reflects the actual situation.

4. Discussion

The primary objective of this study was to evaluate how well a classical random forest machine learning model together with satellite-based environmental estimates can predict relative tsetse abundance in all of Kenya using a spatially limited set of tsetse trapping data. Based on our results, in this section we also pinpoint areas of improvement and opportunities to enhance the precision and reliability of predictions of tsetse fly relative densities within Kenya.

4.1. National scale tsetse mapping with spatially limited data

Different sets of predictors revealed distinct important variables (Fig. 3). In the RFE datasets for 2021, the top predictors were soil moisture, human density, and tree cover. In contrast, for the long-term averages (2011–2021), highly ranked predictors included croplands, mean annual precipitation, and tree cover percentages. In the VSURF dataset, only four variables were retained, and the notable difference in their ranking was that soil moisture held a higher rank than human density in the 2021 set of predictors. Tsetse numbers started to decline when the daily annual average rainfall exceeded 2 mm/day and the

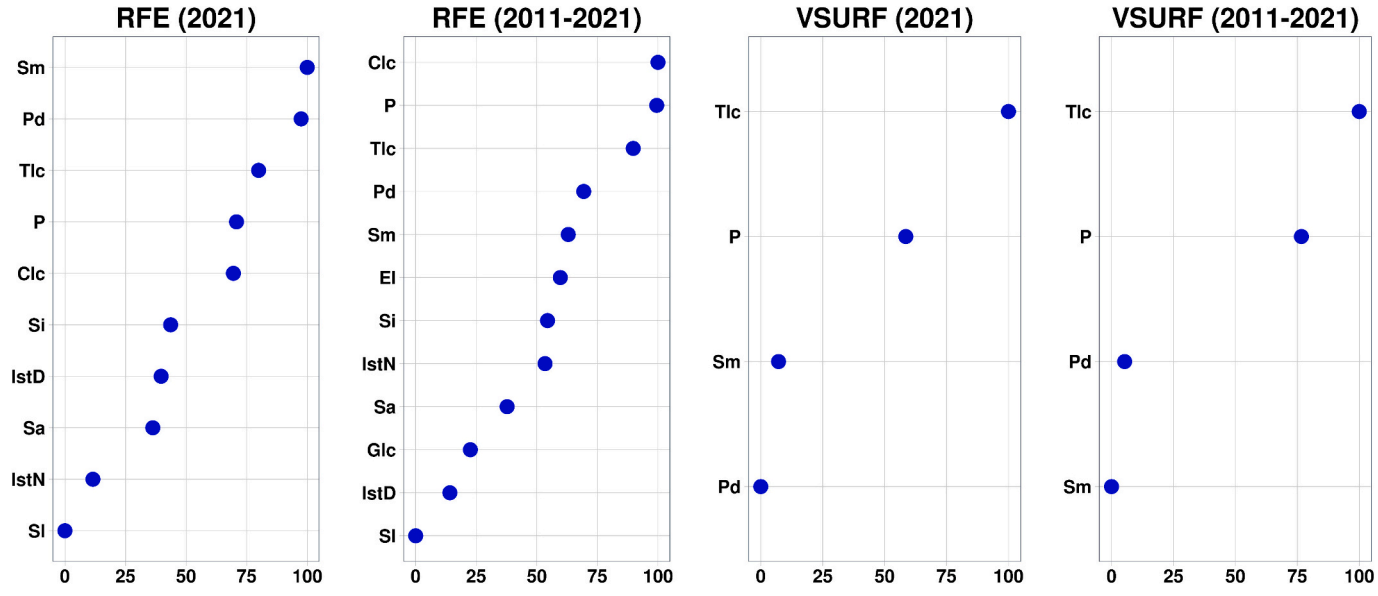


Fig. 3. Variable importance plot for the different dataset combinations and feature elimination techniques.

