Distribution of the Habitat Suitability of the Main Malaria Vector in French Guiana Using Maximum Entropy Modeling
Yi Moua, Emmanuel Roux, Romain Girod, Isabelle Dusfour, Benoit De Thoisy, Frédérique Seyler, Sébastien Briolant

To cite this version:
Yi Moua, Emmanuel Roux, Romain Girod, Isabelle Dusfour, Benoit De Thoisy, et al.. Distribution of the Habitat Suitability of the Main Malaria Vector in French Guiana Using Maximum Entropy Modeling. Journal of Medical Entomology, Entomological Society of America, 2016, <10.1093/jme/tjw199>. <hal-01425541>

HAL Id: hal-01425541
https://hal.archives-ouvertes.fr/hal-01425541
Submitted on 6 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Journal of Medical Entomology

**Distribution of the habitat suitability of the main malaria vector in French Guiana using Maximum Entropy modeling**

Yi Moua1, Emmanuel Roux2, Romain Girod3, Isabelle Dusfour3, Benoit de Thoisy4, Frédérique Seyler2, Sébastien Briolant3,5,6,7

1 Université de Guyane, ESPACE-DEV, UMR 228 (IRD, UM, UR, UA, UG), Cayenne, French Guiana

2 Institut de Recherche pour le Développement, ESPACE-DEV, UMR 228 (IRD, UM, UR, UA, UG), Montpellier, France

3 Unité d’Entomologie Médicale, Institut Pasteur de la Guyane, Cayenne, French Guiana

4 Laboratoire des Interactions Virus Hôtes, Institut Pasteur de la Guyane, Cayenne, French Guiana

5 Direction Interarmées du Service de Santé en Guyane, Cayenne, French Guiana

6 Institut de Recherche Biomédicales des Armées, Unité de Parasitologie et d’Entomologie Médicale, Marseille, France

7 Unité de Recherche en Maladies Infectieuses Tropicales Emergentes, UMR 63, CNRS 7278, IRD 198, INSERM 1095, Faculté de Médecine La Timone, Marseille, France
Abstract

Malaria is an important health issue in French Guiana. Its principal mosquito vector in this region is *Anopheles darlingi*. Knowledge of the spatial distribution of this species is still very incomplete due to the extent of French Guiana and the difficulty to access most of the territory.

Species Distribution Modeling based on the maximal entropy procedure was used to predict the spatial distribution of *An. darlingi* using 39 presence sites. The resulting model provided significantly high prediction performances (mean 10-fold cross-validated partial AUC and continuous Boyce index equal to, respectively, 1.11 – with a level of omission error of 20 % – and 0.42). The model also provided a habitat suitability map and environmental response curves in accordance with the known entomological situation.

Several environmental characteristics that had a positive correlation with the presence of *An. darlingi* were highlighted: non-permanent anthropogenic changes of the natural environment; the presence of roads and tracks; opening of the forest. Some geomorphological landforms and high altitude landscapes appear to be unsuitable for *An. darlingi*.

The Species Distribution Modeling was able to reliably predict the distribution of suitable habitats for *An. darlingi* in French Guiana. Results allowed completion of the knowledge of the spatial distribution of the principal malaria vector in this Amazonian region, and identification of the main factors that favor its presence. They should contribute to the definition of a necessary targeted vector control strategy in a malaria pre-elimination stage, and allow extrapolation of the acquired knowledge to other Amazonian or malaria-endemic contexts.

Keywords: Maxent, species distribution model, presence-only, *Anopheles darlingi*, sampling bias
Malaria is a public health issue in the Amazonian region, with major transmission foci depending on specific local characteristics associated with changing environmental and socio-demographic contexts. French Guiana is a French overseas territory with ~260,000 inhabitants. It remains one of the major malaria foci in the region, despite an improving epidemiological situation during the past ten years. The number of reported clinical cases has significantly dropped from 4,479 in 2005 to 434 in 2015 (Petit-Sinturel et al. 2016), and now corresponds to an incidence rate of two cases for 1,000 inhabitants for the whole territory, making it possible to target the pre-elimination of the disease in 2018 (Agence Régionale de Santé Guyane 2015). *Plasmodium vivax* is at present predominant and this species was responsible for 67% of the diagnosed cases of malaria in the territory in 2014, the others being mainly due to *Plasmodium falciparum* (Musset et al. 2014, Ardillon et al. 2015). However, this epidemiological situation is heterogeneous in space and time. In particular, a recrudescence of malaria cases is currently observed in the inland region (Saül, Cacao, and Régina) and eastern French Guiana (municipalities of Camopi and Saint-Georges-de-l'Oyapock), with a general incidence rate reaching 55.2 cases per 1,000 inhabitants in 2013 (Musset et al. 2014), likely due to the emergence and/or persistence of local foci of high malaria transmission (Berger et al. 2012, Musset et al. 2014). This Amazonian region, especially near the international borders, includes vulnerable populations. Some are hard-to-reach and have poor access to health services and treatment-seeking behaviors that may favor the development of resistance to antimalarial drugs (Musset et al. 2014, Wangdi et al. 2015). Uncontrolled areas of malaria transmission are also prevalent in illegal gold mining areas (Pommier de Santi et al. 2016a, Pommier de Santi et al. 2016b). The epidemiological situation remains quite unstable, and pre-elimination of malaria, corresponding to an incidence rate below one case for 1,000 inhabitants in any locality of French Guiana, remains a major challenge.

In this context, public health authorities must maintain control efforts while targeting them more precisely and objectively in time and space (Alimi et al. 2015). A map of malaria risk in French
Guiana is updated regularly by the regional unit of the French National Public Health Agency, based on the number of cases reported per locality and the data available on movements of human populations at risk, especially due to gold mining activities. This map is validated by the local expert committee of epidemic diseases (Comité d'Experts des Maladies à Caractère Épidémique, CEMCE), which brings together different experts of the disease in the region (from the Health Surveillance Agency, the Pasteur Institute of French Guiana, the Regional Unit of the National Public Health Agency, vector control services, hospitals, and other diagnosis and care centers, and the Defense Health Service in French Guiana). The lack of objective knowledge of several key factors, especially the spatiotemporal distribution of the main malaria vectors and human populations infected by *Plasmodium* and/or carrying gametocytes, makes such a map highly approximate.

Anopheles (Nyssorhynchus) darlingi Root (Diptera: Culicidae) is one of the most efficient malaria vectors in South America and is considered to be the primary malaria vector in French Guiana because of its anthropophilic behavior, natural infectability, high density, and sensitivity to *P. falciparum* (Girod et al. 2008, Hiwat et al. 2010, Fouque et al. 2010). Used entomological data collection for the entire territory, for the mapping of entomological risk indicators at the regional scale, is not feasible. French Guiana occupies a large territory (84,000 km²) which is mostly covered by rain forest (more than 80%) and highly inaccessible. Knowledge of the recent geographical distribution of *An. darlingi* is thus restricted to coastal areas, some villages along the international border rivers, and some illegal gold mining sites (Figure 1).

Species Distribution Modeling (SDM) offers an efficient solution to geographically extrapolate such knowledge to the entire territory (Pearson et al. 2007). Species Distribution Modeling produces maps of species habitat suitability by using known presence locations of the species and relevant environmental data. The use of SDM is thus encouraged to “improve and facilitate the development of alternative vector control strategies” (Alimi et al. 2015). Numerous SDM approaches are
proposed in the literature. Some of them, such as Maximum Entropy (Maxent; Phillips et al. 2006),
Genetic Algorithm for Rule-Set Prediction (GARP; Stockwell 1999), Boosted Regression Trees
(BRT; Friedman et al. 2000), Generalized linear and additive models (GLM and GAM; Guisan, et
al. 2002), and Multivariate adaptive regression splines (MARS; Leathwick et al. 2005), exploit only
species presence information, offering a significant advantage over methods that also require
absence data. Indeed, absence data are often difficult to obtain. According to Peterson (2007) and
Hirzel et al. (2002), absence can result from (1) the non-detection of the species in a suitable
habitat, even if it is present, (2) the actual absence of the species for historical reasons, whereas the
habitat is suitable, and (3) the true absence of the species and the unsuitability of the habitat.
2006, Wisz et al. 2008) show that Maxent is able to fit complex functions between habitat suitability
and predictor variables, is the least sensitive to the size of the presence dataset, and tends to
outperform other comparable methods when the dataset is small.
In this study, the mapping of the habitat suitability of *An. darlingi* at the scale of all of French
Guiana was performed using the Maxent SDM approach. This work aims to provide reliable maps
for improving malaria transmission risk mapping in French Guiana, and to identify the
environmental factors and associated mechanisms that favor the presence of *An. darlingi*.

