Open Journal of Ecology, 2020, 10, 243-263
https://www.scirp.org/journal/oje
ISSN Online: 2162-1993
ISSN Print: 2162-1985

Multi-Level, Multi-Scale Modeling and Predictive Mapping for Jaguars in the Brazilian Pantanal

Eve Bohnett[1,2,3], Dave Hulse[1,3], Bilal Ahmad[4], Thomas Hoctor[1,2]

[1]Department of Landscape Architecture, University of Florida, Gainesville, United States of America
[2]Center for Landscape Conservation Planning, University of Florida, Gainesville, United States of America
[3]Florida Institute for Built Environment Resilience, University of Florida, Gainesville, United States of America
[4]Institute of Agriculture Sciences and Forestry, University of Swat, Swat, Pakistan
Email: evebohnett@ufl.edu

How to cite this paper: Bohnett, E., Hulse, D., Ahmad, B. and Hoctor, T. (2020) Multi-Level, Multi-Scale Modeling and Predictive Mapping for Jaguars in the Brazilian Pantanal. *Open Journal of Ecology*, **10**, 243-263.
https://doi.org/10.4236/oje.2020.105016

Received: April 4, 2020
Accepted: May 8, 2020
Published: May 11, 2020

Abstract

Multi-level multi-scale resource selection models using machine learning were compared and contrasted for generating predictive maps of jaguar habitat (*Panthera onca*) in the Brazilian Pantanal. Multiple spatial scales and temporal movement levels were run within several analytical modeling frameworks for comparison. Included in the analysis were multi-scale raster grains (30 m, 90 m, 180 m, 360 m, 720 m, 1440 m) and GPS collaring temporal movement levels (point, path, and step). Various analytical methods were used for comparison of models that could accommodate data structural levels (group, individual, case-control). Models compared included conditional logistic regression, generalized additive modeling (GAM), and classification regression trees, such as random forests (RF) and gradient boosted regression tree (GBM). The goals of the study were to discuss the potential and limitations for machine learning methods using GPS collaring data to produce predictive habitat suitability mapping using the various scales and levels available. Results indicated that choosing the appropriate temporal level and raster scale improved model outputs. Overall, larger level analytical modeling frameworks and those that used multi-scale raster grains showed the best model evaluation with the inherent condition that they predict a broader scale and subset of data. The identification of the appropriate spatial scale, temporal scale and statistical model need careful consideration in predictive mapping efforts.

Keywords

Machine Learning, Movement Ecology, Habitat Selection, Resource Selection, Multiple Levels, Multiple Scales, Predictive Models, Gradient Boosting Method, Random Forest

1. Introduction

Landscape patterns and processes occur within many spatial and temporal dimensions, and scale is a lens through which to view those dimensions. Landscape ecology examines ecological processes and landscape scales through modeling approaches. Ecological models depend on the entities measured (e.g. which organisms, ecosystems or ecological processes), variables measured (e.g. which environmental or climate covariates), and the processes linking entities and variables. Additionally, they can also respond to changes in spatial extent, spatial grain, temporal duration, and temporal grain used to measure these entities and variables [1] [2] [3]. Wildlife ecologists often employ resource selection models for use-available data with scales defined through nested hierarchical orders of selection, for example, the geographical range of the species (Level I), the home range (Level II), or patch level habitat selection (Level III) [4]. Following resource selection modeling frameworks, one of the approaches researchers apply to integrate scales are buffers of various sizes around the data points to average environmental covariates within a given area. This allows one to assess the effective scale at which the environment shapes animal behavior [5]. McGarigal *et al.* (2016) proposed a multi-scale, multi-level modeling framework to consider the various spatial and temporal scales necessary to address spatial dependencies within various levels of selection. This conceptual framework develops scale optimized multi-level modeling in which multiple scales are tested simultaneously within each level of resource selection and the scales where the effect of each environmental variable most strongly affects selection can emerge [6] [7]. By quantifying the patterns and processes that naturally occur at different scales in time and space, we can reach conclusions regarding the key ecological and evolutionary processes that compose landscapes.

