RESEARCH ARTICLE

# Downscaling local distribution of cattle over Guadeloupe archipelago: An adapted method for disaggregating census data

**Victor Dufleit**[1*], **Laure Guerrini**[1,2], **Marius Gilbert**[3], **Daniele Da Re**[4,5], **Eric Etter**[1]

**1** ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France, **2** CIRAD, UMR ASTRE, Montpellier, France, **3** Université Libre de Bruxelles, **4** Center Agriculture Food Environment, University of Trento, San Michele All' Adige, Italy, **5** Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy

* vic.du-loc@live.fr

## Abstract

Gridded livestock distribution datasets have been produced for several years and are used in various fields, including epidemiology, livestock impact assessment, and territory management. Those datasets are based on census conducted at national/sub-national scale which are then downscaled using machine learning algorithm and relevant spatial-explicit environmental predictors. The most known dataset of livestock disaggregated observations is the Gridded Livestock of the World (GLW), which produces global maps of livestock density at 10 km spatial resolution for several livestock species. Though this spatial resolution can be appropriate to describe livestock distribution at the global scale, it inherently leads to a coarse representation of breeding species density for smaller territories such as the Caribbean Islands. In this study, we propose an adaptation of the GLW methodology that accounts for the spatial autocorrelation in observed cattle distribution, thereby better capturing the specific characteristics of geographically limited areas such as the Guadeloupean archipelago. Cattle census data were collected for the 32 municipalities of the archipelago and associated to environmental predictors derived from remote sensing and land cover datasets. Together with the Random Forest (RF) algorithm used in the standard GLW methodology, we tested the performance of a Geographical Random Forest (GRF), a novel methodology allowing for taking into account the spatial autocorrelation of the response variable. The GRF algorithm demonstrated significantly better performance compared to the RF algorithm, albeit with longer processing times, and allowed us producing cattle distribution maps for the entire Guadeloupe archipelago at a spatial resolution of 225 m using both algorithms. The approach developed holds potential for application to other small territories, including other islands in the Caribbean.

OPEN ACCESS

**Citation:** Dufleit V, Guerrini L, Gilbert M, Da Re D, Etter E (2026) Downscaling local distribution of cattle over Guadeloupe archipelago: An adapted method for disaggregating census data. PLoS One 21(1): e0324695. https://doi.org/10.1371/journal.pone.0324695

**Editor:** Edvard Mizsei, HUN-REN Centre for Ecological Research, HUNGARY

**Received:** April 30, 2025

**Accepted:** January 2, 2026

**Published:** January 21, 2026



**Data availability statement:** Municipal cattle count data used in this study cannot be shared publicly due to confidentiality restrictions imposed by the French Ministry of Agriculture. Data are available for researchers who meet the criteria for access to confidential data by

contacting the French Ministry of Agriculture at marie.bascou@agriculture.gouv.fr. Older censuses are publicly available for Guadeloupe via the French agricultural statistic service "agrestre" (https://agreste.agriculture.gouv.fr/agreste-web/disaron/G_2141/detail/). The R code used in this study and a dummy dataset are available at: https://gitlab.com/vic.duf29/griddedlivestockofguadeloupe.

**Funding:** This study was sponsored by the RACE project (USDA-ARS Project Number: 3022-32000-018-006-S).

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Knowing the spatial distribution of livestock abundance is a key to guide and implement livestock production and sanitary policies. National or local administrations conduct livestock censuses to obtain detailed information on animal densities and distribution, which are then used to assess the development and the economic benefits of the livestock sector, to measure the impact on natural resources or on public health and to control livestock diseases [1].

The Gridded Livestock of the World (GLW) database, first released in 2007, provides a gridded and downscaled representation of livestock densities worldwide [2–4]. Livestock density estimates in GLW are derived from census data collected at different scales (national or sub-national – region, district, communal scale/village) and downscaled using environmental covariates and statistical methods. The first version of the GLW [2] applied a stratified regression approach to estimate livestock densities at a spatial resolution of approximately 5 x 5 km at the equator (3 minutes of arc). In contrast, newer versions of GLW [4, 5] utilize Random Forest (RF) algorithms, which have demonstrated superior prediction accuracy in various contexts [6, 7], to downscale livestock census data globally at a spatial resolution of approximately 10 x 10 km at the equator.

While this spatial resolution choice reduces spatial artifacts in predictions at a global scale, it results in a coarse representation of livestock density in smaller geographical areas like the Caribbean islands, where detailed livestock density information is critical for risk assessment and sanitary policies [8]. For example, in Guadeloupe, an overseas department and region of France in the Caribbean, the GLW dataset represents the entire archipelago with only 15 pixels.

Another limitation of the current GLW is its inability to explicitly account for spatial autocorrelation of the response variable. The RF algorithm does not inherently account for spatial autocorrelation. Though spatial proxies such as coordinates of training observations or distance between observations could be added as predictors to account for potential spatial autocorrelation in the training dataset, this approach has not always led to gain in prediction accuracy or pertinent spatial pattern reproduction [9]. Various model formulation have been proposed to incorporate spatial information into RF [10, 11], among which the Geographical Random Forest approach (GRF) developed by Georganos et al. (2021) [12] stems out. GRF fits local Random Forests for each prediction location, using nearby observations (via a kernel function), which allows relationships between predictors and response to vary across space.

Despite these advances and the different applications of livestock density census downscaling modelling methodologies [13–15], the use of such models in data-scarce and spatially heterogeneous contexts—such as small island territories—remains limited. This is particularly relevant for regions like the Guadeloupe archipelago, where traditional livestock data are often fragmented or unavailable at high spatial resolution. In such settings, evaluating the performance of models like RF and GRF becomes crucial for developing reliable, context-sensitive predictive tools.

In this study, we assessed the applicability of predictive models trained on small datasets and propose a methodological strategy to mitigate issues related to spatial

autocorrelation. Specifically, our goal was to develop a Territorial Livestock Mapping (TLM) model for the Guadeloupe archipelago using both RF and GRF algorithms, to predict cattle densities at a fine spatial resolution of 225 meters. One objective was to compare prediction accuracy of both algorithms. Moreover, the technique developed by the FAO, which can be referred as a "census disaggregation" methodology, was never applied to such a small territory. These high-resolution maps aim to support epidemiological research by offering a detailed spatial representation of livestock distribution across the territory. In addition, we evaluated the impact of methodological choices on the predictive performance of the two algorithms. Specifically, we tested the impact of the choice of i) different sampling strategies and ii) various input variables, including topographic descriptors, land cover data, and environmental parameters, selected based on their ecological relevance to cattle distribution and habitat suitability.