**Materials and Methods**

**Study area**

French Guiana (84,000 km²), a French overseas region located in South America, is separated from
Suriname by the Maroni River and from Brazil by the Oyapock River and the Tumuc-Humac
mountains. More than 80% of the territory is covered by rain forest. The country has an equatorial
climate characterized by two annual dry seasons, from mid-August to mid-November and in March,
and two wet seasons, from mid-April to mid-August and mid-November to February. The average
annual rainfall reaches 4,000 mm and 2,000 mm in the wettest (north-east) and driest (north-west)
areas, respectively (Hammond 2005). The average monthly rainfall is >100 mm for the entire territory throughout the year, except for the three driest months: September, October, and November (Héritier, 2011). The average humidity is between 80% and 90%. The temperature is homogeneous over the entire territory throughout the year, with an average annual temperature of 26°C. The difference between the minimum and maximum daily temperature is more important than the annual variations. For example, in Maripasoula (on the border with Surinam) and Camopi (on the border with Brazil), the annual ranges of the minimum and maximum temperatures were, 4.3°C and 9.6°C (averages over the period 2001-2008), respectively, whereas the mean daily thermal amplitude was 9.8°C (average over the period 2001-2008; Météo-France, 2016). The population of ~260,000 inhabitants is unequally distributed throughout the territory. Approximately 90% of the population lives in the coastal area and most of the rest lives along the Maroni and the Oyapock rivers (Amerindians and Bush-Negroes). However many people live and/or transit through inland and remote areas of the territory (forestry workers, gold miners, and soldiers). According to many studies (Berger et al. 2012, Verret et al. 2006, Queyriaux et al. 2011, Hustache et al. 2007, Stefani et al. 2011, Pommier de Santi et al. 2016a), Amerindians, gold miners, and soldiers may be highly infected by malaria, whereas the areas in which they live and/or transit are those with the poorest knowledge of the presence and density of malaria vectors. It is thus of potential interest to consider the malaria risk, study the distribution of malaria vectors, and implement prevention and control actions over the entire territory of French Guiana.

**Species Records**

Presence sites of *An. darlingi* were provided by surveys of the Medical Entomology Unit of the Pasteur Institute of French Guiana and the Defense Health Service in French Guiana. *Culicidae* collections were performed using either human landing catches or traps (light traps or odor baited traps). Human landing catches consisted of exposing collector's lower leg and collecting landing mosquito with a mouth aspirator. Collectors were members of the Pasteur Institute or Defense
Health Services, they were aware of the risks associated with the method and had given their free consent. Malaria prophylaxis was proposed and information on the medication was explained. Light trap catches were performed with Center for Disease Control and Prevention (CDC) light traps, and odor baited catches were performed with Mosquito Magnet ® traps (Woodstream Corporation, Lititz, PA) baited with Octenol, a combination considered to be the best candidate for Anopheles surveillance in the region (Vezenegho et al., 2015).

Anopheles species were morphologically identified using taxonomic keys specific for the region (Floch and Abonnenc 1951, Faran and Linthicum 1981, Forattini 1962). Only Culicidae collections performed since the year 2000 were precisely geolocated by GPS coordinates and were used for the study (Figure 1). These data correspond to 74 capture sites for the family Culicidae, and to 48 presence sites for the species An. darlingi.

The difficulty in accessing most of the French Guiana territory, and the priority given to the areas at risk of malaria transmission where many people live, led to a significant sampling bias with oversampling of the anthropized region of the territory, notably those easily accessible by roads (Figure 1).

**Ecological knowledge and hypotheses**

The presence of An. darlingi is linked to compositional and configurational features of the land cover and land use, as they partially determine breeding, feeding, and resting sites of the vector (Stefani et al. 2013). The natural environment for this vector in the Amazonian region includes floodable savanna, swamps (Girod et al. 2011, Zeilhofer et al. 2007), and flooded forest (Rozendaal 1992). Larvae are found along river edges, on flooded riverbanks, creeks, and pools formed near river-beds (Rozendaal 1992, Hiwat et al. 2010). Breeding sites are generally situated at low altitude (Mouchet 2004) and solely in freshwater, as An. darlingi is sensitive to salinity (Deane et al. 1948). Hydrological and geomorphological factors are responsible for the formation and destruction of Anopheles breeding sites (Smith et al. 2013).
Human activities, comprising deforestation and fish farming, also contribute to the creation of active breeding sites (Patz et al. 2000, Richard 1987, Stefani et al. 2013, Takken et al. 2005, Terrazas et al. 2015, Vittor et al. 2006, Vittor et al. 2009). Unpaved roads, tracks, and culverts form ideal breeding sites for *An. darlingi* in the Amazon region (Singer and Castro 2001). The presence of *An. darlingi* is also maintained by regular human presence due to its strong anthropophilic behavior. However, the presence and density of *An. darlingi* can either be favored or restricted depending on the type and intensity of the anthropogenic impacts. Stefani et al. (2013) systematically reviewed the literature and showed that all the studies describe the same mechanisms linking deforestation, land use, and the degree of urbanization with malaria transmission risk in the Amazonian region: opening the forest and maintaining a high degree of interaction between forested and deforested areas decreases the distance between feeding, breeding, and resting sites of *An. darlingi*, favoring the presence and high density of the vector (as well as a high probability of contact between humans and vectors); in contrast, intensifying deforestation and creating large urbanized and/or cultivated surfaces tends to decrease suitable habitat for *An. darlingi*. These two antagonistic consequences of human activities were considered in the SDM described here, by explicitly separating favorable and unfavorable factors in the environmental characterization.

The optimum temperature range for *An. darlingi* is between 20 and 30°C with a humidity of above 60% (Martens et al. 1995). Several studies established a minimal monthly rainfall threshold to designate suitable breeding habitats for *Anopheles* (reviewed in Smith et al. 2013). These values vary between 10 and 80 mm and need to be maintained for three or four months.

**Environmental Variables**

Environmental variables chosen as SDM inputs must characterize the ecological factors that influence the presence of *An. darlingi*, previously described. These factors are separated into three types: 1) natural environment features, associated with land cover, land use, and geomorphology for which the impact on the presence of *An. darlingi* depends on specific values or categorical classes;
2) anthropogenic activities that non-permanently alter the natural environment on a highly local scale and favor the presence of *An. darlingi*; 3) urbanization, corresponding to human presence and activities that permanently alter the natural environment over large areas and hinders the presence of the vector. Meteorological variables were not included in the model, because the temperature, rainfall, and humidity fall within the optimal ranges for presence of the species in French Guiana. Thus, these variables cannot significantly explain differences in the time average habitat suitability distribution over the year (this point is extensively discussed in the Discussion section).

**Raw Geographic Data.** Variables chosen as SDM inputs were derived from the following raw geographic data:

- Geomorphological landscape (*GLS*) and Geomorphological landforms (*GLF*) from the French Forest Office (ONF) (Guitet et al. 2013);
- Landscape types (*LS*) from the French Agricultural Research Centre for International Development (CIRAD) (Gond et al. 2011). This provides the distribution of landscape types in French Guiana, most being forested landscapes;
- Altitude (*ATL*) derived from the Digital Elevation Model provided by the Shuttle Radar Topography Mission (SRTM, spatial resolution: 30 meters) of the United States Aeronautics and Space Administration (NASA);
- Human footprint (*HFP*): An integrated human activity index that gives a general measure of the extent of expected threats on biodiversity, by assigning a score depending on the nature of the disturbance. It combines sublayers spatializing human population density, urban areas, legal and illegal mining sites, agriculture, forest settlements and camps, tourist camps, logged areas (forest activities), and potential hunting areas corresponding to a zone of two kilometers around roads, tracks and rivers, likely to be used by humans. The total disturbance score is the sum of all human activity scores (de Thoisy et al. 2010);
- Roads and tracks from the BD TOPO® database of the French Institute of Geographical
Table 1 summarizes the main features of these raw geographic data.