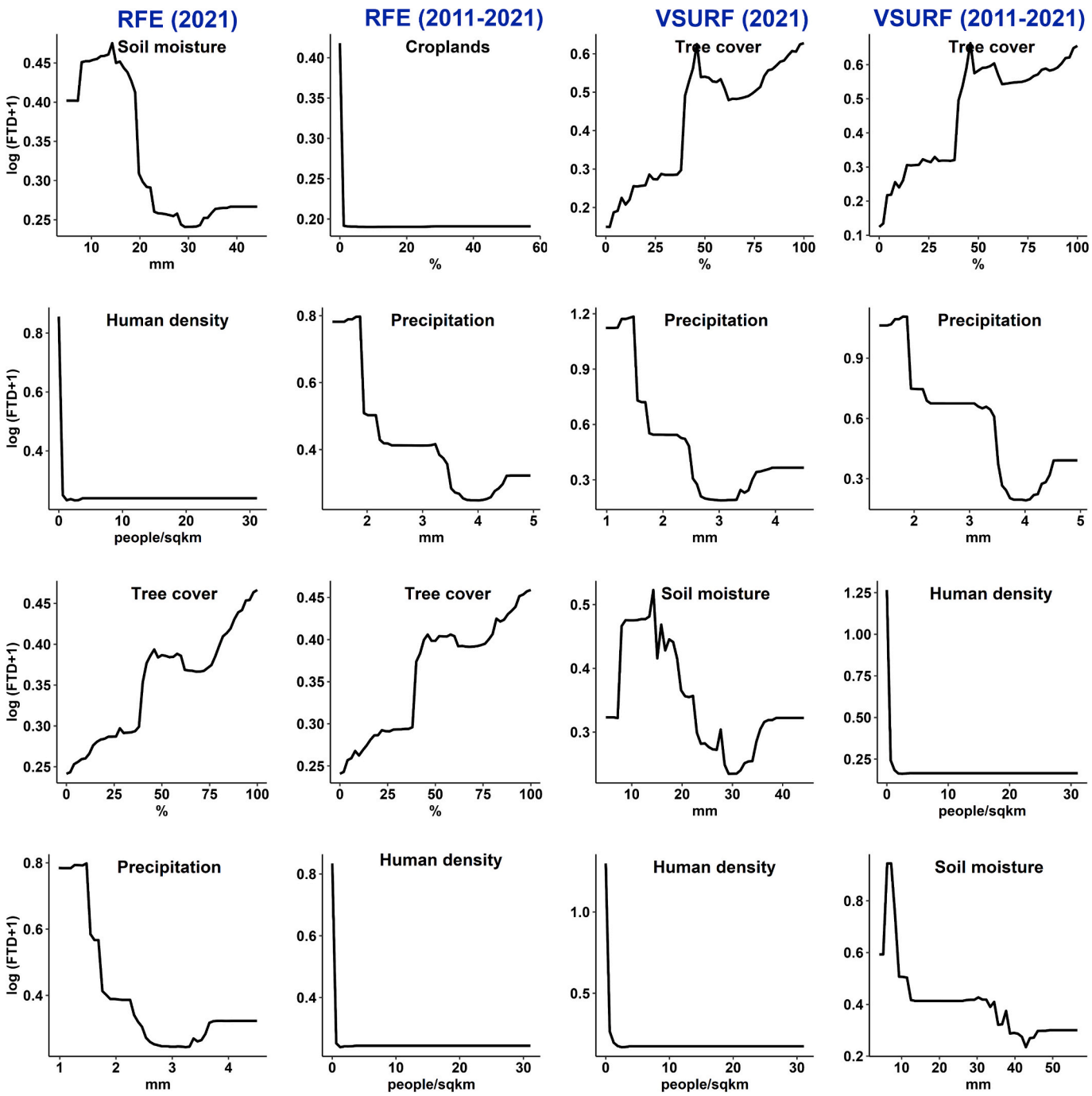


Fig. 4. Partial dependent plots for the top four important predictor variables showing their estimated influence on tsetse abundance for the different set of predictors.

volumetric soil moisture exceeded 10 mm (Fig. 4). This finding is consistent with prior research that examined temporal patterns of abundance, such as by Gachoki et al. (2023a), who found that tsetse numbers rose with increased rainfall but then declined when rainfall increased for more than a month. Intense rainfall can lead to excessive water accumulation, increasing soil moisture, which, in turn, can lead to the submerging or dislodging of buried pupae, ultimately causing a decrease in tsetse populations (Lukaw et al., 2014; Ngonyoka et al., 2017a, 2017b; Omoogun et al., 1989; Signaboubo et al., 2021). Additionally, during periods of heavy rainfall, the behavioral activity of tsetse flies actively seeking a host for feeding is likely to decrease thereby lowering the probability of entrapment.

In densely populated areas, heightened human activities such as clearing land for cultivation and settlements are likely to disrupt the favorable environments for tsetse fly resting and breeding, which explains the observed negative relationship. Conversely, a higher percentage of tree cover offers suitable conditions for tsetse resting and breeding and other research also found tsetse numbers to positively correlate with abundant vegetated areas. For example, Lord et al. (2018) reported that high *G. pallidipes* abundance in Serengeti National Park, Tanzania, correlated with areas rich in vegetation, and Mugenyi et al. (2021) documented that another tsetse species, *G. fuscipes fuscipes*, also exhibited high numbers in vegetated regions. Shaded areas, such as those with ample tree cover, create a cooler microclimate that is essential for tsetse flies breeding and resting (Gachoki et al., 2021; Isherwood and Duffy, 1959). These areas also provide refuge for the animal hosts that tsetse flies rely on for blood meals (Isherwood et al., 1961). The combination of these factors may explain the observed