## Materials and methods

### Study area

The archipelago of Guadeloupe is located in the Caribbean and is part of the French West Indies. It comprises "mainland" Guadeloupe, with two islands, Basse-Terre (848 km², to the west) and Grande Terre (588 km², to the east), plus the administrative dependencies of the islands of La Désirade (21 km²), Marie-Galante (158 km²) and Les Saintes (14 km²). The tropical climate is influenced by the sea and trade winds with microclimatic variability driven by topography. Basse-Terre is dominated by the volcanic massif of La Soufrière (1467 m), resulting in a rainy climate, with annual rainfall ranging from 1000 to over 7000 mm, peacking at the volcano's summit. In contrast, Grande Terre's lack of significant relief, results in a drier climate with annual rainfall between 1000 and 1500 mm [16]. The administrative dependencies, share a similar climate to Grande Terre. Guadeloupe has a population of around 400,000 with tourism as its primary economic activity. Agriculture remains important, with two main crops, sugar cane and bananas [17]. Livestock farming has declined since the 1980s [18] due to factors like urban development, tropical parasites such as ticks (e.g., *Amblyomma variegatum, Rhipicephalus microplus* [18]) and related disease (e.g., *anaplasmosis, heartwater*), feral dog attacks [19] and livestock theft [20]. While tethered dominates, grazing is also practiced. Traditionally tied to sugar cane cultivation, cattle are now used for subsistence and leisure activities, with the tradition of "ox pulling". These challenges, combined with land access issues, have contributed to the decline of cattle farming in Guadeloupe.

### Census data

The cattle census data, extracted from the database collected by the departmental livestock institute (2021 census; Agriculture, Alimentation and Forest Direction (DAAF)), and obtained from DAAF Guadeloupe in 2023, provided a list of 5 320 cattle breeders, representing a total of 37 139 cattle heads. Data without information on the location of the breeders (i.e., municipalities) were not included in the survey. Data included the identification of the breeder (EDE number), its location (municipality name) and the number of cattle per breeder. Cattle density per municipality was calculated (Fig 1). Administrative boundaries were obtained from BD TOPO [21]. Moran Index (Moran I) was computed to account for potential spatial autocorrelation in the dataset [22].

### Suitability raster and masks

Masks are commonly employed in studies or modelling processes to restrict the area of interest and the area of training and extrapolation of the model. In this study, a mask was created to delineate areas within the Guadeloupean territory where cattle breeding was either feasible or infeasible; this mask will be referred as the "suitability raster". The construction of this mask was guided by the methodology outlined below:

- Land cover classification: The original land cover classes from the Karucover dataset were reclassified into two categories: "Suitable" or "Unsuitable" for cattle breeding (S1 in S1 File). Land cover classes were analysed to determine the

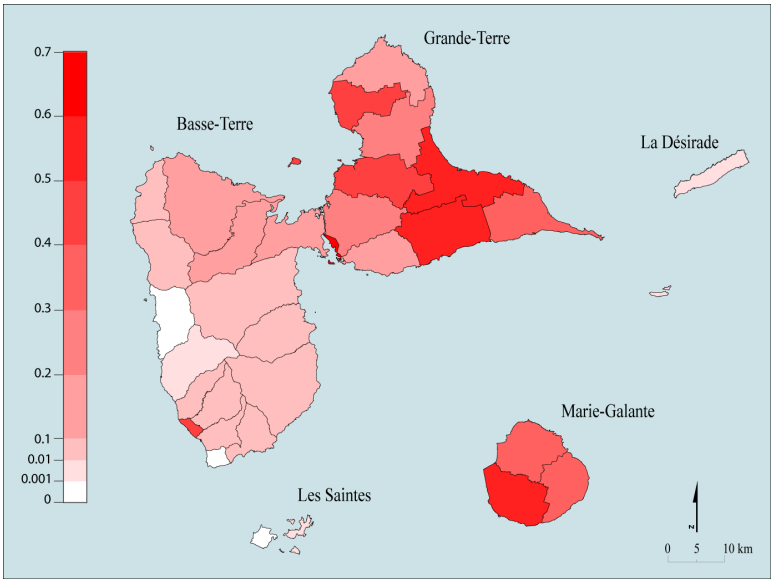


**Fig 1. Cattle densities in Guadeloupe municipalities were obtained from farmers' declarations at the communal level.** The left-hand scale shows the cattle density (in head/ha). Created by Victor Dufleit using BD_TOPO® shapefile data. BD_TOPO® are open access data published by French National Geographic Service under "Etalab 2.0" licence.

https://doi.org/10.1371/journal.pone.0324695.g001

suitability of natural vegetated areas, such as grasslands and forests. Land use data were incorporated to improved classification accuracy (S1a in S1 File). Agricultural land and low-density home gardens were classified as suitable, as cattle are occasionally raised on fallow agricultural land or in-home gardens. Specific vegetated public infrastructures (e.g., roadside vegetation or roundabouts) were also considered suitable for grazing, particularly for farmers with limited herds or land availability (S1b in S1 File).

- Exclusion of unsuitable areas: Areas where cattle breeding is unfeasible, such as the National Park and urban zones (S2 in S1 File), were excluded from the analysis (S1c in S1 File; S2 in S1 File).
- Suitability raster creation: A raster with a resolution of 225 x 225 m was created to cover the entire study area. Pixels were assigned a value of 1 (suitable) if they contained at least one suitable polygon (S1d in S1 File).

### Predictors

Spatially explicit predictors are crucial for generating cattle density prediction maps through downscaling approaches [6] (Table 1).

Topographic predictors were derived from a Digital Elevation Model (DEM) for Guadeloupe, provided by the geological and mining research bureau (Bureau de Recherche Géologique et Minière; BRGM). This included elevation, slope, and the Topographical Ruggedness Index (TRI), potentially influencing cattle distribution, by determining accessibility, ease of movement, and suitability for grazing. Steeper slopes or rugged terrain may restrict cattle grazing activities.

Land use and land cover data (LULC) [23] from the high-resolution Karucover 2017 dataset [24], which includes 24 land cover classes and 53 land use classes, provides essential information about vegetation and land use types critical to cattle. From this dataset, ten land cover classes relevant to cattle breeding were identified in consultation with livestock specialists and used to generate 225 m resolution raster layers (see S3, S4 and S5 Tables in S1 File for class conversions).

**Table 1. Predictor variables used in the model. The column "Predictor dataset" shows the variables included in the models model I and model II. See also: Experimental design.**

| Variable name | Type | Use | Source | Predictor dataset |
|---|---|---|---|---|
| Digital Elevation Model (DEM) | Topographic | Spatial predictor | BRGM [29] | I; II |
| Slope | Topographic | Spatial predictor | Created from DEM | I; II |
| Topographical Ruggedness Index (TRI) | Topographic | Spatial predictor | Created from DEM | I; II |
| Ten land cover classes derived from Karucover 2017 | Land cover | Spatial predictor | Karugéo [24] | I |
| Barn distance raster | Barn distance raster | Spatial predictor | Karugéo [24] | I |
| Ponds Distance raster | Ponds Distance raster | Spatial predictor | Karugéo [24] | I |
| Road Distance raster | Road Distance raster | Spatial predictor | BD TOPO [21] | I; II |
| River distance raster | River distance raster | Spatial predictor | BD TOPO [21] | I; II |
| 10 Fourier-derived variables from Normalized Difference Vegetation Index (NDVI) | Vegetation | Spatial predictor | MODIS [26] | I; II |
| 10 Fourier-derived variables from Enhanced Vegetation Index (EVI) | Vegetation | Spatial predictor | MODIS [26] | I; II |
| 10 Fourier-derived variables from Day Land Surface Temperature (DLST) | Thermal | Spatial predictor | MODIS [25] | I; II |
| 10 Fourier-derived variables from Night Land Surface Temperature (NLST) | Thermal | Spatial predictor | MODIS [25] | I; II |

A total of 57 predictor rasters were created for the entire archipelago of Guadeloupe.