**Definition of Environmental Variables Used as Inputs for SDM.** Several variables were extracted from the previously described raw data to better reflect the ecological knowledge and hypotheses mentioned above. The reference spatial resolution (pixel size) permitting the integration of all environmental layers was set to 1 by 1 km, *i.e.*, the coarsest resolution of the available layers, associated with the LS map.

The length of roads and tracks outside of urban areas (*ROADS*) was computed in the 1 km-cell grid from the BD TOPO® database.

The sublayers composing the *HFP* were first rasterized into 30-m grid cells, the smaller polygon of the *HFP* having a size of approximately 40 by 10 m. Distinct attributes were then extracted:

- The percentage of urbanization (*PER_URB*) within the 1 km grid cells;
- The percentage of urbanization within the eight neighbor cells of each urban cell (*PER_URB_NEIGH*), which permits distinguishing small from large urban areas. This layer was obtained for each 1 km-cell considered to be urban (*i.e.*, with *PER_URB* $\geq 50$%), by averaging the *PER_URB* values for the eight (1 km side) neighbor pixels;
- The human activities which non-permanently alter the natural environment (*HA*), by first summing the scores of the following sublayers: tourist and forest camps, mining activities and logged areas, hunting areas nearby rivers, and then, by computing the minimum, median, and maximum values within the 1-km grid cells.

The agriculture sublayer from *HFP* was not used because it covers only the coastal area. The population density sublayer was also excluded because it did not have sufficient level of detail. The sublayer of potential hunting areas near roads and tracks were not used to avoid duplication of the length of roads and tracks outside of urban areas computed previously.

For each 1-km grid cell, the majority class of the categorical variables *GLS* and *GLF*, and the
minimum, median, and maximum altitude (ALT) values were computed.

Eventually, some corrections of the LS layer were performed as it did not identify urban areas and did not distinguish flooded forests associated with freshwater from those of the coastal strip associated with brackish water (mangroves): LS cells with an PER_URB value greater than or equal to 50% were reclassified into a new LS class referred to as Urban; LS cells classified as Flooded forest and corresponding to mangroves according to the coastal land use map provided by the ONF (Office National des Forêts Direction Régionale de Guyane, 2013) were reclassified as Mangrove. The variable PER_URB was excluded from the input SDM variables, as the urban areas were mapped, and their extent quantified, by the corrected LS and PER_URB_NEIGH layers, respectively.

Table 1 lists and describes the environmental variables used to build the model.

Maxent Model Principle

Maxent is an SDM which requires environmental variables and species presence-only data. It is based on the principle of maximum entropy to estimate an (a priori) unknown probability distribution over the entire study area. This probability distribution assigns a value that is proportional to the probability of the presence of the species to each pixel of the study area. It is therefore interpreted as a habitat suitability index (HSI) across the study area (Phillips et al. 2006). The Maximum Entropy principle consists of approximating the unknown probability distribution by finding the one that maximizes entropy and satisfies the constraints imposed by the environmental features at the known sites of presence. Environmental features are a set of input environmental variables chosen according to their expected relevance for the studied taxon (Phillips et al. 2006, Elith et al. 2011). The constraints ensure that the environmental values expected under the approximated probability distribution are consistent with environmental information observed at the presence points.

In practice, the Maxent distribution is defined on a set of points called background points. These
points should reflect the available environmental conditions of the study area and are chosen by uniform random sampling. This approach assumes that the presence data are not biased and that environmental conditions are uniformly sampled (Yackulic et al. 2013). However, in practice, some areas are more intensively sampled than others, and environmental conditions are not uniformly distributed and may imply a strong sampling bias. Phillips et al. (2009) proposed selecting the background points with the same environmental bias as the presence dataset to correct the effect of this sampling bias.

Model Building and Evaluation

Eleven environmental variables and 48 An. darlingi presence points (their coordinates were in the table in supplementary material S1) are used as inputs for Maxent. Only one presence site was selected to build the model when more than one occurred in the same pixel. As a result, only 39 presence sites were actually used for building the model. Hinge and categorical features were selected for the environmental variables. A hinge feature provides a good compromise between simplicity and the quality of the approximation of the species response curves (Elith et al. 2011, Phillips and Dudík 2008).

In this study, the distribution of the background points was biased so that the selection bias corresponds to that of the sampling. The sampling bias was defined as the relative sampling effort in the environmental space, and was estimated by considering the capture locations of Culicidae, obtained using the same capture techniques and supposed to be subjected to the same sampling bias as the An. darlingi species. The details of the method to create the relative sampling effort map are described in supplementary material S2.

The model was computed using version 3.3.3k of Maxent. The recommended values derived from Phillips and Dudík (2008) concerning the regularization parameters and the background set size, were applied. Regularization parameter values were set to 0.25 and 0.5 for categorical and hinge features, respectively, and the size of the background was set to 10,000. The extrapolation option
was not selected to avoid making predictions in environmental domains in which the model was not trained. The model was fitted using the full data set and evaluated using a 10-fold cross-validation procedure. The Receiver Operating Characteristic (ROC) curves and the associated Areas Under the ROC Curve (AUC) were computed. This was completed by computing the mean partial AUC ratios (Peterson et al., 2008), consisting of the ratios of the partial AUCs of the model over the null AUC (corresponding to random prediction), for omission errors ($E$) of 20, 10, and 5%. The Continuous Boyce Index (CBI), considered to better adapted to presence-only models than the AUC (Hirzel et al., 2006), was also computed. The gain (regularized training gain) was also used to evaluate the performance of the model prediction. It is a measure of the likelihood of the sample, and indicates how much better the estimated distribution fits the presence points than the uniform distribution, which corresponds to a null gain (Yost et al. 2008).

The importance of each variable was estimated using two methods, a heuristic method and the jackknife test. The heuristic method computes the percentage contribution of each variable to the model. During the training process, the increase of the gain is due to the adjustment of the feature weights and this increase is assigned to the environmental variable that the feature depends on. The sum of these increases in gain indicates the percentage contribution of each environmental variable. The jackknife test evaluates the individual contribution of each variable to the model by estimating the difference of the gain when removing each variable, one by one, and when considering the given variable alone to build the model.

**Results**

The mean AUC was 0.93, and the mean partial AUC ratios were 1.08, 1.03, and 1.01 for maximum omission errors sets to 20, 10, and 5% respectively. The mean CBI was 0.356 and the mean gain was 3.14. Three variables cumulatively contributed >80% (Table 2): the length of roads and tracks outside of urban areas ($ROADS$), the percentage of urbanization of neighboring pixels ($PER\_URB\_NEIGH$), and landscape ($LS$). The maximum value of the human activities which non-
permanently alter the natural environment (\(HA_{\text{MAX}}\)), geomorphological landscape (\(GLS\)), minimum altitude (\(ALT_{\text{MIN}}\)), and geomorphological landform (\(GLF\)) contributed moderately to the model, with contributions of 6.84\%, 5.35\%, 1.34\%, and 1.19\%, respectively. The following input variables contributed very little to the model: minimum and median values of human activities which non-permanently alter the natural environment (\(HA_{\text{MIN}}\) and \(HA_{\text{MED}}\); 0.35 and 0.24\%, respectively); and median and maximum values of altitude (\(ALT_{\text{MED}}\) and \(ALT_{\text{MAX}}\); 0.69 and 0.06\%, respectively).

The results of the Jackknife test confirmed the non-significant contribution of the input variables \(HA_{\text{MIN}}\), \(HA_{\text{MED}}\), \(ALT_{\text{MED}}\), and \(ALT_{\text{MAX}}\) (Table 2).