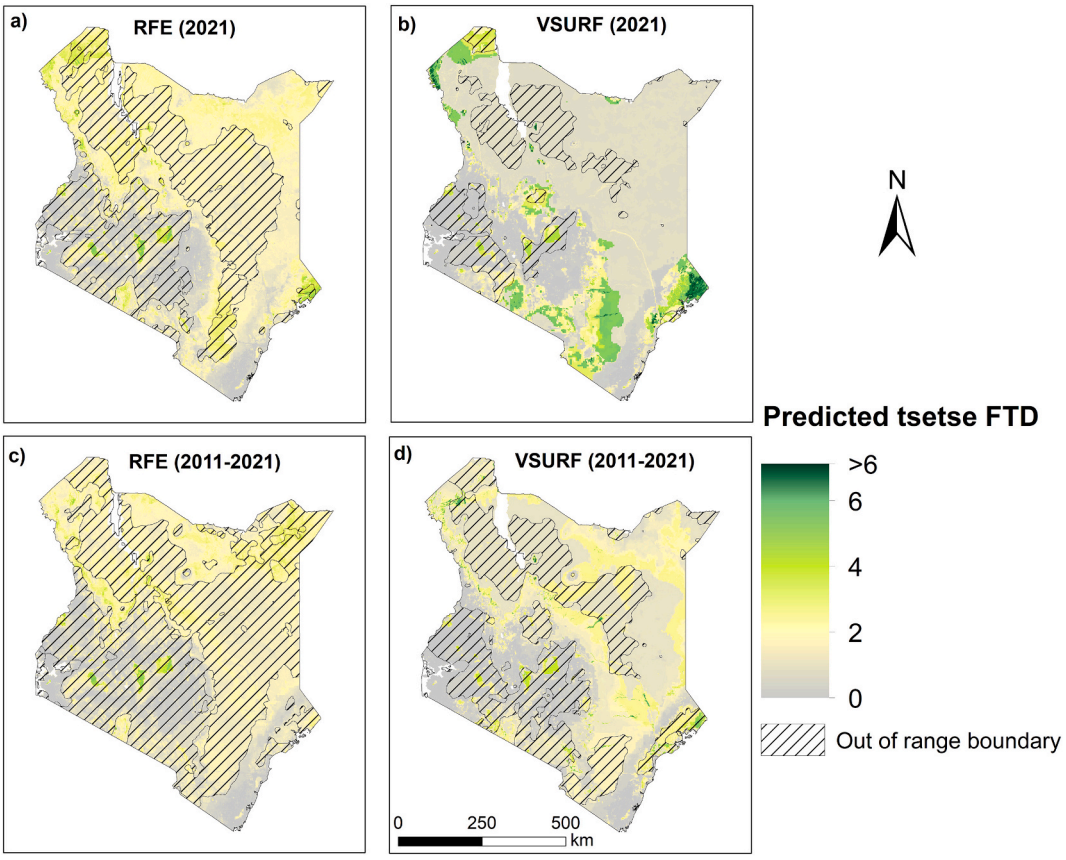


Fig. 5. Extrapolated G. pallidipes FTD. The hatched black lines show areas that were outside the range of the environmental conditions in the training data according to the Area of Application methods (Meyer and Pebesma, 2021).

positive relationship between tsetse flies and tree cover percentage.

The attained performance of the trained models (R^2 values ranging from 0.38 to 0.41) demonstrates the potential for predicting tsetse fly relative numbers using machine learning methodologies. Our analysis indicates that using environmental and weather data near the period of tsetse monitoring yields more accurate predictions compared to longer-term averages. When extrapolating tsetse number predictions based on different sets of predictors, significant disparities emerge. VSURF-retained variables of 2021 reveal pronounced tsetse hotspots (>6 FTD) within known tsetse belts (McCord et al., 2012). On the other hand, predictions based on RFE-retained variables did not identify prominent hotspots, and most of these predictions fell outside the range of the environmental and weather data used for model training. Notably, for RFE-retained variables and VSURF long-term variables, extrapolations indicate higher tsetse number predictions (>2 FTD) in the northwestern (Fig. 5) region of Kenya, which is not a historically known tsetse fly belt.

Previous research such as Lord et al. (2018) also reported overestimations of tsetse numbers by GLM-based models in regions beyond those monitored in Serengeti National Park. They attributed this overestimation to a mismatch between the period tsetse data was collected and when the environmental variables were estimated. However, in our study, most of the predictions of high tsetse numbers occurred in regions for which no tsetse data were available and thus not included in the model training, and where environmental predictors had values outside the range of our training data. Using predictive modeling techniques to extrapolate beyond the training data can lead to less accurate predictions due to the model's limited fitting of the response variable in those conditions (Gutzwiller and Serno, 2023; Muckley et al., 2023). We expected that creating a mask to delineate the "area of applicability" for the trained model (Meyer and Pebesma, 2021) would successfully filter out a significant portion of regions lacking training data, particularly in the Northern and Eastern regions where high tsetse predictions were evident. However, we found that most of these areas still fell within the range where the trained model's accuracy remained valid.

4.2. *Prospects for enhancing large-scale spatial prediction of tsetse abundance*

While this research analysis does establish a basis for predicting tsetse numbers for large areas, the reliability of the current predictions remains uncertain. Consequently, to guarantee that future national-level spatial maps of tsetse abundance are accurate and reliable, it is imperative to undertake several critical steps.

As earlier mentioned, extrapolating predictive modeling techniques beyond the data range used for model training can result in poor predictions because the model lacks knowledge of how the response variable behaves in such conditions (Gutzwiller and Serno, 2023; Muckley et al., 2023). In this study, the utilized tsetse fly data did not encompass all environmental conditions in Kenya, highlighting the need for additional trapping data covering a wide range of such conditions. The recently published Kenya tsetse atlas reveals that additional data exist from various sources (Ngari et al., 2020). However, trapping data is lacking in certain areas, particularly in the northern and eastern regions, where our current models consistently predict high tsetse abundance (Fig. 5a, c, d). Without trap data from these localities, it becomes impossible to validate the current models and this may equally hinder the development of improved predictive models, even when incorporating data from the atlas. While initiatives like COMBAT (Controlling and progressively Minimizing the Burden of Animal Trypanosomiasis; Boulangé et al., 2022) can use these research findings to identify areas requiring increased sampling efforts, these regions might still be extensive, leading to high tsetse sampling costs. A cost-effective

alternative would be to consider the implementation of citizen mapping (Hamer et al., 2018). In this approach, local communities would receive training on identifying tsetse flies in set traps and reporting their findings over time. Similar programs have proven effective in mapping other disease vectors, such as mosquitoes (Cohnstaedt et al., 2016; Palmer et al., 2017) and ticks (Laaksonen et al., 2017; Xu et al., 2016).

Tsetse flies rely on blood from both wild and domesticated mammals for survival (Ducheyne et al., 2009; Rogers, 1979), but this study lacked information on the distribution of animal hosts that tsetse flies feed on. Consequently, the predictive results in this study only explain tsetse densities based on environmental variables. This has a drawback that unsampled areas may have environmental and weather conditions favorable to tsetse, but, where nonetheless tsetse flies will not be present due to host absence. Therefore, incorporating data on the distribution of animal hosts can help exclude such areas, refining the extent of tsetse distribution. However, obtaining animal distribution data is challenging. While animal tracking seems like a viable method, the associated costs and potential reluctance from wildlife managers, who view tsetse flies as "guardians of Africa's biodiversity," make this approach less likely (Rogers and Randolph, 1988). An alternative solution could involve using publicly available information on protected zones as a predictor variable, given that many of these areas serve as refuges for wildlife. If localities beyond the protected zones are identified as hosting high tsetse fly densities, ground-truthing efforts may be necessary.