A correlation matrix of the predictors and correlation between predictors and cattle densities at municipality level were calculated, based on a non-parametric measure of rank correlation (Spearman's ρ).

https://doi.org/10.1371/journal.pone.0324695.t001

Distance rasters for barns, ponds (extracted from the Karucover 2017 dataset) as well as roads and rivers (extracted from the BD TOPO® dataset [21]), influencing cattle accessibility and breeding feasibility, were included.

Environmental variables were extracted from Moderate-resolution Imaging Spectroradiometer (MODIS) data [25, 26], covering five years (2015–2019). Day/Night Land Surface Temperature (DLST, NLST) and vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), capture seasonal dynamics in vegetation growth and thermal conditions. Temporal Fourier Analysis (TFA) of these indices allows the integration of ecological dynamics such as the annual growth cycle of crops (e.g., sugarcane fields) that impact cattle feeding areas. Therefore, the MODIS time series were resampled to a 225 m resolution when necessary and summarised using TFA [27]. This process enabled the calculation of metrics such as the mean, annual minimum and maximum, standard deviation as well as amplitude and phase of the annual, biannual and triannual cycles for the four selected indicators, which were then included in the modelling procedure [28].

### Disaggregating census Data: Experimental design

The methodology used was adapted from the GLW3 model [5] to a smaller scale to develop a Territorial Livestock Mapping model for the Guadeloupe islands. A summary of the method is presented in Fig 2.

Preliminary exploration of census data revealed that the highest cattle densities were recorded in urban areas of Basse-Terre and Pointe-à-Pitre, reflecting the practice of breeders residing in urban centres while maintaining livestock in the surrounding rural areas. To mitigate potential artefacts, cattle were redistributed from urban municipalities to neighbouring rural areas using the following equation (Eq 1):

$$C_i^* = C_i + \left( C_c \times \frac{L_{c-i}}{\sum_0^N L_{c-i}} \right) \quad \text{Eq.1}$$

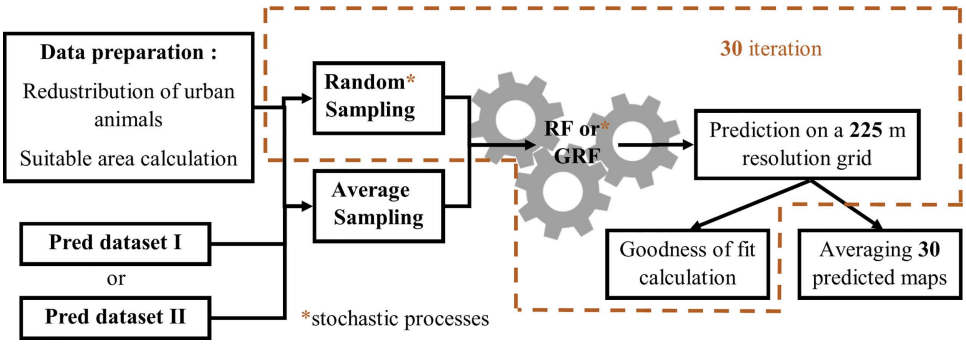


**Fig 2. Flowchart of the study; The data preparation steps are summarised in *S6* in S1 File. Predictors from the different datasets are then sampled using the techniques descibed in "*Sampling predictors*" before being used to create RF or GRF models for cattle density mapping.**

https://doi.org/10.1371/journal.pone.0324695.g002

Where $C_i$ is the number of cattle in the $i^{th}$ neighbouring municipality, $C_c$ is the number of cattle in the $c^{th}$ urban municipality, N is the total number of neighbouring municipalities of an urban municipality, and $L_{c-i}$ is the length of the border between the $c^{th}$ urban municipality and the $i^{th}$ neighbouring municipality. Using the redistributed data and suitable areas (calculated as the total area of all suitable pixels per municipality based on the suitability mask), cattle densities were estimated for suitable areas. These calculations formed part of the data preparation process illustrated in Fig 2. Moran I was adjusted with the obtained 'adjusted' municipal cattle densities.

### Sampling predictors

We employed two predictor-sampling approaches (Fig 3) in order to assess the effect of this methodological choice on the models' prediction accuracy [5].

The first approach, the average sampling method, involves calculating the average of predictor values over all the pixels within the polygon. This average sampling approach was previously used in studies by Li et al. (2021) and

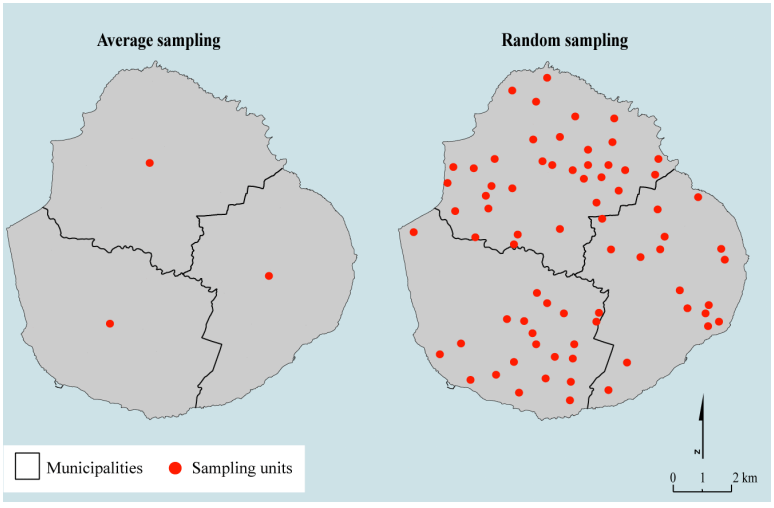


**Fig 3. Presentation of the two predictors sampling methodologies; Created by Victor Dufleit using BD_TOPO® shapefile data. BD_TOPO® are open access data published by French National Geographic Service under "Etalab 2.0" licence.**

https://doi.org/10.1371/journal.pone.0324695.g003

Da Re et al. (2020) [5, 14]. The second approach, the random sampling method instead involves generating a randomly distributed number of points in the area of interest, in our case with a density of 50 points per 10 000 ha. Each point was associated with the predictor values derived from the pixel where it was located. The latter approach increased the size and variability of the training dataset. Random sampling was repeated 30 times to assess variability, while average sampling was conducted once, given the lack of stochasticity inherent to the latter process. For each dataset obtained through random sampling, one RF/GRF was trained, resulting in a total of 30 models. Using the average sampling dataset, 30 random forests were generated.