A second model was built using only the most highly contributing environmental variables: \(ROADS\), \(LS\), \(PER_{\text{URB\_NEIGH}}\), \(HA_{\text{MAX}}\), \(GLS\), \(GLF\), and \(ALT_{\text{MIN}}\). The overall performance of this simpler model was very similar to the previous one, with the mean AUC and partial AUC ratios equal to 0.93 and 1.11, 1.05, and 1.03, respectively. The mean gain was equal to 3.19 and the mean CBI was 0.421. Relative contributions of the input variables were also very similar (Table 3). The response curves of the environmental variables are represented in Figures 2 and 3. They show that the HSI is maximal when the \(PER_{\text{URB\_NEIGH}}\) is below 8\%. Above this value, the HSI decreases progressively towards 0. The HSI increases as \(ROADS\) increases up to 7,000 meters, reaches a plateau value, and then tends to decrease above 10,000 meters. Among all \(LS\) classes, Woodland savanna/dry forest and Open forest contribute the most to the high HSI values. The geomorphological landscape classes Coastal flat plain and Plain with residual relief and the geomorphological landform classes Small-size and flat wet land, Small-size rounded hill, and Lowered half-orange relief – a tropical relief type corresponding to a hill with convex flanks giving to it a roughly hemispherical shape (George, 1972) and usually linked to flat or swampy lowlands drained by streams with meanders – are also associated with high HSI values. The HSI is maximal when \(ALT\) is ~0, with a rapid decrease as altitudes increase. The \(HA_{\text{MAX}}\) response curve presents a
more complicated profile. The HSI increases for \( HA_{\text{MAX}} \) values between 0 and 8, decreases until \( HA_{\text{MAX}} \) reaches 24, and then again increases as values continue to climb above 24.

The map of habitat suitability for \textit{An. darlingi}, based on all the presence data for modeling, shows six main areas (A – F) with a high HSI and a seventh area (G) corresponding to an epidemiological interest area (see Figure 4). A qualitative analysis was performed to determine the characteristics of the environmental variables of the areas with high HSI values (Table 4).

In the coastal area (A), where 90% of the Guyanese population lives, the HSI tends to be higher along the main road representing the main traffic route in French Guiana. Focusing on the main urban areas, represented in Figure 5, the HSI values within the highly urbanized districts of Cayenne and Kourou (rectangles in Figure 5) are lower than those of the surrounding pixels that are not considered to be highly urbanized. A very high HSI was predicted within the urban area of Saint-Laurent-du-Maroni. However, none of the pixels characterizing this city has a \( PER_{\text{URB} \_\text{NEIGH}} \) value higher than or equal to 50%. The high HSI values in areas B, D, E, and G are characterized by the environmental variables \textit{ROADS}, \( HA_{\text{MAX}} \), the classes \textit{Open forest} and \textit{Mixed high and open forest}, and flat or moderately hilly terrain. The high HSI in areas C and F is essentially linked to \textit{Open forest} and flat terrain.

The areas for which the model did not predict the HSI, due to the choice to not extrapolate to environmental domains not used to train the model, correspond to areas with an altitude higher than 400 meters. They represent a small number of pixels of the study area.

**Discussion**

The prediction performances of the model are excellent and significantly greater than those of the null model. The following discussion focuses on the ecological interpretation of the results and the methodological choices and alternatives.

**Environmental Factors Explaining the Habitat Suitability**

The geographic distribution of habitat suitability is consistent with existing knowledge of the
entomological situation despite the small number of presence points. The high HSI values can be explained by different environmental contexts depending on the geographical locations. In most areas (A, B, D, and E), the HSI values depend on human presence and activities, characterized by the environmental variables $HA_{MAX}$ and $ROADS$ (in areas D, E, and B, most roads are not paved and correspond mostly to tracks). The significantly positive correlation between the variable $ROADS$ and the HSI confirms that road and track opening, accompanied by deforestation and pooling of rainwater at the roadside, may favor breeding sites (Singer and Castro 2001). The response curve for the variable $ROADS$ (Figure 3) reaches a plateau above 7,000 meters of road per square kilometer and decreases thereafter. The decrease of the HSI at values above 7,000 meters suggests that the density of the road network leads to an improvement of the road quality (paved road eliminating culverts, adding sidewalks), thus limiting the availability of breeding and/or resting sites, in the same way as urbanization. Indeed, the response curve of the $PER_{URB\_NEIGH}$ variable confirms that highly urbanized areas provide a poorly suitable habitat for An. darlingi (Figure 3). Intensive urbanization implies concrete paving, the decrease or removal of green areas and forests, and consequently, the destruction of breeding and resting sites for An. darlingi (Stefani et al. 2013). This phenomenon is observed in the highly urbanized areas of Cayenne and Kourou (Figure 5). In contrast, Saint-Laurent-du-Maroni, the second largest urban area of French Guiana in terms of urbanization size and density, has high HSI values. In fact, unlike Cayenne and Kourou, this area is not considered to be highly urbanized using the criterion of this study ($PER_{URB\_NEIGH} \geq 50\%$). However, the result for Saint-Laurent-du-Maroni seems unlikely because the presence of An. darlingi has not yet been reported in an urban area. Further field works could confirm the presence of this species in the city. The sensitivity of the model for the criterion that defines a highly urbanized area may also merit further study.

The values of $HA_{MAX}$ in areas D and E were essentially associated with mining activity. In French Guiana, this activity is responsible for forest loss reaching 2,000 hectares per year (Office National
des Forêts Direction Régionale de Guyane, 2014). Between 2001 and 2013, Alvarez-Berríos and Aide (2015) estimated that the largest forest loss due to gold mining in the tropical and subtropical moist forest in South America was situated in the Guianan region including French Guiana. This suggests that this activity, resulting in deforestation and creating sources of standing water such as mining pits, combined with the presence of a large number of people, creates suitable conditions for *An. darlingi*. The high HSI in these two areas is also explained by the *Mixed high and open forest* landscape which is associated with human disturbance (Gond et al. 2011). Indeed, this landscape is described as a forest environment linked to young or unstable vegetation mostly due to first stages of anthropization. These results confirm the important role of human presence in the creation of suitable habitats for *An. darlingi*, which is also consistent with the strong anthropophilic behavior of this vector.

Some landscape types which are not directly associated with human presence or activities were also associated with a HSI. The *Woodland savanna/dry forest* class appears to highly contribute to high HSI values (Figure 2). It corresponds to the driest landscape in French Guiana (Gond et al. 2011), but can be seasonally inundated due to its poor drainage, creating breeding sites (Rosa-Freitas et al. 2007). The high HSI values in this area are in accordance with previous studies (Vezenegho et al. 2015, Dusfour et al. 2013), which reported finding *An. darlingi* in the coastal savanna environments of French Guiana. In uninhabited areas (zones F and C in Figure 4), a high HSI is associated with the *Open forest* class (LS layer) and flat terrain. This LS class can be associated with different land cover types in French Guiana (Gond et al. 2011) depending on the geographical location.

Consequently, this LS class may differentially affect *An. darlingi* habitat suitability. The *Open forest* in area C mainly corresponds to wetlands (classified as *Flooded forest* according to the coastal land use map provided by the Office National des Forêts Direction Régionale de Guyane, 2013), whereas in area F, it corresponds to *Large surfaces of bamboo thicket and forbs*. *Anopheles darlingi* was found in flooded forest; however, to our knowledge, no information is available concerning its
presence in large areas of bamboo thicket and forbs. The prediction in these areas should be taken
with precaution as a more precise description of the habitats within *Open forest* class is required.
Overall, this information highlights that natural environment could form highly suitable habitats
despite the high anthropophily of *An. darlingi*.