Another improvement to consider is developing a land cover map with classes specifically associated with the tsetse species under consideration. In this study, the primary tsetse species was *G. pallidipes*, which is positively correlated with woodlands, a land cover class that was absent from the freely available land cover layer we utilized. Also, the way satellite-based data is integrated into the models is of paramount importance. The prevailing approach in most tsetse predictive mapping models involves establishing a direct correlation between tsetse presence or abundance and various attributes related to vegetation cover. These attributes are derived from the actual and static land cover observed at the trapping site. However, given that tsetse flies move within their geographic range in search of a host to feed on (Brightwell et al., 1992) and tsetse traps are strategically positioned in areas where tsetse flies perceive them as potential hosts (Fuentes, 2017), it becomes highly likely that the land cover at tsetse trap locations may not accurately depict the genuine environmental conditions sustaining the tsetse population. While studies like the one conducted by Lord et al. (2018) made attempts to incorporate the tsetse dispersal range by employing a buffer to calculate averages of dynamic variables like LST, when it comes to land cover classes, it might prove more advantageous to calculate the overall percentage of each land cover class within a radius that corresponds to the typical movement range of tsetse flies (Gachoki et al., 2021). This adjustment has the potential to significantly enhance model performance, particularly when dealing with categorical data such as land cover.

The selection of appropriate parameters for model tuning is crucial when constructing predictive models. While it is widely acknowledged that machine learning methods, such as the random forest, excel at handling multicollinear data, the inclusion of irrelevant variables can significantly diminish model performance. Furthermore, the choice of a variable elimination method should be made with careful consideration, considering the representativeness of the training data in relation to the broader environmental conditions to which the model will be extrapolated. Our study demonstrates that a variable elimination method that emphasizes retaining variables on how well they reduce the prediction error yields more reliable (based on known tsetse belts; McCord et al., 2012) results compared to methods that retain variables solely based on their importance in explaining the training data. Furthermore, future studies should explore the possibility of aligning tsetse observations with environmental and weather data collected during the same period as tsetse monitoring, as this is likely to enhance model performance.

Lastly, when evaluating model performance, the common practice involves the separation of the training and test data beforehand. However, this approach can introduce bias, particularly when dealing with data that contains many zero values. Therefore, we strongly recommend that future research efforts adopt spatial cross-validation techniques to bolster model robustness. Spatial cross-validation operates by randomly selecting blocks of data for training while reserving others for testing. This process is repeated multiple times based on the number of specified folds. Through this iterative approach, the model refines its estimation of prediction errors, consequently enhancing overall model performance (Meyer et al., 2019).

5. Conclusion

Our research presents a framework for the prediction of relative *G. pallidipes* densities for large areas. Our findings indicate that to achieve a more reliable relative tsetse abundance map in Kenya, additional tsetse sampling is essential. This necessity arises because our models predicted high tsetse numbers in regions lacking in-situ tsetse trap data, potentially indicating an underrepresentation of environmental conditions in the training data. The method employed for eliminating irrelevant variables is crucial when extrapolating predictions beyond monitored regions. The VSURF elimination method, which retains variables based on their ability to reduce prediction errors, offered a more reliable approach for extrapolation. While the accuracies of the extrapolated predictions in this analysis remain uncertain, our comprehensive map of tsetse relative densities for Kenya serves as a valuable tool that relevant organizations can effectively leverage to optimize and strategically deploy their surveillance efforts.

Funding

The authors gratefully acknowledge the financial support for this research by the following organizations and agencies: the German Federal Ministry for Economic Cooperation and Development (BMZ) commissioned and administered through the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Fund for International Agricultural Research (FIA), grant number 81235250; the Royal Netherlands Academy of Arts and Sciences (KNAW) Ecology Fund, grant number KNAWWF/747/ECO2021–7; the Schlumberger Foundation (Faculty for the Future fellowship; 2023–2024); the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Australian Centre for International Agricultural Research (ACIAR); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya. The views expressed herein do not necessarily reflect the official opinion of the donors.

CRediT authorship contribution statement

Stella Gachoki: Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Thomas A. Groen:** Writing – review & editing, Visualization, Supervision, Methodology, Conceptualization. **Anton Vrieling:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Andrew Skidmore:** Writing – review & editing, Supervision. **Daniel Masiga:** Writing – review & editing, Supervision, Funding acquisition.

Data availability

All tsetse data and the associated R-programming scripts will be made publicly available through the Data Archiving and Network Services (DANS) repository (https://doi.org/10.17026/PT/YUVINN).

Acknowledgment

The authors would like to extend their gratitude to the *icipe* animal

health team and the KENTTEC team for their dedication to installing traps and collecting tsetse fly data in both Ruma National Park and Nguruman Conservancy. We would also like to extend our gratitude to Mr. Willem Nieuwenhuis of Faculty ITC for the assistance accorded in generating the land cover fractions.