### Modelling

The training sample, independently from the sampling strategy and predictors choice, was provided to one of the selected RF algorithms. The *ranger* and *spatialML R* packages [30, 31] were selected for their respective implementations of the standard RF model and Geographical Random Forest [32]. Both models have hyperparameters derived from standard RF. They were set following Nicolas et al., 2016 [6] indications:

- **Number of trees (ntree)**: determines the number of trees to be grown and was set as the number of training points divided by 20, with a minimum of 100 trees.
- **The node size**: influences the minimum number of samples required at each node after a split during tree construction. If after a split, a node has a size (number of samples) less than the node size, no further splits will be performed on that subsample. The node size was set to the number of samples divided by 1000 with a minimum node size of 3.
- **mtry**: defines the number of predictors used to build each tree. It was set as the number of predictors divided by 3 with a minimum of 3 predictors.

The GRF model incorporated a "bandwidth" parameter, which defines the maximum distance between a data point and its neighbouring observations. Local models were created for each training sample location using an "adaptive" kernel [32]. For each local model, points were collected around the locations using the K nearest neighbour rules. The "grf.bw" function from the *spatialML* R package was used to optimise the bandwidth values with a minimum value of 1/3 and a maximum value of 2/3 of the training data points (only five steps were tested to optimize the bandwidth). For each bandwidth tested, the coefficient of determination ($r^2$) was calculated and the bandwidth with the highest $r^2$ value was selected for the final model run. For the GRF model predictions, the weight given to the global and local models for the predicted value had to be chosen. As recommended by Georganos et al. (2011) [12], the weights were set to 0.75 and 0.25 for the global and the local models respectively.

For both algorithms, we decided to test two model formulations using two predictor dataset:

- model I including all the 57 predictors
- model II with a reduced set of 45 predictors, excluding Karucover derived predictors

The second model formulation was used to assess the accuracy of the predictions without high-resolution land cover data. This ensures the applicability of the model to other Caribbean islands with limited land cover information.

### Validation

At the end of all bootstraps, a 225 m-resolution predicted map was generated. To evaluate the relative importance of predictors, Permutation variable importance (PVI) was calculated during model training [33]. While PVI highlights the relative contribution of each predictor to model accuracy, it does not indicate whether the effects on the dependent variable are positive or negative. To complement this analysis, PVI was compared with the Spearman correlation (absolute value) between observed municipal cattle density and predictors (averaged at the municipal level), calculated prior to modelling.

Additionally, correlations between predictors were calculated using the full raster dataset. Predicted density maps were converted to animal count maps by multiplying the predicted cattle density by the pixel area. These maps were then aggregated at the municipality level to derive the total predicted animal counts per municipality. To evaluate model performance, the Root Mean Squared Error (RMSE) and the Pearson Correlation Coefficient (PCC) were calculated between observed and predicted municipal counts. RMSE assessed prediction accuracy (i.e., how closely predictions matched observed values), while PCC measured the linear association between the observed and predicted values.

To compare the performance of GRF and RF, the distributions of RMSE and PCC were analysed across the eight combinations of modelling procedures. A one-way ANOVA and the associated Tukey's honestly significant difference (HSD) were used to determine statistically significant differences between methods. Observed and predicted values were standardised to the interval [0;1] using min-max standardisation [34], and correlation plots were then generated to compare outputs of the different methods.

### Post-processing

The means and standard deviations of all the predicted maps were calculated to produce a final raster. This final raster served as a weighting layer to redistribute census data across the territory. This post-processing procedure, inspired by the GLW methodology, adjusted the predicted animal counts to ensure that the sum of all pixel values equalled the total number of animals in the Guadeloupe Archipelago. The adjusted pixel value $X_i$ was calculated using the following equation (Eq.2):

$$X_i = C \times \frac{x_i}{\sum_1^n x_i} \quad \text{Eq. 2}$$

Where $C$ represents the total animal count from the DAAF Census, n is the total number of pixels in the study area and $x_i$ is the predicted animal count for the $i^{th}$ pixel.

The entire methodology was implemented using R (v4.2.3) [35].

## Results

### Produced maps and predictor importance

To run the models, 4 971 breeders, representing 34 790 cattle were finally used. A Moran I of 0.61 ($p < 0.05$) was found using raw cattle densities data. Results of data preparations conducted before modelling (city animal redistribution and corrected density calculation) are displayed in supplementary materials (S6 in S1 File). Corrected cattle densities gave a Moran I of 0.51 ($p < 0.05$). The predicted maps generated by the GRF model are shown in Fig 4 (RF results are not displayed as they exhibit similar pattern to those of GRF). The simulated animal count per pixel ranged from 0 to 2.93 using the random sampling method, and from 0 to 2.5 using the average sampling method.

For both sampling methods, the models effectively reproduced the observed distribution of cattle. The highest densities were observed on Marie-Galante and in the central region of Grande-Terre. These patterns correspond to the municipal densities calculated after redistributing animals from urban areas.

The Spearman correlations between cattle densities and the predictors, as well as the correlations between the predictors, are shown in supplementary materials (S7 in S1 File). The PVI is summarised in Fig 5. Significant positive correlations ($p < 0.05$) were found between cattle density and the breeding area predictor, which had the greatest impact on model I performance for both the RF and GRF algorithms. Significant positive correlations were also observed with agricultural land cover, which did not appear to be an important predictor in either model predictions. Distance to barn, distance to pond, daytime land-surface temperature standard deviation (DLST SD), slope and TRI were initially significantly negatively correlated with cattle densities ($p < 0.05$). The predictor Distance to barn was important in model I particularly

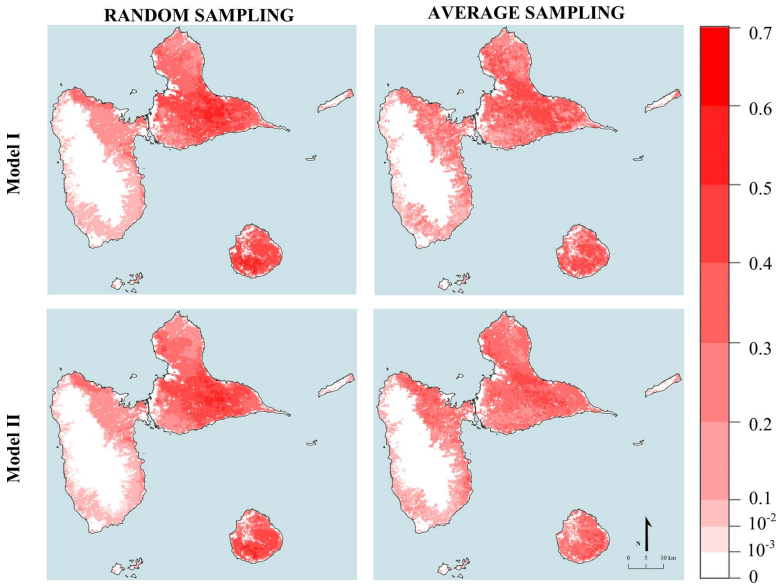


**Fig 4. Predicted maps produced with GRF using two sampling methods and two predictor dataset.** The maps show that models constructed using the random sampling methodology appear capable of predicting cattle density values over a wider range than those constructed using the average sampling. Furthermore, the highest predicted cattle densities were found in Marie-Galante and central Grande-Terre, with consistent results across the different methodologies. Created by Victor Dufleit using BD_TOPO® shapefile data. BD_TOPO® are open access data published by French National Geographic Service under "Etalab 2.0" licence.