**Meteorological Variables**

In this study, meteorological variables were not used to build the model. Temperatures fall within
the optimal range for the species presence, and were considered to be geographically and
temporally too homogeneous to explain differences in the spatial distribution of habitat suitability.
Such a hypothesis is common in the Amazonian context. Olson et al. (2009) report that in their
study region (Amazon basin), “monthly temperatures were between 24.6°C and 29.4°C (well within
the range for optimal malaria transmission) for 95% of the observations,” and consequently did not
include temperatures in their model. In French Guiana, several studies also used rainfall data to
study the intra-annual variations in *An. darlingi* density (Hiwat et al., 2010, Girod et al., 2011). The
exclusion of rainfall data is more debatable, as rainfall clearly influenced the intra-annual density of
*An. darlingi* in the study region (Hiwat et al., 2010, Girod et al., 2011, Vezenegho et al., 2015) even
if the relationship was not systematically observed (Girod et al., 2011). The evidence for this impact
on densities is that *An. darlingi* habitat suitability varies at an intra-annual scale, due to the
alternation of dry and wet seasons. However, the entire study area is subject to this alternation.
Moreover, given the high density of the French Guiana hydrological network and that the driest area
(north-west) still receives 2,000 mm a year, it can be reasonably assumed that *An. darlingi* can find
suitable conditions within the entire territory throughout most of the year. In French Guiana, the
geomorphological landscape highly influences the availability of breeding sites, and therefore their
spatial distribution, whereas the rainfall quantities influence the intra-annual variations of *An.
darlingi* densities. As a consequence, on an average over the year, we assume that the significant
factor influencing the distribution of habitat suitability is not the quantity of rainfall, but the
capacity of the landscape to provide suitable breeding sites when it rains.

**Model Parametrization**

The model was run by using the regularization parameter values and background set size recommended by Phillips and Dudík (2008), instead of those determined from specific experiments, as suggested by Merow et al. (2013). Phillips and Dudík (2008) tested a set of regularization parameter values with 48 species datasets that contained 11 to 13 environmental variables and a small number of categorical variables (1-3, as they considered discrete ordinal variables to be categorical). Nine of these datasets contained between 30 and 60 occurrences. The characteristics of the dataset exploited in our study (39 occurrence records; 13 and seven environmental variables including three categorical ones) are assumed to be quite similar of those of the datasets used by Phillips and Dudík (2008). We thus assumed that the pseudo-optimal parameters proposed by Phillips and Dudík (2008) could be confidently used in our study. Similarly, the background size was set to 10,000 based on the tests realized by Phillips and Dudík (2008), with 226 species and a median number of 57 presence sites. Better prediction performance may have been obtained by tuning the regularization values and background size and adding input environmental variables and features. However, the risk would have been to favor overfitting to the detriment of the bi-ecological interpretation of the model (see for example Merow et al., 2013). According to the entomologists who participated in the study, the model appears to be a good compromise between overfitting (that would have predicted suitable areas near occurrence points only) and being too general (that would have predicted suitable areas in too many environmental contexts for which the specialists have no species presence evidence).

**Correction of the sampling bias effect**

In this study, the effect of sampling bias was corrected by selecting background points with the same environmental bias as the sampled points. This approach appeared to be useful when applied to *An. darlingi* in French Guiana. Without a bias effect correction, the model predicted very high
HSI values in highly urbanized areas whereas these areas are known to be unsuitable for this vector (see above). The biased background set is more concentrated around the sampled points (in the environmental space) than the uniform random background, and is not likely to include environmental conditions that are highly dissimilar to those encountered at the sampled points. As a result, environmental conditions highly dissimilar to those of the sampled points can be subjected to extrapolation, which may lead to erroneous habitat suitability predictions and bio-ecological interpretations. This justifies not using the extrapolation option for modeling. The predicted HSI map from the model with a biased background contains several excluded areas, whereas that of the model with a uniform random background does not. Excluded areas correspond to high altitude areas which are unsuitable for *An. darlingi* (Mouchet 2004).

When using a uniform random background, the three most contributive variables (cumulative contribution equal to 85.5%) were all directly linked to human presence and territory accessibility (*ROADS*: 38.1%, *PER_URB_NEIGH*: 34.9%, and *HA_MAX*: 12.5%). Thus, apart from urban areas, high HSI values were associated with high *HA_MAX* and *ROADS* values. However, when correcting the sampling bias effect, the *Landscape (LS)* variable was the second most contributive variable (14.1%), the *ROADS* variable contribution increased to 62.6%, and the *PER_URB_NEIGH* variable contribution decreased to 11.1% (see Table 3).

From a quantitative point of view, the two approaches (with and without applying the correction of the sampling bias effect) resulted in identical AUC and partial AUC ratios. However, the regularized gain and the CBI were lower without correction, with values equal to 2.81 (vs. 3.18) and 0.284 (vs. 0.421), respectively.

Thus, correction of the sampling bias effect gave better results: both more consistent with knowledge from the field and more accurate in terms of prediction. The fact that the contributions of the *LS* and *HA_MAX* variables respectively increased and decreased with the use of the biased background, tends to show that the correction method actually manages to counterbalance the over-
representation of inhabited areas (cities, villages, and gold mining areas) in the sampled data. However, the very high contribution of the variable ROADS may be a residual effect of sampling bias, as sampling is essentially performed in the vicinity of accessible roads and tracks. Further studies are necessary to objectively and quantitatively assess the actual performance of the proposed methodology for correcting the effect of sampling bias.

**Habitat Suitability and Malaria in French Guiana**

Alimi et al. (2015) highlighted the utility of SDMs for gaining a better understanding of the geographical range and distribution of vectors for eliminating malaria and preventing outbreaks. The coastal strip in French Guiana is generally malaria free, although some cases resulting from local transmission are regularly diagnosed (Ardillon et al. 2015). This study, as well as that of Vezenegho et al. (2015), shows that the savanna in French Guiana may be highly suitable for *An. darlingi*. In the forest, Pommier de Santi et al. (2016c) found a link between mining, malaria cases, and the presence of *An. darlingi*. Indeed, >74% of malaria cases in French army soldiers were associated with operations to counteract illegal gold mining (Pommier de Santi et al. 2016a).

According to the results of the present study, some areas associated with intense gold mining activity, known to be malaria transmission foci, are not necessarily associated with very high HSI values. In the village of Camopi, the annual malaria prevalence was 70% for children younger than seven years of age between 2000 and 2002 (Carme et al. 2005), reaching 100% in 2006 (Hustache et al. 2007). However, only some pixels on the border of the Camopi and Oyapock rivers have high values on the HSI map (area G in Figure 4). This is consistent with the study of Girod et al. (2011), which showed that the number of *An. darlingi* caught in this village was very low relative to the incidence of malaria cases. These findings collectively highlight two important points. First, the HSI map shown in Figure 4 does not correspond to a map of malaria transmission risk. Transmission risk depends on many factors that were not taken into account here, such as the parasitic charge and immunological status of the local population, compositional and
configurational features of the landscape (Stefani et al, 2013; Li et al, 2016), and behavioral factors.

Second, this highlights that malaria transmission can occur in areas where there is a very low density of \textit{An. darlingi}. This may be due to the presence of other \textit{Anopheles} species such as \textit{An. (Nys.) nuneztovari} Galbaldón, \textit{An. (Nys.) oswaldoi} Peryassú, \textit{An. (Nys.) intermedius} Peryassú, \textit{An. (Nys.) marajoara} Galvão and Damasceno, or \textit{An. (Nys.) ininii} Sénévet and Abonnenc (Diptera: Culicidae), already known to be naturally infected with \textit{Plasmodium} species and/or described as efficient secondary malaria vectors (Dusfour et al. 2012, Pommier de Santi et al, 2016c).

Environmental Characterization

A significant limitation of this study was the spatial resolution of the environmental data. Capture campaigns are generally carried out at a local scale (villages or camps; Vezenegho et al. 2015, Dusfour et al. 2013). The spatial resolution of the study was not sufficient to take into account the heterogeneity of the environment at the capture scale. The use of environmental data with higher spatial resolution, such as the canopy height estimation from Fayad et al. (2014) or finer characterization of the land cover could improve future studies. However, these data are not consistently available across the entire territory.

In conclusion, the results of this study help to complete our knowledge on the spatial distribution of the principal malaria vector in this Amazonian region, and to identify the main factors that favor its presence. These results can be exploited to define the necessary targeted vector control strategies in a malaria pre-elimination context, and to extrapolate the acquired knowledge to other Amazonian contexts. They also suggest areas that need to be targeted to complete the field knowledge, validate the prediction and strengthen the model. Eventually, these proposed methodological developments can be applied to other species, including other disease vectors.