1.1. Representation of Movement in Resource Selection Approaches

Resource selection may also depend on the how the process of interest is represented. For GPS data, resource selection functions may be based on different sampling units used to represent animal movement: points (locations in space), steps (displacements), and paths (sequences of displacements). Binary response variables include (1) animal movement data and (0) background points generated on the landscape within several levels by using a “used vs. available” sampling design broadly referred to as resource selection function (RSF) [8]. Fine-scale GPS collaring trajectories allow the extension of traditional RSF’s to point, step, and path selection functions. Studies with these movement representations should also assess multiple scales along with levels of selection [7]. In terms of classic habitat selection, for the point selection functions, points are subset into different levels within a broader geographic range, the species range, or the home range [4]. These levels are different for step and path selection functions. Animal steps and paths may be sectioned into periods that represent one or several displacement events, e.g. hourly or daily sections, respectively. Then, background points are generated for steps or paths *not* traveled, yet were available for animal movement and not chosen by the animals that were sampled [8] [9], and habitat selection inferences are made by comparing realized vs. not used steps or paths. Temporal differences in the generation of background points for path selection have been shown to optimize the scale of effect for large carnivore dispersal studies [10] [11]. For instance, paths refer to sets of locations along 12 hours, 24 hours, or several days for the analysis, depending on the research questions. Overall, temporal separation (point, hourly, daily, or otherwise) in the data can potentially reveal behavioral differences in the temporal scale of animal habitat use.

1.2. Statistical Analysis and Modeling Options

Parametric, semi-parametric, and non-parametric models are important options for analysis on movement ecology. Parametric models generally aim at estimating cause-effect relationships between an organism's movement features and the environmental covariates. For example, conditional logistic regression is often used in point and step selection approaches to find out which sets of environmental conditions are selected or avoided by organisms. However, these parametric approaches are not recommended for predictive mapping or predicting results on new data. Predictive mapping efforts that utilize non-parametric approaches such as machine learning, for example, are most suitable for predicting on new data [12] [13].

Emerging studies on machine learning (ML) methods for landscape ecology have used classification and regression trees (CART), showing they outperform other methods for multi-scale modeling [14]. At continental scales where GPS collaring data may be sparse, random forest (RF) models performed adequately [15]. Various statistical methods (Occupancy, GAM, CART) have been used for single scale or multi-scale modeling, determining the effect of an environmental covariate or distance effect on models [16].

Previous research has also applied ML methods like RF for movement studies for point selection functions for mule deer [17] and Florida panther [18]. Zeller (2018) explored non-parametric and semi-parametric statistical modeling approaches in resource selection functions to derive predictive distribution maps for large carnivore conservation [19]. They compared RSFs generated from points, steps, and paths using conditional logistic regression, to point selection functions using machine learning methods [19].

How predictive modeling, like machine learning methods, may apply to step and path selection functions is unknown. This study extends previous research on movement ecology and machine learning by applying non-parametric and semi-parametric techniques to multi-scale, multi-level methods for GPS collaring data using also path and step selection functions. Step and path selection functions traditionally use conditional logistic regression, and have recently begun exploring alternative methods such as machine learning. We believe that multi-scale predictive modeling and mapping may improve through the use of machine learning and generalized additive modeling approaches. ML methodologies are becoming more useful to ecologists collating big data of high dimensions from data repositories. Both remote sensing data (e.g. Google Earth Engine) and animal movement data (e.g. MOVEBANK) repositories are making data widely available, which also demands for more sophisticated processing and modeling approaches [20] [21] [22] [23].

Here we aimed at understanding how parametric, semi-parametric, and non-parametric models can contribute to habitat selection estimates and predictive mapping (**Figure 1**). Since steps and paths are most common to represent third-order habitat selection at the level of the resource patches, here we focus at this level of analysis. In this study, multi-scale, multi-level modeling of habitat selection is explored using jaguars in the Brazilian Pantanal as a case study.

2. Materials and Methods

2.1. Study Area

The study area is the Brazilian Pantanal region surrounding the Taiama Ecological