https://doi.org/10.1371/journal.pone.0324695.g004

for the average sampling method. DLST SD proved to be an important predictor when applying the random sampling method particularly with model II. Slope appeared to be an important predictor in model II whatever the sampling procedure. Distance to pond and TRI did not appear to be important predictors regardless of modelling method. Notably, grass cover, distance to river and mean night temperature showed relevance (Fig 5) in the random sampling method, although no significant correlations with cattle density were initially detected.

### Goodness of Fit measures (GOF)

The RMSE and PCC metrics produced consistent results across the tested methodologies. Fig 6 shows that higher PCC values mean lower RMSE values, suggesting that the model reliably follows the observed values. Both sampling methods resulted in high Pearson's *r* values (> 0.9). Among the GOF metrics, the random sampling method demonstrated superior predictive performance, regardless of the random forest type (GRF and RF) or the set of predictors used (model I or model II). Within both average and random sampling methods, the GRF model provided more accurate predictions of cattle density compared to the RF model.

The inclusion of land cover predictors into the model (model I vs. model II) did not improve the GOF metrics when the random sampling method was used. Conversely, when the average sampling method was used, model I showed significantly higher PCC values than model II.

Fig 7 compares the standardised simulated animal counts with observed counts at the municipal level. For both models, the two random forest methods combined with the two sampling methods, overestimated animal counts in municipalities with the lowest observed animals and underestimated it in the municipalities with the highest observed counts. The random sampling method provided the best fit, better simulating higher animal counts in municipalities with the largest observed population.

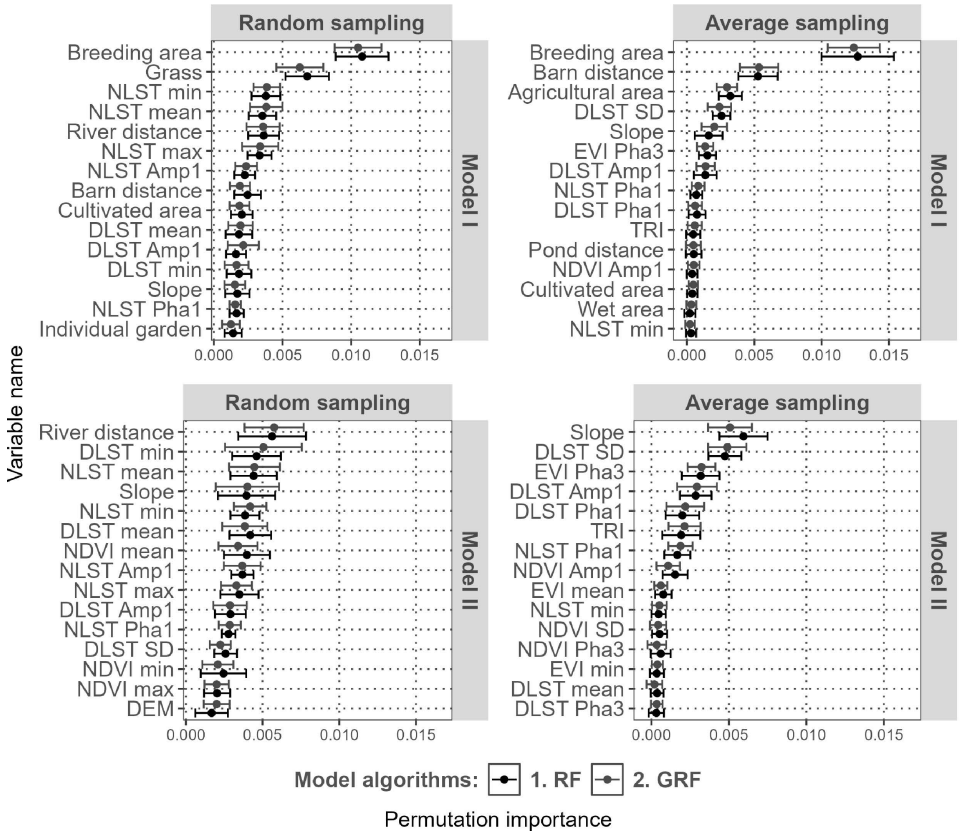


**Fig 5. Permutation variable importance calculated by RF and GRF models.** Only the 15 most important variables are presented here. DLST = Day Land Surface Temperature; NLST = Night Land Surface Temperature; SD = standard deviation; Amp1 = Amplitude of the annual harmonic; Amp2 = Amplitude of the biannual harmonic; Amp3 = Amplitude of the triannual harmonic; Pha1 = Phase of the annual harmonic; Pha2 = Phase of the biannual harmonic; Pha3 = Phase of the triannual harmonic (16).

https://doi.org/10.1371/journal.pone.0324695.g005

## Discussion

The RF and GRF census disaggregation methods were successfully applied to the Guadeloupean territory, resulting in the first cattle density map with a spatial resolution of 225 meters. This work represents a significant advance in livestock census disaggregation. Indeed, to our knowledge, this method had never before been implemented on such a small territory, as evidenced by the case study of the Guadeloupean archipelago.

### RF vs GRF

The GRF significantly outperformed classical RF in terms of predictive accuracy, regardless of datasets of predictors used and whether average or random sampling was used. Similar improvements in predictive performance with GRF over RF have been reported in different research areas [36], but with an increase in computational time [37]. This highlights the critical role of spatial heterogeneity in improving predictive accuracy for population density modelling.

In fact, since spatial autocorrelation was observed in our census dataset with Moran I = 0.61 after city animal redistribution (S6B in S1 File) and a Moran I = 0.51 after calculation of density with suitable pixels (S6D in S1 File). The inclusion of spatial proxies for our case study was relevant.