Acknowledgements

This study was funded by the Fonds Social Européen (FSE), Centre National d'Etudes Spatiales
(CNES), and Collectivité Territoriale de Guyane. Financial support was partially provided by the “Investissement d’Avenir” grants managed by the Agence Nationale de la Recherche (Center for the study of Biodiversity in Amazonia, ANR-10-LABX-0025) and by the GAPAM-Sentinela project of the Franco-brazilian scientific and academic cooperation program Guyamazon (funds: IRD, CIRAD, French Embassy in Brazil, Territorial Collectivity of French Guiana, Brazilian State-level research agencies of Amapá, Amazonas and Maranhão).

References cited


Météo-France. 2016. Données pluviométriques disponibles au 01/01/2016


648
**Peterson, A. T. 2007.** Ecological niche modelling and understanding the geography of disease
650
**Peterson, A. T., M. Papeš, and J. Soberón. 2008.** Rethinking receiver operating characteristic
652
654
656
658
661
664
668
**Pommier de Santi, V., R. Girod, M. Mura, A. Dia, S. Briolant, F. Djossou, I. Dusfour, A. Mendebil, F. Simon, X. Deparis, and F. Pagès. 2016c.** Epidemiological and entomological studies of a malaria outbreak among French armed forces deployed at illegal gold mining sites
671


Figure 1. *Culicidae* capture points and *Anopheles darlingi* presence points (from 2000 to 2013).
Figure 2. Response curves of categorical environmental variables.
Figure 3. Response curves of numerical environmental variables. Dashed lines show the mean values and the grey regions represent the interval between the maximum and minimum values.
Figure 4. Habitat suitability index map. Six main areas with a high habitat suitability index (A to F) and Camopi village (G) are circumscribed by the red circles and rectangles.
Figure 5. Zoom of urban areas. a, d, and g: habitat suitability index maps. b, e, and h: landscape type. c, f, and i: percentage urbanization of neighbor pixels. Rectangles correspond to highly urbanized areas (LS class is Urban and PER_URB_NEIGH ≥ 50%). Cayenne and Kourou include highly urbanized areas, but Saint-Laurent-du-Maroni does not.
### Table 1. Raw environmental data and derived variables used to build the model.

<table>
<thead>
<tr>
<th>Number of input variable</th>
<th>Producer, reference</th>
<th>Raw environmental data</th>
<th>Derived from (information source)</th>
<th>Date(s)</th>
<th>Original spatial resolution or interpretation scale</th>
<th>Derived SDM input variable(s)</th>
<th>Type of feature extraction for each 1x1 km pixel</th>
<th>Classes or range of values and units</th>
<th>Environment types</th>
<th>A priori effect on An. darlingi presence and bibliographic references</th>
<th>Input variable type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Forest National Office (ONF), (Guittet et al. 2013)</td>
<td>Geomorphological landscape (GLS)</td>
<td>SRTM</td>
<td>2000</td>
<td>≥ 5000 m</td>
<td>Geomorphological landscape (GLS)</td>
<td>Majority class</td>
<td>12 classes</td>
<td>Natural environment</td>
<td>( / )</td>
<td>Categorical</td>
</tr>
<tr>
<td>2</td>
<td>Forest National Office (ONF), (Guittet et al. 2013)</td>
<td>Geomorphological landform (GLF)</td>
<td>SRTM</td>
<td>2000</td>
<td>≥ 200 m</td>
<td>Geomorphological landform (GLF)</td>
<td>Majority class</td>
<td>15 classes</td>
<td>Natural environment</td>
<td>( / )</td>
<td>Categorical</td>
</tr>
<tr>
<td>3</td>
<td>Agricultural Research Centre for International Development (CIRAD), (Gond et al. 2011)</td>
<td>Landscape types (LS)</td>
<td>Spot-Vegetation</td>
<td>2000</td>
<td>1000 m</td>
<td>Landscape types (LS)</td>
<td>Correction of pixels corresponding to urban areas and mangroves</td>
<td>14 classes</td>
<td>Natural environment</td>
<td>( / )</td>
<td>Categorical</td>
</tr>
<tr>
<td>4, 5, 6</td>
<td>National Aeronautics and Space Administration (NASA)</td>
<td>Altitude (ALT)</td>
<td>SRTM</td>
<td>2000</td>
<td>30 m</td>
<td>Altitude</td>
<td>Statistical computation</td>
<td>0 – 832 m</td>
<td>Natural environment</td>
<td>( - )</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

*Note: The table lists various environmental data sources and how they were used to build a model for predicting the habitat suitability of *An. darlingi* in French Guiana.*
<table>
<thead>
<tr>
<th>Number of input variable</th>
<th>Producer, reference</th>
<th>Raw environmental data</th>
<th>Derived from (information source)</th>
<th>Date(s)</th>
<th>Original spatial resolution or interpretation scale</th>
<th>Derived SDM input variable(s)</th>
<th>Type of feature extraction for each</th>
<th>Classes or range of values and units</th>
<th>Environment types</th>
<th>A priori effect on An. darlingi presence</th>
<th>Input variable(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>National Institute of Geographic and Forestry Information (IGN)</td>
<td>Road and track network</td>
<td>BD TOPO®</td>
<td>2011</td>
<td>≥ 1000 m</td>
<td>Length of roads and tracks outside of urban areas (ROADS)</td>
<td>Computation of road/track lengths</td>
<td>0 – 12545 m</td>
<td>Non-permanent anthropogenic changes</td>
<td>(+ )</td>
<td>Singer and Castro (2001)</td>
</tr>
<tr>
<td>8</td>
<td>Association Kwata 'Study and Conservation of French Guianan Wildlife' (de Thoisy et al. 2010)</td>
<td>Human footprint (HFP)</td>
<td>From various sources *</td>
<td>2005</td>
<td>≥ 1000 m</td>
<td>Percentage of urbanization of neighboring pixels (PER_URB_NEIGH)</td>
<td>Percentage of urbanization within the eight neighbor cells</td>
<td>0-100%</td>
<td>Urbanization</td>
<td>( )</td>
<td>Stefani et al. (2013)</td>
</tr>
<tr>
<td>9, 10, 11</td>
<td>Association Kwata 'Study and Conservation of French Guianan Wildlife' (de Thoisy et al. 2010)</td>
<td>Human footprint (HFP)</td>
<td>From various sources *</td>
<td>2005</td>
<td>≥ 1000 m</td>
<td>Human activities which non-permanently alter natural environment (HA) - minimum (HA_MIN) - maximum (HA_MAX) - median (HA_MED)</td>
<td>Statistical computation</td>
<td>0-30</td>
<td>Non-permanent anthropogenic changes</td>
<td>( )</td>
<td>Vittor et al. (2009)</td>
</tr>
</tbody>
</table>

* French Institute for Statistical and Economic studies (INSEE); Regional Departments for Food, Agriculture and the Forest (DAAF); ONF; Regional Equipment, Habitat and Planning Authority (DDE) and Hammond et al. (2007).

See the section on environmental variables in Materials and Methods.

A priori effect on An. darlingi presence: (+) favorable; (-) unfavorable; (/) depends on categorical variable values.
Table 2. Mean contributions and jackknife results of the eleven input environmental variables.