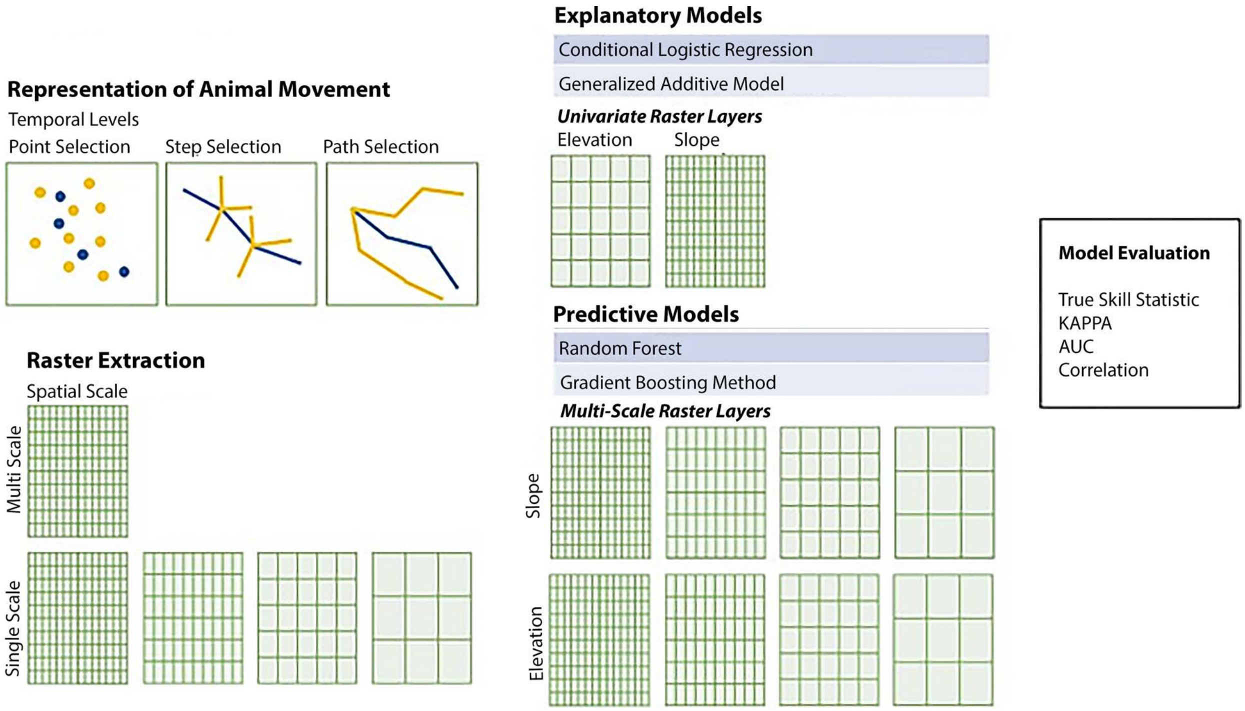

Figure 1. Chart to illustrate the point, step, and path selection functions subsets of the GPS collaring data. Illustrations of raster extraction scales of data extraction. Illustration of how many data layers are included in a univariate and multi-scale modeling workflow.

Station (17.712061S, 57.415956W.) The station is deep in the Brazilian Pantanal, the world's largest freshwater wetlands located in the state of Mato Grasso do Sul and Mato Grosso, in Western Brazil. Several major tributaries of the Paraguay River incur a seasonal flooding regime from January to July [24]. Vegetation is mainly semi-deciduous forest, open forest, closed forest, savanna (Cerrado), and aquatic or swamp terrain. The area is rich in biodiversity and has a high abundance of jaguars. In the Pantanal region, Taiama Ecological Station is a remote area with few roads or disturbances. It is an ideal place to study the jaguar habitat in its semi-natural form.

2.2. Environmental Variables

Remote sensing information was particularly useful in this study, and the availability of data layers globally has made compiling such large datasets much easier. A total of 12 raster-based environmental data layers were extracted from Google earth engine and open source data layers available online, namely elevation, slope, aspect, land use and land cover, forest non-forest, total canopy cover, roads, water and hydrology, human density, and cattle density (**Table 1**). Data

Table 1. Environmental covariates listed with layer name and original source, including some descriptive information where necessary.

Layer Name	Source
Elevation	SRTM Digital Elevation Model
Slope	SRTM Digital Elevation Model
Aspect	SRTM Digital Elevation Model
Global Cover	ENVISAT's Medium Resolution Imaging Spectrometer (MERIS) Level 1B Land Cover
Forest Non-Forest	JAXA L-band PALSAR SAR and ALOS mosaics using backscatter coefficient to determine "Forest" and "Non-Forest"
Total Canopy Cover	Landsat Vegetation Continuous Fields (VCF) total cover of vegetation of woody plants above 5 m in height.
SAR Land Cover	HH and HV L-band ALOS/PALSAR, and HH and HV C-band. RADARSAT-2 using a hierarchical object-based image analysis approach to hydrologically variant subregions.
Water	The Global Inland Water dataset shows inland surface water bodies, including fresh and saline lakes, rivers, and reservoirs.
Roads	Center for International Earth Science Information Network - CIESIN - Columbia University - Distance Raster generated from Shapefile with Global Roads Open Access Data Set, Version 1 (gROADSv1) 1980-2010. From this distance to roads were calculated.
Hydrology	WWF HydroSHEDS Raster - Hydrological data and maps based on Shuttle Elevation Derivatives at multiple scales.
Human Density	WorldPop Raster with estimated the number of people per hectare by 2015, with national totals adjusted to match the estimates of the UN population division.
Cattle Density	FAO, ILRI, the University of Oxford and the Université Libre de Bruxelles. Raster with Global Distribution of Livestock

were extracted for time periods suitable for the study, unless unavailable, then previous layers were used, for example the roads layer is current to 2010. Collinearity was checked using pairwise comparison for those models where this may be an issue, any variables that had greater than 0.7 pairwise correlation would be rejected. None of the variables met this criterion, and were uncorrelated. Machine learning models do not consider collinearity to be an issue so all data layers can be included.