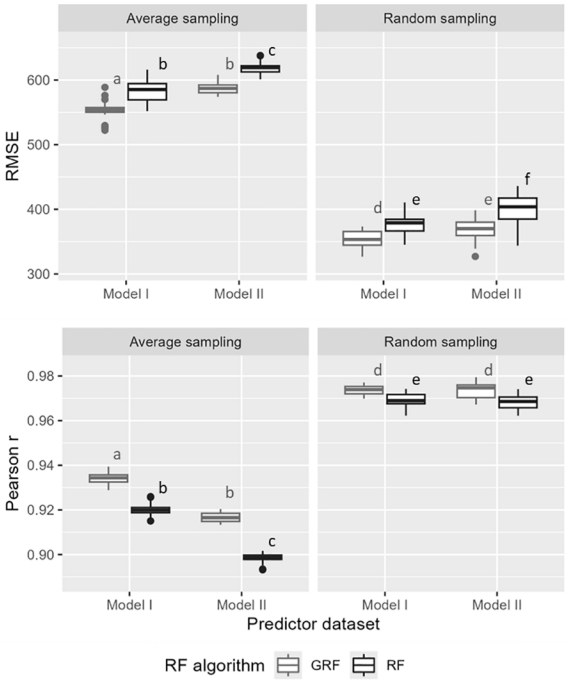


**Fig 6. Synthesis of RMSE and Pearson correlation coefficients for the tested methodologies.** Different letters indicate statistically significant differences in the means, as determined by one-way ANOVA followed by Tukey's HSD.

https://doi.org/10.1371/journal.pone.0324695.g006

### Average sampling vs Random sampling

We also assessed the effect of the choice of two methods for sampling predictors on the random forest models predictive accuracy. Random sampling showed better GOF metrics and produced a wider range of predicted density values (0 to 2.92 animals per pixel compared to 0 to 2.46 with average sampling). This improvement may be attributed to the higher predicted animal counts in the most densely populated municipalities (Fig 7). However, no major differences were observed for less densely populated municipalities. The better performance of the random sampling method could also reflect that this method might better took into account the spatial heterogeneity within municipality compared with the average sampling using the suitability mask to set cattle density of training points located in unsuitable area to 0. It should also be noted that this sampling methodology is more computationally demanding, because the number of entities used to build the model is artificially increased by creating a dense point layer to capture predictor values and local cattle densities. In context with limited computational resources, this parameter should be optimised to achieve accurate predictions with a minimal sampling point density. In this study, a density of 50 points per 10,000 ha was used. This parameter was not optimised, as priority was given to other aspects of the experimental process. Nevertheless, based on the stochastic sampling process, some assumptions can be made, for example, less dense point layers are likely to produce greater variability in RF and GRF model outputs, as they capture a smaller portion of potential training points. A possible approach to optimize this parameter would be to incrementally increase point density and monitor output variability until these values reach a user-defined quality threshold.

The model output maps showed differing density patterns between the two sampling methods. This may be due to differences in the ranking of predictors according to the variable permutation importance when using different sampling methods. With average sampling, predictors previously identified as significantly correlated with cattle density (S7 in S1 File) also had high importance in RF models such as breeding area. Conversely, random sampling emphasised the

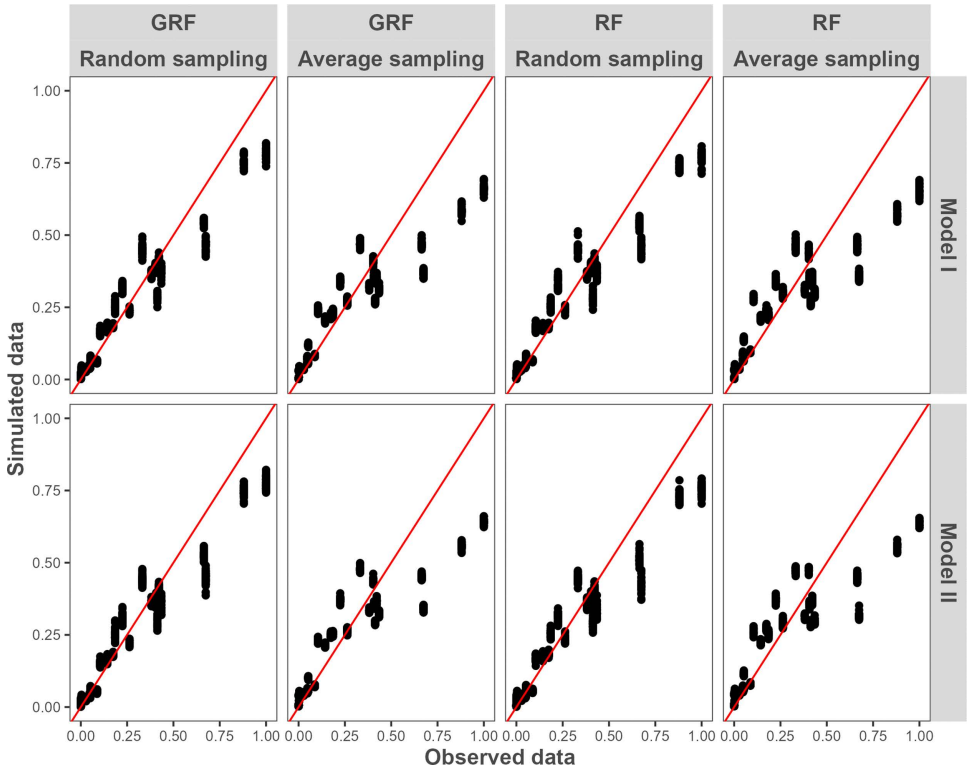


**Fig 7. Standardized observed and simulated municipal cattle counts.** The red line represents the equation f(x) = x.

https://doi.org/10.1371/journal.pone.0324695.g007

importance of alternative predictors that may not have previously shown significant correlations with cattle density. With the whole dataset of predictors (model I) breeding area remained the most important predictor regardless of the sampling methodology used. For the other predictors of importance, they differed dramatically comparing both sampling methodologies regardless of the datasets of predictors used. This may highlight how predictor sampling methodology can influence model outcomes and the relative importance of variables. Nevertheless, several studies [38–40] have questioned the relevance and possible interpretation of the PVI when correlations between predictors exist. We chose to follow the initial GLW methodology for the validation of our study to facilitate comparison with GLW results. However, implementing SHAP analysis in the validation process could improve the census disaggregation methodologies by providing more accurate estimates of the predictor contributions to predictions, as proposed by Lee et al. [41] in 2024.

### Predictors

Regarding the predictors and the PCC GOF metrics (Fig 6), when using the random sampling method, there was no difference between the full predictor dataset (model I) compared to model II without the Karucover derived predictors. However, when using the average sampling method, an improvement in prediction accuracy was observed when using the model I datasets. Nevertheless, as demonstrated previously, random sampling yielded superior results in both the model I and the model II datasets. This finding suggests the potential for achieving comparable prediction accuracy with fewer predictors. This could be particularly beneficial in the Caribbean region, where the model is intended for wider application. Even in upper-middle income countries, high-resolution land cover datasets such as Karucover are often not available, so the ability to perform well with limited predictors is a practical and valuable consideration.