Appendix A. Appendices

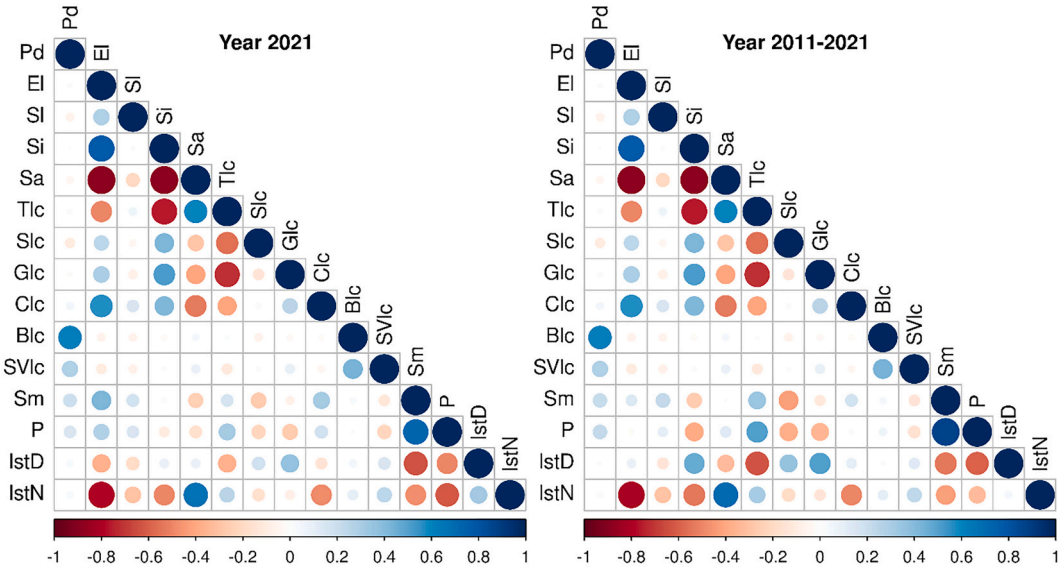


Fig. A.1. The correlation matrix displays the relationships between different predictor variables. "Year 2021" represents the temporary variables observed in 2021, while "Year 2011–2021" represents the averages across multiple years.

References

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5 (1), 1–12. https://doi.org/10.1038/sdata.2017.191.

Are, E.B., Hargrove, J.W., 2020. Extinction probabilities as a function of temperature for populations of tsetse (Glossina spp.). PLoS Negl. Trop. Dis. 14 (5), e0007769 https://doi.org/10.1371/journal.pntd.0007769.

Asfaw, N., Hiruy, B., Worku, N., Massebo, F., 2022. Evaluating the efficacy of various traps in catching tsetse flies at Nech Sar and Maze National Parks, Southwestern Ethiopia: an implication for Trypanosoma vector control. PLoS Negl. Trop. Dis. 16 (12), e0010999 https://doi.org/10.1371/JOURNAL.PNTD.0010999.

Bishop, A.P., Amatulli, G., Hyseni, C., Pless, E., Bateta, R., Okeyo, W.A., et al., 2021. A machine learning approach to integrating genetic and ecological data in tsetse flies (*Glossina pallidipes*) for spatially explicit vector control planning. Evol. Appl. 14 (7), 1762–1777. https://doi.org/10.1111/EVA.13237.

Boulangé, A., Lejon, V., Orcid, D.B., Thévenon, S., Gimonneau, G., Desquesnes, M., et al., 2022. The COMBAT project: Controlling and progressively minimizing the burden of vector-borne animal trypanosomosis in Africa [version 2; peer review: 3 approved]. Available at: http://localhost:8080/xmlui/handle/20.500.12562/1758 (Accessed: 12 November 2023).

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Brightwell, R., Dransfield, R.D., Kyorku, C., Golder, T.K., Tarimo, S.A., Mungai, D., 1987. A new trap for *Glossina pallidipes*. Int. J. Pest Manag. 33 (2), 151–159. https://doi.org/10.1080/09670878709371136.

Brightwell, R., Dransfield, R.D., Williams, B.G., 1992. Factors affecting seasonal dispersal of the tsetse flies *Glossina pallidipes* and *G. longipennis* (Diptera: Glossinidae) at Nguruman, south-West Kenya. Bull. Entomol. Res. 82 (2), 167–182. https://doi.org/10.1017/S0007485300051695.

Carrasco-Escobar, G., Moreno, M., Fornace, K., Herrera-Varela, M., Manrique, E., Conn, J.E., 2022. The use of drones for mosquito surveillance and control. Parasit. Vectors 15 (1), 473. https://doi.org/10.1186/s13071-022-05580-5.

Ceia-Hasse, A., Sousa, C.A., Gouveia, B.R., Capinha, C., 2023. Forecasting the abundance of disease vectors with deep learning. Eco. Inform. 78, 102272 https://doi.org/10.1016/J.ECOINF.2023.102272.

Cohnstaedt, L.W., Ladner, J., Campbell, L.R., Busch, N., Barrera, R., 2016. Determining mosquito distribution from egg data: the role of the citizen scientist. Am. Biol. Teach. 78 (4), 317–322. https://doi.org/10.1525/ABT.2016.78.4.317.

Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19 (S1), 65. https://doi.org/10.1186/s12863-018-0633-8.

De Beer, C.J., Dicko, A.H., Ntshangase, J., Moyaba, P., Taioe, M.O., Mulandane, F.C., et al., 2021. A distribution model for Glossina brevipalpis and Glossina austeni in Southern Mozambique, Eswatini and South Africa for enhanced area-wide integrated pest management approaches. PLoS Negl. Trop. Dis. 15 (11), e0009989 https://doi.org/10.1371/JOURNAL.PNTD.0009989.

Desquesnes, M., Dia, M.L., 2003. Trypanosoma vivax: mechanical transmission in cattle by one of the most common African tabanids, Atylotus agrestis. Exp. Parasitol. 103 (1–2), 35–43. https://doi.org/10.1016/S0014-4894(03)00067-5.

DeVisser, M.H., Messina, J.P., 2009. Optimum land cover products for use in a Glossina-morsitans habitat model of Kenya. Int. J. Health Geogr. 8 (1), 39. https://doi.org/10.1186/1476-072X-8-39.

Dlamini, S.N., Beloconi, A., Mabaso, S., Vounatsou, P., Impouma, B., Fall, I.S., 2019. Review of remotely sensed data products for disease mapping and epidemiology. Remote Sens. Appl. Soc. Environ. 14, 108–118. https://doi.org/10.1016/J.RSASE.2019.02.005.

Dransfield, R.D., Brightwell, R., 2001. Trap efficiency for *Glossina pallidipes* (Diptera: *Glossinidae*) at Nguruman, south-West Kenya. Bull. Entomol. Res. 91 (6), 429–444. https://doi.org/10.1079/BER2001127.