<table>
<thead>
<tr>
<th>Environmental variables</th>
<th>Contribution (%)</th>
<th>Cumulative contribution (%)</th>
<th>Gain with the variable only</th>
<th>Decrease of the gain without the variable (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ROADS</strong></td>
<td>51.45</td>
<td>51.45</td>
<td>2.20</td>
<td>-7.98</td>
</tr>
<tr>
<td><strong>PER_URB_NEIGH</strong></td>
<td>17.17</td>
<td>68.62</td>
<td>1.86</td>
<td>-0.41</td>
</tr>
<tr>
<td><strong>LS</strong></td>
<td>15.32</td>
<td>83.94</td>
<td>2.23</td>
<td>-4.67</td>
</tr>
<tr>
<td><strong>HA</strong></td>
<td>7.43</td>
<td>91.37</td>
<td>min: 0.02</td>
<td>min: -0.06</td>
</tr>
<tr>
<td></td>
<td>(min: 0.35;</td>
<td></td>
<td>median: 0.15</td>
<td>median: -0.22</td>
</tr>
<tr>
<td></td>
<td>median: 0.24;</td>
<td></td>
<td>max: 0.43</td>
<td>max: -2.10</td>
</tr>
<tr>
<td></td>
<td>max: 6.84)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GLS</strong></td>
<td>5.35</td>
<td>96.72</td>
<td>1.40</td>
<td>-2.59</td>
</tr>
<tr>
<td><strong>ALT</strong></td>
<td>2.09</td>
<td>98.81</td>
<td>min: 1.12</td>
<td>min: -1.04</td>
</tr>
<tr>
<td></td>
<td>(min: 1.34;</td>
<td></td>
<td>median: 1.04</td>
<td>median: -0.39</td>
</tr>
<tr>
<td></td>
<td>median: 0.69;</td>
<td></td>
<td>max: 0.76</td>
<td>max: -0.03</td>
</tr>
<tr>
<td></td>
<td>max: 0.06)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GLF</strong></td>
<td>1.19</td>
<td>100</td>
<td>0.80</td>
<td>-0.21</td>
</tr>
</tbody>
</table>
Table 3. Mean contributions and jackknife results of the seven input environmental variables of the simpler model.

<table>
<thead>
<tr>
<th>Environmental variables</th>
<th>Contribution (%)</th>
<th>Cumulative contribution (%)</th>
<th>Gain with the variable only</th>
<th>Decrease of the gain without the variable (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROADS</td>
<td>62.61</td>
<td>62.61</td>
<td>2.31</td>
<td>-8.61</td>
</tr>
<tr>
<td>LS</td>
<td>14.10</td>
<td>76.71</td>
<td>2.35</td>
<td>-6.23</td>
</tr>
<tr>
<td>PER_URB_NEIGH</td>
<td>11.15</td>
<td>87.86</td>
<td>2.05</td>
<td>-0.58</td>
</tr>
<tr>
<td>HA_MAX</td>
<td>5.39</td>
<td>93.25</td>
<td>0.37</td>
<td>-1.74</td>
</tr>
<tr>
<td>GLS</td>
<td>3.84</td>
<td>97.09</td>
<td>1.44</td>
<td>-1.90</td>
</tr>
<tr>
<td>GLF</td>
<td>2.1</td>
<td>99.19</td>
<td>1.01</td>
<td>-0.32</td>
</tr>
<tr>
<td>ALT_MIN</td>
<td>0.88</td>
<td>100</td>
<td>1.27</td>
<td>-1.29</td>
</tr>
</tbody>
</table>
Table 4. Characterization of areas with a high HSI

*ns.* signifies that the high HSI of the concerned area was not driven by that environmental variable, (+) signifies that when the value of the variable increases, the HSI also increases also, (-) signifies that when the value of the variable decreases, the HSI increases, and cells with classes name signifies that the presence of the given class implies a high HSI.

<table>
<thead>
<tr>
<th>Area</th>
<th>ROADS</th>
<th>LS classes</th>
<th>PER_URB_NEIGH</th>
<th>HA_MAX</th>
<th>GLS classes</th>
<th>GLF classes</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(+)</td>
<td>- Woodland savanna / Dry forest</td>
<td>(-)</td>
<td>(+)</td>
<td>- Coastal plain with low relief</td>
<td>- Small size and flat wetland</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Mixed high and open forest</td>
<td></td>
<td></td>
<td>- Plain with residual reliefs (back coastal)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>(+)</td>
<td>- Open forest</td>
<td><em>ns.</em></td>
<td>(+)</td>
<td>- Peneplain with moderate hills</td>
<td>- Wet hillok (low base-level)</td>
<td>(-)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Mixed high and open forest</td>
<td></td>
<td></td>
<td></td>
<td>- Large flattened relief</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td><em>ns.</em></td>
<td>- Open forest</td>
<td><em>ns.</em></td>
<td><em>ns.</em></td>
<td>- Coastal flat plain</td>
<td>- Large flattened and wet relief</td>
<td>(-)</td>
</tr>
<tr>
<td>D</td>
<td>(+)</td>
<td>- Mixed high and open forest</td>
<td><em>ns.</em></td>
<td>(+)</td>
<td><em>ns.</em></td>
<td><em>ns.</em></td>
<td>(-)</td>
</tr>
<tr>
<td>E</td>
<td>(+)</td>
<td>- Mixed high and open forest</td>
<td><em>ns.</em></td>
<td>(+)</td>
<td>- Peneplain with moderate hills</td>
<td>- Large flattened relief</td>
<td>(-)</td>
</tr>
<tr>
<td>F</td>
<td><em>ns.</em></td>
<td>- Open forest</td>
<td><em>ns.</em></td>
<td><em>ns.</em></td>
<td>- Peneplain with moderate hills</td>
<td>- Lowered half-orange</td>
<td>(-)</td>
</tr>
<tr>
<td>G</td>
<td>(+)</td>
<td>Mixed high and open forest</td>
<td><em>ns.</em></td>
<td>(+)</td>
<td><em>ns.</em></td>
<td><em>ns.</em></td>
<td>(-)</td>
</tr>
</tbody>
</table>
## S1. Coordinates of *Anopheles darlingi* presence sites (Coordinate system: RGFG95/UTM22N)