### Method

In line with Da Re et al. (2020) [5], this study used the finest available census data in France—at the municipal level—for model development. However, these data are based on breeders' declarations, and associate cattle with the breeders' residence, which does not always reflect the actual location of their animals. To address this geographical bias, particularly in urban areas, the methodology employed in this study included specific adjustments. The "urban animal" redistribution equation (Eq. 1) considered in this analysis assumes uniform accessibility between urban and adjacent rural municipalities based on the shared border length. However, this process does not account for potential geographic barriers such as the steep "Rivière des Pères" valley between the Basse-Terre and Baillif municipalities for example. Alternative proxies for redistributing urban animals, such as the number of roads connecting municipalities A and B and their relative importance (e.g., average daily traffic), might have provided a more realistic representation. However, shared border length was a readily available dataset that did not require additional investigation to account for accessibility differences between urban areas and their adjacent rural municipalities. Nevertheless, the question remains as administrative boundaries do not always correspond to biological realities of cattle farming.

It is also noteworthy that, in all the explored methods, models exhibited a tendency to overestimate cattle counts in municipalities where low densities are observed and underestimate cattle counts in high cattle densities municipalities. This regression-to-the-mean pattern was previously described in census disaggregation methodologies [5,6], and may stem from different factors. One of the them could be the limitation of machine learnings approach, such as RF and GRF, to extrapolate estimated values (in our case cattle density) outside of the range defined by the training dataset. Evidence for this phenomenon can be found in the range of observed municipalities, which varied from 0 to 0.65 head/ha. In comparison, the predicted pixel density values ranged from 0 to 0.58 head/ha. This results in a maximum value of 2.93 animals per pixel. This value may appear low in the context of field observations. This is due to the fact that the number of individuals observed on surfaces measuring less than 225 m x 225 m can reach 20 or more. It is therefore hypothesised that the value may reflect a temporal mean rather than an exact value. It is important to note that the proposed predictive maps, constructed utilising RF and GRF models, must be interpreted as statistical estimates of the potential distribution of cattle within the Guadeloupe territory. Notwithstanding the aforementioned limitations, this approach facilitated estimations and discrimination of potential high/low cattle density areas in Guadeloupe.

Moreover, as highlighted by Robinson et al. (2014) [3], validating model predictions at the pixel level would require more detailed census data collected consistently over at least one year. Such an effort, however, would demand substantial coordination and resources from both the scientific community and the livestock sector. Participatory mapping (i.e., asking breeders to report the locations of their animals or the number of head per plot) could potentially help to create such datasets and would enhance the overall methodology. This approach would provide data for model validation, and could also enable the development of alternative predictive models beyond the downscaling approach used here.

## Conclusion

This study highlighted the value of applying census disaggregation methods to smaller territories, such as Guadeloupe, which has a relatively small number of administrative units. The maps produced by this study will have significant applications in the field of epidemiology. They could also find applications in conservation studies, for example, to simulate grazing pressure as proposed by Wang et al. (2025) [42]. The high-resolution livestock density maps produced at a 225m scale represent a significant improvement over previously available GLW maps [4], which offered a resolution of 5 minutes of arc (approximately 9 km x 9 km over the Guadeloupean archipelago). This work also provided an updated source of livestock data for the region, using recent census information. The methodology developed here will be adapted to other Caribbean islands; however, cattle census data and relevant predictors will need to be collected in these regions. The

random sampling method of predictor selection showed advantages over the average sampling method, producing maps of comparable quality with a reduced dataset, even in the absence of high-resolution land cover data. This approach may be particularly useful in regions lacking detailed datasets such as Karucover. A potential next step to improve livestock density mapping would be to be independent of census data collected at the administrative level. Mapping densities assessments based on grazing or production areas, and explicitly identifying areas where livestock are absent such as, urban, natural or protected zones, could provide valuable insights for training datasets. This approach could also serve as a means to validate the downscaling of administrative census data to grid scale, further improving the accuracy and applicability of territorial livestock mapping models.

## Supporting information

**S1 File. Supplementary figures and tables.** Figures and tables of this document are referenced in the text as S1 to S7 in this manuscript.
(PDF)

## Acknowledgments

We acknowledge the DAAF Guadeloupe for providing list of breeder contacts and the animal census. We would also like to thank the geological and mining research bureau (Bureau de Recherche Géologique et Minière; BRGM) of Guadeloupe for providing us with the MNT layer.

## Author contributions

**Conceptualization:** Victor Dufleit, Laure Guerrini, Daniele Da Re.

**Data curation:** Victor Dufleit.

**Formal analysis:** Victor Dufleit.

**Methodology:** Victor Dufleit, Laure Guerrini, Marius Gilbert, Daniele Da Re.

**Project administration:** Eric Etter.

**Resources:** Victor Dufleit, Marius Gilbert, Daniele Da Re.

**Software:** Marius Gilbert, Daniele Da Re.

**Supervision:** Laure Guerrini, Eric Etter.

**Validation:** Victor Dufleit, Laure Guerrini, Daniele Da Re.

**Visualization:** Victor Dufleit, Laure Guerrini.

**Writing – original draft:** Victor Dufleit, Laure Guerrini, Eric Etter.

**Writing – review & editing:** Laure Guerrini, Daniele Da Re, Eric Etter.

## References

1. Franceschini G, Robinson TP, Morteo K, Dentale D, Wint W, Otte J. The global livestock impact mapping system (GLIMS) as a tool for animal health applications. Vet Ital. 2009;45(4):491–9. PMID: 20391413
2. Robinson TP, Franceschini G, Wint W. The food and agriculture organization's gridded livestock of the world. Vet Ital. 2007;43(3):745–51. PMID: 20422554
3. Robinson TP, Wint GRW, Conchedda G, Van Boeckel TP, Ercoli V, Palamara E, et al. Mapping the global distribution of livestock. PLoS One. 2014;9(5):e96084. https://doi.org/10.1371/journal.pone.0096084 PMID: 24875496
4. Gilbert M, Nicolas G, Cinardi G, Van Boeckel TP, Vanwambeke SO, Wint GRW, et al. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Sci Data. 2018;5:180227. https://doi.org/10.1038/sdata.2018.227 PMID: 30375994