Dransfield, R.D., Williams, B.G., Brightwell, R., 1991. Control of tsetse flies and trypanosomiasis: myth or reality? Parasitol. Today 7 (10), 287–291. https://doi.org/10.1016/0169-4758(91)90099-A.

Ducheyne, E., Mweempwa, C., De Pus, C., Vernieuwe, H., De Deken, R., Hendrickx, G., et al., 2009. The impact of habitat fragmentation on tsetse abundance on the plateau of eastern Zambia. Prev. Vet. Med. 91 (1), 11–18. https://doi.org/10.1016/J.PREVETMED.2009.05.009.

Duque-Lazo, J., Van Gils, H., Groen, T.A., Navarro-Cerrillo, R.M., 2016. Transferability of species distribution models: the case of Phytophthora cinnamomi in Southwest Spain and Southwest Australia. Ecol. Model. 320, 62–70. https://doi.org/10.1016/j.ecolmodel.2015.09.019.

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al., 2007. The shuttle radar topography mission. Rev. Geophys. 45 (2), 2004. https://doi.org/10.1029/2005RG000183.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., et al., 2014. Log-transformation and its implications for data analysis. Shanghai Arch. Psychiatry 26 (2), 105. https://doi.org/10.3969/J.ISSN.1002-0829.2014.02.009.

Fuentes, A., 2017. Colors of Doom: What does the tsetse fly see? Available at: https://socobilldurham.stanford.edu/sites/g/files/sbiybj10241/f/final_fuentes_colorsofdoom_sophomorecollege2017.pdf (Accessed: 26 May 2020).

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci. Data 2 (1), 1–21. https://doi.org/10.1038/sdata.2015.66.

Gachoki, S., Groen, T., Vrieling, A., Okal, M., Skidmore, A., Masiga, D., 2021. Satellite-based modelling of potential tsetse (*Glossina pallidipes*) breeding and foraging sites using teneral and non-teneral fly occurrence data. Parasit. Vectors 14 (1), 1–18. https://doi.org/10.1186/S13071-021-05017-5.

Gachoki, S., Groen, T., Vrieling, A., Skidmore, A., Masiga, D., 2023a. Evidence-based advice on timing and location of tsetse control measures in Shimba Hills National reserve, Kenya. PLoS Negl. Trop. Dis. 17 (6) https://doi.org/10.1371/JOURNAL.PNTD.0011398.

Gachoki, S., Groen, T., Vrieling, A., Skidmore, A., Masiga, D., 2023b. Transferability of tsetse habitat models between different regions in Kenya and Rwanda. Ecol. Model. 486, 110548 https://doi.org/10.1016/J.ECOLMODEL.2023.110548.

Genuer, R., Jean-Michel, P., Christine, T.-M., 2022. Variable Selection Using Random Forests (VSURF), CRAN repository. Available at: https://cran.r-project.org/web/packages/VSURF/VSURF.pdf (Accessed: 26 January 2023).

González Jiménez, M., Babayan, S.A., Khazaeli, P., Doyle, M., Walton, F., Reedy, E., et al., 2019. Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. Wellcome Open Res. 4, 76. https://doi.org/10.12688/wellcomeopenres.15201.3.

Greenwell, B., 2022. Package 'pdp'- Partial Dependence Plots. Available at: https://cran.r-project.org/web/packages/pdp/pdp.pdf.

Gutzwiller, K.J., Serno, K.M., 2023. Using the risk of spatial extrapolation by machine-learning models to assess the reliability of model predictions for conservation. Landsc. Ecol. 38 (6), 1363–1372. https://doi.org/10.1007/s10980-023-01651-9.

Hamer, S.A., Curtis-Robles, R., Hamer, G.L., 2018. Contributions of citizen scientists to arthropod vector data in the age of digital epidemiology. Curr. Opin. Insect Sci. 28, 98–104. https://doi.org/10.1016/J.COIS.2018.05.005.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., et al., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11 (1), 1–18. https://doi.org/10.1038/s41598-021-85639-y.

Ibañez-Justicia, A., Cianci, D., 2015. Modelling the spatial distribution of the nuisance mosquito species Anopheles plumbeus (Diptera: Culicidae) in the Netherlands. Parasit. Vectors 8 (1), 1–9. https://doi.org/10.1186/s13071-015-0865-7.

Isherwood, F., Duffy, B., 1959. Resting *G. pallidipes* in the Lambwe Valley. East African Trypanosomiasis Research Organization, Kenya.

Isherwood, F., Duffy, B.J., Glasgow, J.P., Lee-Jones, F., Weitz, B., 1961. Further Studies of the Food of Tsetse Flies. J. Anim. Ecol. 30 (2), 259–266. https://doi.org/10.2307/2297.

Joshi, A., Miller, C., 2021. Review of machine learning techniques for mosquito control in urban environments. Eco. Inform. 61, 101241 https://doi.org/10.1016/J.ECOINF.2021.101241.

Jung Kjær, L., Soleng, A., Edgar, K.S., Lindstedt, H.E.H., Paulsen, K.M., Andreassen, Å.K., et al., 2019. Predicting the spatial abundance of *Ixodes ricinus* ticks in southern Scandinavia using environmental and climatic data. Sci. Rep. 9 (1), 1–16. https://doi.org/10.1038/s41598-019-54496-1.

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), 1361–1371. https://doi.org/10.1371/JOURNAL.PPAT.0030116.

Kaur, I., Sandhu, A.K., Kumar, Y., 2021. Analyzing and minimizing the effects of vector-borne diseases using machine and deep learning techniques: A systematic review. In: Proceedings of the IEEE International Conference Image Information Processing, 2021-November, pp. 69–74. https://doi.org/10.1109/ICIIP53038.2021.9702662.