2.2.1. Multi-Scale Environmental Data

The study attempts to understand how various modeling approaches perform, considering spatial information at multiple scales and model levels (Supplementary Information). Point, step (1 hour displacements), and path (24 hour trajectories) approaches were used to represent movement (Table 2(a)) at different model levels (group, individual ID strata, case-control) (Table 2(b)), furthermore incorporating multi-scale raster grain data to compare single grain (30 m) and multi-grain raster data (30 m, 90 m, 180 m, 360 m, 720 m, 1440 m) (Table 2(c)). For clarity, this refers to a study that investigates multiple levels of both temporal frequencies and levels using multi-scale raster grain covariates in the models. A combination of several multi-level multi-scale modeling definitions, where levels are hierarchies of organization in time or space, and scales as the scale and extent of the organization [25].

Table 2. Descriptions of the multi-levels and scales for GPS collaring temporal levels, data structural levels and raster grains. (a) GPS collaring temporal levels; (b) Model and data structural levels; (c) Spatial scale.

(a)

Model	Level
Point Selection	Home Range
Step Selection	1 hour steps
Path Selection	24 hour paths

(b)

Model	Resource Selection Order
Gradient Boosting Method	Group
Random Forest	Individual ID
GAM	Case-Control/Individual ID
Conditional Logistic	Case-Control/Individual ID

(c)

Model	Raster Scales
Gradient Boosting Method	All Scales 84 variables
Random Forest	All Scales 84 variables
GAM	Univariate Scales 12 variables
Conditional Logistic	Univariate Scales 12 variables

2.2.2. Raster Grain

To represent the raster layers at multiple scales, a Gaussian kernel smoother was used to average the layers at multiple extents (90 m, 180 m, 360 m, 720 m, and 1440 m), resulting in a total of 84 variables. If the original raster layers were not in a 30 m grain, they were disaggregated to 30 m for analysis. Specifically, the MERIS global land cover (300 m) grain and the SAR land cover (50 m) were disaggregated to 30 m. Assessing the functional grain of analysis, or the grain at which the organism is responding to the landscape, for connectivity studies has been useful to create resistance maps of habitat preferences [26]. Raster spatial grain is a problem in the multi-scale paradigm that is often not considered within multi-level, multi-scale studies for resource selection [27]. Although similar in technique to expanding distance buffers and averaging the pixels around a point or line.

The conditional regression model and GAM models were fit univariately to determine the adequate scale [7]. Model selection was performed using AIC and ΔAIC, building an optimal multi-scale model with one chosen raster grain for each environmental covariate. The RF and GBM models run with all multi-scale data layers because of the inherent tree system used to build the models, producing variable importance plots to determine the most valuable raster grain to the model.

2.3. Study Species

Modeling methods rely on niche habitat concepts for environmental covariates to construct accurate models of jaguar distributions on the landscape [28]. In other studies, jaguars exist in primary forest habitat, or areas with high forest cover, far from deforested patches or other human activities like cattle pastures, roads, or croplands. Jaguar populations are shown to decline with increasing human population density [29] and roads [30]. Jaguars prefer areas having topography with moderate slopes. They also prefer riparian areas with high amounts of water [31] [32].

2.4. Movement Data

GPS collars (Lotek Globalstar and Iridium Collars) for jaguars (n = 11, five females and six males) are the largest group of animals monitored in the Northern Pantanal. Data were made freely available by Morato *et al.* (2018), and capture procedures and permits were described in Morato *et al.* 2016. Monitoring occurred from October 2013 to February 2016 for 909 total days of data collection with individuals ranging from a minimum of 26 days and a maximum of 597 days. Collars were programmed to collect one relocation every hour, summing up 42,741 observed locations for all animals. Data collected followed protocols approved by Instituto Chico Mendes de Conservação da Biodiversidade (Ministério do Meio Ambiente, Brazil (ICMBio-SISBIO)). All procedures followed guidelines approved by the American Society of Mammologists [30].