5. Da Re D, Gilbert M, Chaiban C, Bourguignon P, Thanapongtharm W, Robinson TP, et al. Downscaling livestock census data using multivariate predictive models: sensitivity to modifiable areal unit problem. PLoS One. 2020;15(1):e0221070. https://doi.org/10.1371/journal.pone.0221070 PMID: 31986146
6. Nicolas G, Robinson TP, Wint GRW, Conchedda G, Cinardi G, Gilbert M. Using random forest to improve the downscaling of global livestock census data. PLoS One. 2016;11(3):e0150424. https://doi.org/10.1371/journal.pone.0150424 PMID: 26977807
7. Almeida FPS, Castelli M, Côrte-Real N. Leveraging feature sets and machine learning for enhanced energy load prediction: a comparative analysis. Emerg Sci J. 2024;8(6):2120–43.
8. USDA. Assessing the Risk of Arthropod-Borne Disease in the Caribbean and the Americas (RACE; Risk of Arthropod-borne diseases in the Caribbean). Agricultural Research Service; 2025. https://www.ars.usda.gov/research/project/?accnNo=440817
9. Milà C, Ludwig M, Pebesma E, Tonne C, Meyer H. Random forests with spatial proxies for environmental modelling: opportunities and pitfalls. Geosci Model Dev. 2024;17(15):6007–33. https://doi.org/10.5194/gmd-17-6007-2024
10. Saha A, Basu S, Datta A. Random forests for spatially dependent data. J Am Stat Assoc. 2023;118(541):665–83.
11. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 2018;6:e5518. https://doi.org/10.7717/peerj.5518 PMID: 30186691
12. Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, et al. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Intern. 2019;36(2):121–36. https://doi.org/10.1080/10106049.2019.1595177
13. Neumann K, Elbersen BS, Verburg PH, Staritsky I, Pérez-Soba M, De Vries W, et al. Modelling the spatial distribution of livestock in Europe. Landsc Ecol. 2009;24(9):1207–22.
14. Li X, Hou J, Huang C. High-resolution gridded livestock projection for Western China based on machine learning. Remote Sens. 2021;13(24):5038. https://doi.org/10.3390/rs13245038
15. Tian Y, Yue T, Zhu L, Clinton N. Modeling population density using land cover data. Ecol Model. 2005;189(1–2):72–88.
16. Bleuse N, Mandar C. Le régime pluviométrique de la Guadeloupe. Météorologie. 1993;8(4):42.
17. Fanchone A, Nelson L, Dodet N, Martin L, Andrieu N. How agro-environmental and climate measures are affecting farming system performances in Guadeloupe?: Lessons for the design of effective climate change policies. Int J Agric Sustain. 2022;20(7):1348–59.
18. Galan F, Julien L, Duflot B. Panorama filières animales et typologie systèmes Guadeloupe. 2008. https://www.odeadom.fr/wp-content/uploads/2018/04/Panorama-fili%C3%A8res-animales-et-typologie-syst%C3%A8mes-Guadeloupe.pdf
19. Daaf G. Sois son héros: on peut tout dire à son chien, sauf débrouille-toi tout seul. 2019. https://daaf.guadeloupe.agriculture.gouv.fr/sois-son-heros-on-peut-tout-dire-a-son-chien-sauf-debrouille-toi-tout-seul-a852.html
20. Kermorgant P. Principales maladies et dangers sanitaires pour les élevages de Guadeloupe, de Saint-Martin et de Saint-Barthélemy. 2018. https://daaf.guadeloupe.agriculture.gouv.fr/principales-maladies-et-dangers-sanitaires-pour-les-elevages-de-guadeloupe-de-a862.html
21. Institut National Geographique I. BDTOPO®. 2022. Accessed 2023 February 6. https://geoservices.ign.fr/bdtopo
22. Mimoto N, Zitikis R. The Atkinson index, the moran statistic, and testing exponentiality. J J Stat Soc. 2008;38(2):187–205.
23. Rattanarat J, Jaroensutasinee K, Jaroensutasinee M, Sparrow EB. Government policy influence on land use and land cover changes: a 30-year analysis. Emerg Sci J. 2024;8(5):1783–97.
24. Robillard A. Occupation du sol à grande échemme en 2 dimendion de Guadeloupe. 2022. https://catalogue.karugeo.fr/geosource/consultation?id=21421625
25. Didan K. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 global 250m SIN grid V061. NASA EOSDIS Land Processes DAAC; 2021. https://modis.gsfc.nasa.gov/data/dataprod/mod13.php
26. Zhengming W, Hook S, Hulley G. MOD11A2 MODIS/Aqua land surface temperature/Emissivity 8-Day L3 global 1km SIN grid V061. NASA EOSDIS Land Processes DAAC. 2021. https://modis.gsfc.nasa.gov/data/dataprod/mod11.php
27. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW, et al. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. PLoS One. 2008;3(1):e1408. https://doi.org/10.1371/journal.pone.0001408 PMID: 18183289
28. Rogers DJ, Hay SI, Packer MJ. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. Ann Trop Med Parasitol. 1996;90(3):225–41. https://doi.org/10.1080/00034983.1996.11813049 PMID: 8758138
29. Bureau de Recherche Géologiques et Minières. Bureau de recherche géologiques et minières (BRGM). 2023. https://www.brgm.fr/en
30. Wright MN. ranger: a fast implementation of random forest for high dimensional data in C++ and R. J. Stat. Softw. 2017;77(1):1-17.
31. Kalogirou S, Georganos S. Spatial machine learning. 2022. https://CRAN.R-project.org/package=SpatialML
32. Georganos S, Kalogirou S. A forest of forests: a spatially weighted and computationally efficient formulation of geographical random forests. IJGI. 2022;11(9):471.
33. Wright MN, Ziegler A, König IR. Do little interactions get lost in dark random forests?. BMC Bioinformatics. 2016;17:145. https://doi.org/10.1186/s12859-016-0995-8 PMID: 27029549
34. Mohamad IB, Usman D. Standardization and its effects on k-means clustering algorithm. RJASET. 2013;6(17):3299–303.

35. R Core Team. R: a language and environment for statistical computing. Vienna, Austria. 2023.
36. Quevedo RP, Maciel DA, Uehara TDT, Vojtek M, Rennó CD, Pradhan B, et al. Consideration of spatial heterogeneity in landslide susceptibility mapping using geographical random forest model. Geocarto Inter. 2021;37(25):8190–213. https://doi.org/10.1080/10106049.2021.1996637
37. Lotfata A, Grekousis G, Wang R. Using geographical random forest models to explore spatial patterns in the neighborhood determinants of hypertension prevalence across chicago, illinois, USA. Environ Plann B: Urban Analyt City Sci. 2023;50(9):2376–93. https://doi.org/10.1177/23998083231153401
38. Efron B. Prediction, estimation, and attribution. Int Statistical Rev. 2020;88(S1). https://doi.org/10.1111/insr.12409
39. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27(3):659–78.
40. Debeer D, Strobl C. Conditional permutation importance revisited. BMC Bioinformatics. 2020;21(1):307. https://doi.org/10.1186/s12859-020-03622-2 PMID: 32664864
41. Lee E, Ong TS, Lee Y. Evaluating household consumption patterns: comparative analysis using ordinary least squares and random forest regression models. HighTech Innov J. 2024;5(2):489–507. https://doi.org/10.28991/hij-2024-05-02-019
42. Wang Y, Huang H, Tian Y, Yang G, Li L, Yuan C, et al. A grazing pressure mapping method for large-scale, complex surface scenarios: integrating deep learning and spatio-temporal characteristic of remote sensing. Elsevier BV; 2025. https://doi.org/10.2139/ssrn.5087520