Kaur, I., Sandhu, A.K., Kumar, Y., 2022. Artificial intelligence techniques for predictive modeling of vector-borne diseases and its pathogens: a systematic review. Arch. Comput. Methods En. 29 (6), 3741–3771. https://doi.org/10.1007/s11831-022-09724-9.

Keshavamurthy, R., Dixon, S., Pazdernik, K.T., Charles, L.E., 2022. Predicting infectious disease for biopreparedness and response: a systematic review of machine learning and deep learning approaches. One Health 15, 100439. https://doi.org/10.1016/J.ONEHLT.2022.100439.

Khun, M., 2022. Classification and Regression Training: Package caret. Available at: https://cran.r-project.org/web/packages/caret/caret.pdf (Accessed: 26 January 2023).

Laaksonen, M., Sajanti, E., Sormunen, J.J., Penttinen, R., Hänninen, J., Ruohomäki, K., et al., 2017. Crowdsourcing-based nationwide tick collection reveals the distribution of Ixodes ricinus and I. persulcatus and associated pathogens in Finland. Emerg. Microb. Infect. 6 (5) https://doi.org/10.1038/EMI.2017.17.

Lindh, J.M., Torr, S.J., Vale, G.A., Lehane, M.J., 2009. Improving the cost-effectiveness of artificial visual baits for controlling the Tsetse Fly Glossina fuscipes fuscipes. PLoS Negl. Trop. Dis. 3 (7), e474 https://doi.org/10.1371/JOURNAL.PNTD.0000474.

Lord, J.S., Torr, S.J., Auty, H.K., Brock, P.M., Byamungu, M., Hargrove, J.W., et al., 2018. Geostatistical models using remotely-sensed data predict savanna tsetse decline across the interface between protected and unprotected areas in Serengeti, Tanzania. J. Appl. Ecol. 55 (4), 1997–2007. https://doi.org/10.1111/1365-2664.13091.

Lukaw, Y.S., Abdelrahman, M.M., Mohammed, Y.O., Ochi, E.B., Elrayah, I.E., 2014. Factors influencing seasonal abundance of *Glossina fuscipes fuscipes (Glossina: Glossinidae)* in Kajo-Keji County, South Sudan. Curr. Res. J. Biol. Sci. 6 (6), 222–228. https://doi.org/10.19026/CRJBS.6.5197.

Makridou, T., Arvanitakis, G., Tsaprailis, K., Fornasiero, D., Kontoes, C., 2023. Understanding the Area of Applicability of Data Driven Mosquito Abundance Prediction Models. EGU23 [Preprint]. https://doi.org/10.5194/EGUSPHERE-EGU23-15398.

McCord, P.F., Messina, J.P., Campbell, D.J., Grady, S.C., 2012. Tsetse fly control in Kenya's spatially and temporally dynamic control reservoirs: a cost analysis. Appl. Geogr. 34, 189–204. https://doi.org/10.1016/j.apgeog.2011.11.005.

Mechan, F., Bartonicek, Z., Malone, D., Lees, R.S., 2023. Unmanned aerial vehicles for surveillance and control of vectors of malaria and other vector-borne diseases. Malar. J. 22 (1), 23. https://doi.org/10.1186/s12936-022-04414-0.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815 https://doi.org/10.1016/J.ECOLMODEL.2019.108815.

Mihok, S., Maramba, O., Munyoki, E., Kagoiya, J., 1995. Mechanical transmission of Trypanosoma spp. by African Stomoxyinae (Diptera: Muscidae). Trop. Med. Parasitol. 46 (2), 103–105. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8525279 (Accessed: 30 June 2019).

Muckley, E., Saal, J., Bryce Meredig, S., Roper, C., Martin, J., 2023. Interpretable models for extrapolation in scientific machine learning. Dig. Dis. https://doi.org/10.1039/D3DD00082F [Preprint].

Mudele, O., Frery, A.C., Zanandrez, L.F.R., Eiras, A.E., Gamba, P., 2021. Modeling dengue vector population with earth observation data and a generalized linear model. Acta Trop. 215, 105809 https://doi.org/10.1016/J.ACTATROPICA.2020.105809.

Mugenyi, A., Muhanguzi, D., Hendrickx, G., Nicolas, G., Waiswa, C., Torr, S., et al., 2021. Spatial analysis of G.f.fuscipes abundance in Uganda using Poisson and Zero-Inflated Poisson regression models. PLoS Negl. Trop. Dis. 15 (12), e0009820 https://doi.org/10.1371/JOURNAL.PNTD.0009820.

Ngari, N.N., Gamba, D.O., Olet, P.A., Zhao, W., Paone, M., Cecchi, G., 2020. Developing a national atlas to support the progressive control of tsetse-transmitted animal trypanosomosis in Kenya. Parasit. Vectors 13 (1), 286. https://doi.org/10.1186/s13071-020-04156-5.

Ngonyoka, A., Gwakisa, P.S., Estes, A.B., Nnko, H.J., Hudson, P.J., Cattadori, I.M., 2017a. Variation of tsetse fly abundance in relation to habitat and host presence in the Maasai Steppe, Tanzania. J. Vector Ecol. 42 (1), 34–43. https://doi.org/10.1111/JVEC.12237.

Ngonyoka, A., Gwakisa, P., Estes, A., Salekwa, L., Nnko, H., Hudson, P., et al., 2017b. Patterns of tsetse abundance and trypanosome infection rates among habitats of surveyed villages in Maasai steppe of northern Tanzania. Infect. Dis. Poverty 6 (1), 1–12. https://doi.org/10.1186/S40249-017-0340-0.

Omoogun, G.A., Dipeolu, O.O., Akinboade, O.A., 1989. Distribution and seasonal variation of tsetse population in the egbe area of Kwara state, Nigeria. Int. J. Trop. Insect Sci. 10 (05), 713–718. https://doi.org/10.1017/S174275840002186X.