<table>
<thead>
<tr>
<th>Number of sites</th>
<th>Locality</th>
<th>Longitude</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cayodé</td>
<td>175866.785122246</td>
<td>374453.165383</td>
</tr>
<tr>
<td>2</td>
<td>Taluène</td>
<td>162884.684665956</td>
<td>372648.888139721</td>
</tr>
<tr>
<td>3</td>
<td>Cacao</td>
<td>336895.086300153</td>
<td>505521.800163598</td>
</tr>
<tr>
<td>4</td>
<td>Midenangalanti</td>
<td>121041.015986301</td>
<td>557342.72581239</td>
</tr>
<tr>
<td>5</td>
<td>Grand Santi</td>
<td>124374.105534813</td>
<td>473261.338110987</td>
</tr>
<tr>
<td>6</td>
<td>Bois Martin</td>
<td>118927.138507594</td>
<td>552587.496776593</td>
</tr>
<tr>
<td>7</td>
<td>Flavien Campou</td>
<td>126212.611562883</td>
<td>477960.065899793</td>
</tr>
<tr>
<td>8</td>
<td>Régina</td>
<td>374698.048396995</td>
<td>476928.730177523</td>
</tr>
<tr>
<td>9</td>
<td>Camopi</td>
<td>352395.046055802</td>
<td>350764.216480451</td>
</tr>
<tr>
<td>10</td>
<td>Alikéné</td>
<td>322929.978284572</td>
<td>360135.527835386</td>
</tr>
<tr>
<td>11</td>
<td>Mine Boulanger</td>
<td>343189.582357019</td>
<td>504590.723307373</td>
</tr>
<tr>
<td>12</td>
<td>Carbet Légion crique Sikini</td>
<td>359046.553479</td>
<td>361996.097871</td>
</tr>
<tr>
<td>13</td>
<td>Cogneau - 23. Lot. Aquavilla</td>
<td>354091.053713437</td>
<td>538837.801956188</td>
</tr>
<tr>
<td>14</td>
<td>41. rue des Ixoras - Lot. Cogneau</td>
<td>351322.960529913</td>
<td>541116.10235073</td>
</tr>
<tr>
<td></td>
<td>Larivot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Attila Cabassou</td>
<td>354987.967015231</td>
<td>540678.87835953</td>
</tr>
<tr>
<td>16</td>
<td>La Chaumièr</td>
<td>347961.540051952</td>
<td>539402.864452834</td>
</tr>
<tr>
<td>17</td>
<td>1228. Ch. de La Chaumièr</td>
<td>349626.610158034</td>
<td>540136.658859129</td>
</tr>
<tr>
<td>18</td>
<td>Saint-Georges</td>
<td>411052.118458134</td>
<td>429833.606330522</td>
</tr>
<tr>
<td>19</td>
<td>Quartier Espérance - Saint-Georges</td>
<td>411113.960068634</td>
<td>430709.665437749</td>
</tr>
<tr>
<td>20</td>
<td>Village Martin - Saint-Georges</td>
<td>411519.804866871</td>
<td>432628.481492214</td>
</tr>
<tr>
<td>21</td>
<td>Boulangerie – Saül</td>
<td>254644.262333224</td>
<td>400558.641871421</td>
</tr>
<tr>
<td>22</td>
<td>Chemin Mogès</td>
<td>352032.329517247</td>
<td>529456.717653458</td>
</tr>
<tr>
<td>23</td>
<td>Dorlin</td>
<td>216742.863044</td>
<td>415732.862195</td>
</tr>
<tr>
<td>24</td>
<td>Maripasoula</td>
<td>163199.637191727</td>
<td>403271.422431979</td>
</tr>
<tr>
<td>25</td>
<td>Repentir</td>
<td>234327.380239</td>
<td>427769.677349</td>
</tr>
<tr>
<td>Number of sites</td>
<td>Locality</td>
<td>Longitude</td>
<td>Latitude</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>26</td>
<td>Stoupan</td>
<td>352127.057631364</td>
<td>527017.037636427</td>
</tr>
<tr>
<td>27</td>
<td>Camp Pararé – Nouragues</td>
<td>311267.546406115</td>
<td>446010.858172732</td>
</tr>
<tr>
<td>28</td>
<td>Village Blondin - Saint-Georges</td>
<td>409492.529993713</td>
<td>428392.815435875</td>
</tr>
<tr>
<td>29</td>
<td>Quartier Adimo - Saint-Georges</td>
<td>410525.149504269</td>
<td>430989.662548913</td>
</tr>
<tr>
<td>30</td>
<td>Camp Bernet/ Légion étrangère</td>
<td>410437.343147834</td>
<td>429860.701524782</td>
</tr>
<tr>
<td>31</td>
<td>La ferme de Lait-Quateur</td>
<td>332757.212761399</td>
<td>552628.404024885</td>
</tr>
<tr>
<td>32</td>
<td>Grand Usine</td>
<td>288898.531939378</td>
<td>372330.073251345</td>
</tr>
<tr>
<td>33</td>
<td>Dagobert</td>
<td>226797.35968</td>
<td>438864.200286</td>
</tr>
<tr>
<td>34</td>
<td>Cacao</td>
<td>336515.199831583</td>
<td>505801.969900718</td>
</tr>
<tr>
<td>35</td>
<td>Saut-Maripa – Camp militaire</td>
<td>401579.145107468</td>
<td>420354.782186144</td>
</tr>
<tr>
<td>36</td>
<td>Eau-Claire</td>
<td>213145.824694</td>
<td>398626.838699</td>
</tr>
<tr>
<td>37</td>
<td>Cacao</td>
<td>336956.199791048</td>
<td>504651.969979654</td>
</tr>
<tr>
<td>38</td>
<td>Cacao</td>
<td>337909.199733996</td>
<td>506008.969874801</td>
</tr>
<tr>
<td>39</td>
<td>Cacao</td>
<td>336576.199808207</td>
<td>503413.970071365</td>
</tr>
<tr>
<td>40</td>
<td>Cacao</td>
<td>336555.199807874</td>
<td>503184.970087929</td>
</tr>
<tr>
<td>41</td>
<td>Impasse de la rafinerie – Cogneau</td>
<td>354091.053713437</td>
<td>538837.801956188</td>
</tr>
<tr>
<td>42</td>
<td>Quartier Bambou</td>
<td>411277.041477492</td>
<td>430178.275876772</td>
</tr>
<tr>
<td>43</td>
<td>Quartier Maripa - Saint-Georges</td>
<td>410252.974346512</td>
<td>430188.463058511</td>
</tr>
<tr>
<td>44</td>
<td>Quartier Savane - Saint-Georges</td>
<td>411024.737954334</td>
<td>430940.054473617</td>
</tr>
<tr>
<td>45</td>
<td>Cacao</td>
<td>336826.199809151</td>
<td>505764.969900894</td>
</tr>
<tr>
<td>46</td>
<td>Cacao</td>
<td>337091.199787723</td>
<td>505441.969921949</td>
</tr>
<tr>
<td>47</td>
<td>Cacao</td>
<td>336186.199848992</td>
<td>505049.969957266</td>
</tr>
<tr>
<td>48</td>
<td>Quartier Onozo- Saint-Georges</td>
<td>411339.040785181</td>
<td>430641.896115415</td>
</tr>
</tbody>
</table>
S2. Creation of the relative sampling effort map

Capture data of Culicidae (74 capture sites) were used to estimate the sampling effort of An. darlingi. The collection methods were identical and the sampling bias for the family was assumed to be representative of that for the focal species.

The sampling bias was defined as the relative sampling effort in the environmental space. For a pixel \( i \), it corresponds to the ratio of the number of sampled pixels over the total number of pixels, within the environmental neighborhood of \( i \).

First, all pixels of the study area were represented in the environmental variable space. This was accomplished by performing a Factorial Analysis of Mixed Data (FAMD) (Pagès, 2004). This analysis jointly takes into account numerical and categorical variables and makes it possible to represent the pixels within an Euclidean, orthonormal space defined from the whole set of environmental variables.

The membership degree of a pixel \( j \) to the neighborhood of pixel \( i \), denoted \( w_{ij} \), was defined by a Gaussian-like membership function:

\[
 w_{ij} = 0.5 \left( \frac{d_{ij}}{D_{\text{min}}} \right)^2
\]

with \( d_{ij} \) the euclidean distance between \( i \) and \( j \) in the factorial space, and \( D_{\text{min}} \) the threshold distance over which \( j \) does not significantly belong to the environmental neighborhood of \( i \), i.e. over which \( w_{ij} < 0.5 \). The membership degree \( w_{ij} \) has the following properties:

- \( w_{ij} \in [0,1] \);
- \( w_{ij} = 1 \) if \( d_{ij} = 0 \);
- \( w_{ij} < 0.5 \) if \( d_{ij} > D_{\text{min}} \).
The parameter $D_{\text{min}}$ was set from a priori knowledge of *An. darlingi* bio-ecology. As highly urbanized areas are not suitable for *An. darlingi* (see § I.3), we stated that a pixel associated with *An. darlingi* presence cannot belong to a highly urbanized pixel. Reciprocally, a pixel considered to be highly urbanized cannot belong to the environmental neighborhood of a pixel where *An. darlingi* was observed.

Consequently, given $P$, the set of pixels where the species was observed and $U$, the set of pixels belonging to highly urbanized areas, $D_{\text{min}}$ was defined as follows:

$$D_{\text{min}} = \min \left( d_{pu} \right)_{P \cap P \cup U}$$

(2)

A pixel is considered to be highly urbanized if it belongs to the LC class *Urban* and if its eight neighboring pixels present an average urbanization percentage ($PER_{\text{URB}_NEIGH}$) higher than or equal to 50%.

The concepts of environmental space and neighborhood, as well as the key method parameters are schematically represented in Figure S1.

Given $X$, the set of pixels of the study area, and $c = \{c_i\}_{i \in X}$, a vector such that $c_i = 1$ if $i$ is sampled and $c_i = 0$ otherwise, the relative sampling effort at pixel $i$, $z_i$, is then defined as:

$$z_i = \sum_{j \in X} w_j \cdot c / \sum_{j \in X} w_j$$

(3)

The relative sampling effort was computed for each pixel of the study area. The resulting map was used to bias the random selection of background points. Consequently, for a given pixel, the greater the relative sampling effort, the higher the chance of selecting the pixel as a background point.

Reference

Figure S1. Neighborhood of a pixel $i$ in the environmental space represented by the first and second factorial axes. The environmental neighborhood of point $i$ is represented by the Gaussian function. The blue lines define the limit of the neighborhood of $i$. Only point $j$ is situated above these lines. Thus $j$ is in the neighborhood of $i$ in the first factorial plane.