Palaniyandi, M., Sharmila, T., Manivel, P., Thirumalai, P., Anand, P., 2021. Multispectral satellite data and GIS for mapping vector ecology, monitoring, risk assessment, and forecast of vector borne disease epidemics: a systematic review. Appl. Ecol. Environ. Sci. 9 (8), 751–760. https://doi.org/10.12691/AEES-9-8-6.

Palmer, J.R.B., Oltra, A., Collantes, F., Delgado, J.A., Lucientes, J., Delacour, S., et al., 2017. Citizen science provides a reliable and scalable tool to track disease-carrying mosquitoes. Nat. Commun. 8 (1), 1–13. https://doi.org/10.1038/s41467-017-00914-9.

Rahman, M.S., Pientong, C., Zafar, S., Ekalaksananan, T., Paul, R.E., Haque, U., et al., 2021. Mapping the spatial distribution of the dengue vector Aedes aegypti and predicting its abundance in northeastern Thailand using machine-learning approach. One Health 13. https://doi.org/10.1016/J.ONEHLT.2021.100358.

Rogers, D., 1979. Tsetse Population Dynamics and Distribution: A New Analytical Approach. J. Anim. Ecol. 48 (3), 825–849. https://doi.org/10.2307/4197.

Rogers, D.J., Randolph, S.E., 1986. Distribution and abundance of tsetse flies (Glossina Spp.). J. Anim. Ecol. 55 (3), 1007. https://doi.org/10.2307/4430.

Rogers, D., Randolph, S., 1988. Tsetse flies in Africa: bane or boon? Conserv. Biol. 2 (1), 57–65. https://doi.org/10.1111/J.1523-1739.1988.TB00335.X.

Rosà, R., Tagliapietra, V., Manica, M., Arnoldi, D., Hauffe, H.C., Rossi, C., et al., 2019. Changes in host densities and co-feeding pattern efficiently predict tick-borne encephalitis hazard in an endemic focus in northern Italy. Int. J. Parasitol. 49 (10), 779–787. https://doi.org/10.1016/J.IJPARA.2019.05.006.

Schneider, J., Greco, A., Chang, J., Molchanova, M., Shao, L., 2022. Predicting West Nile Virus Mosquito Positivity Rates and Abundance: A Comparative Evaluation of Machine Learning Methods for Epidemiological Applications. Authorea Preprints [Preprint].. https://doi.org/10.1002/ESSOAR.10509422.1.

Shutt, D.P., Goodsman, D.W., Martinez, K., Hemez, Z.J.L., Conrad, J.R., Xu, C., et al., 2022. A process-based model with temperature, water, and lab-derived data improves predictions of daily culex pipiens/restuans mosquito density. J. Med. Entomol. 59 (6), 1947–1959. https://doi.org/10.1093/JME/TJAC127.

Signaboubo, D., Payne, V.K., Moussa, I.M.A., Hassane, H.M., Berger, P., Kelm, S., et al., 2021. Diversity of tsetse flies and trypanosome species circulating in the area of Lake Iro in southeastern Chad. Parasit. Vectors 14 (1), 293. https://doi.org/10.1186/s13071-021-04782-7.

Speiser, J.L., 2021. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J. Biomed. Inform. 117, 103763 https://doi.org/10.1016/J.JBI.2021.103763.

Speiser, J.L., Miller, M.E., Tooze, J., Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst. Appl. 134, 93–101. https://doi.org/10.1016/J.ESWA.2019.05.028.

Taconet, P., Porciani, A., Soma, D.D., Mouline, K., Simard, F., Koffi, A.A., et al., 2021. Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso. Parasit. Vectors 14 (1). https://doi.org/10.1186/S13071-021-04851-X.

Talbot, B., Slatculescu, A., Thickstun, C.R., Koffi, J.K., Leighton, P.A., McKay, R., et al., 2019. Landscape determinants of density of blacklegged ticks, vectors of Lyme

disease, at the northern edge of their distribution in Canada. Sci. Rep. 9 (1), 1–12. https://doi.org/10.1038/s41598-019-50858-x.

Vale, G.A., Hursey, B.S., Hargrove, J.W., Torr, S.J., Allsopp, R., 1984. The use of small plots to study populations of tsetse (Diptera: Glossinidae): difficulties associated with population dispersal. Int. J. Trop. Insect Sci. 5 (5), 403–410. https://doi.org/10.1017/S1742758400008730.

Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W.W.L., Edgar, G.J., Mouillot, D., et al., 2022. A quantitative review of abundance-based species distribution models. Ecography 2022 (1). https://doi.org/10.1111/ECOG.05694.

Williams, B., Dransfield, R., Brightwell, R., 1992. The control of tsetse flies in relation to fly movement and trapping efficiency. J. Appl. Ecol. 29 (1), 163. https://doi.org/10.2307/2404359.

World Health Assembly, 2017. Global Vector Control response 2017–2030: An integrated approach for the control of vector borne diseases. World Health Organization. Available at: https://iris.who.int/handle/10665/275708 (Accessed: 5 September 2023).

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1) https://doi.org/10.18637/JSS.V077.I01.

Xu, G., Mather, T.N., Hollingsworth, C.S., Rich, S.M., 2016. Passive surveillance of Ixodes scapularis (say), their biting activity, and associated pathogens in Massachusetts. Vector Borne Zoonotic Dis. 16 (8), 520. https://doi.org/10.1089/VBZ.2015.1912.

Yu, Z., Wang, K., Wan, Z., Xie, S., Lv, Z., 2022. Popular deep learning algorithms for disease prediction: a review. Clust. Comput. 26 (2), 1231–1251. https://doi.org/10.1007/S10586-022-03707-Y.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., et al., 2021. ESA WorldCover 10 m 2020 v100. https://doi.org/10.5281/ZENODO.5571936.