2.4.1. Point Selection Functions (PSF)

Minimum convex polygons have been the traditional method for home range estimation, drawing a polygon around the point locations for the animal, to be used as the area available to animals in their routine movements. This technique provided a crude estimate of the home range, most commonly reported with 95 percent of data points [33]. Here minimum convex polygons 95% (MCP95) were determined using the MCP function in the *adehabitatHR* package in R [34]. Random points within the MCP were generated using the *dismo* package in R [35].

2.4.2. Step Selection Functions (SSF)

The GPS collars captured one point every hour. These hourly point data were subset into "steps", using the first point as the start of the step, and the next point as the end of the step. Background steps had the same original step distance projected into a different angle around the starting point. Steps from hourly data had a total of 85,388 presences and absences, generating one background step for each step. Step selection was performed in program R, using the package AdeHabitatLT [34].

2.4.3. Path Selection Functions (PathSF)

Animal trajectories were subset into 24-hour time sequences of hourly steps, generating longer paths for daily intervals, with a total of 4409 present and absent paths. Absences were generated using a correlated random walk (CRW) method. The CRW began at the starting point, simulating a trajectory of a similar length at alternate angles. CRWs are a completely randomized simulation of blind jaguar movement at any chance direction. CRW does not account for any decisions the jaguars normally encounter in time or space.

Nonetheless, the correlated random walk provided a randomized path to understand the simplest baseline from which to compare the actual jaguar movement. The mean of all steps values in the path aggregated to one single value to represent each path extracted covariate. Paths were generated in program R, package SiMRiv [36].

2.5. Study Design

Since the use-available (presence-absence) data for steps and paths were generated from the original data points, then these steps and paths can be paired in a case-control framework for the analysis. The conditional logistic and GAM models fit in a case-control framework allowing for a direct comparison between these two models.

However, the RF model used a higher order level at the individual ID strata and was not case-control. Therefore, RF performs the analysis for all steps and all background steps of individual animals during the duration of the study, and not matching presence steps with the generated absence steps them directly. The GBM algorithm is an even larger level using the entire group's paths and steps together without any subsetting into individual strata or case-control (Table 3). The two levels (group and individual stratum) inherent in the machine learning algorithms make models not directly comparable to the conditional logistic and GAM which have a case-control level.

2.6. Statistical Modeling Frameworks

2.6.1. Conditional Logistic Regression

Traditional resource selection functions for GPS collaring studies use an explanatory modeling approach such as conditional logistic for case-control for the steps and paths. Hooten *et al.* (2014) developed a point process model. In a very basic interpretation of the model, the probability density function for use $[x]_u$ is equal to a weighted distribution of availability $[x]_a$, then further indexing resource observations by relocation at time t (Equation (1) and Equation (2)) [37].

$$x\left[\left(s_t\right)\right]_u = \frac{g\left(x\left(s_t\right),\beta\right)\left[x\left(s_t\right)\right]_a}{\int g\left(x\left(s\right),\beta\right)\left[x\left(s\right)\right]_a ds} = \left[x\left(s_t\right)\middle|\beta\right]_u \tag{1}$$

The likelihood is maximized for resource coefficients β.

$$\prod_{t=1}^{T}\left[x\left(s_t\right)\middle|\beta\right]_u \tag{2}$$

A vector of resource covariates, β is a set of regression coefficients, x, a normalizing constant, and $g\left(x,\beta\right)$ is a resource selection function [37]. The weighted distribution framework is shown to account for the high amount of autocorrelation in GPS telemetry data [8]. All conditional logistic regression models used the mclogit package in program R. The mclogit package includes case-control and individual level strata for model fitting.

2.6.2. Generalized Additive Model

The generalized additive models are a semi-parametric extension of the generalized linear models that allow for non-linear functions of the environmental covariates. This method assumes that functions are additive and components smoothed. It estimates an additive approximation to the multivariate regression function, employing univariate smoothers and using individual estimates to

Table 3. Data Structures of models that use group level, individual id, and case-control levels.

Data Structure					
Group Level - GBM	Individual ID Level - Random Forest		Paired Case Control Level - Conditional Logistic Regression - Generalized Additive Model		
Use	Use	Individual ID	Use	Individual ID	Case Control
1 (Present)	1 (Present)	1	1 (Present)	1	1
0 (Absent)	0 (Absent)	1	0 (Absent)	1	1
1 (Present)	1 (Present)	3	1 (Present)	3	2
0 (Absent)	0 (Absent)	3	0 (Absent)	3	2

explain relationships between variables.

$$y_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_q(x_{qi}) + \epsilon_i \tag{3}$$

The smoothing splines, thin-plate splines, cubic splines, and splines with variable knots ($k = 3$, $k = 8$) applied to applicable covariates (Equation (3)). In this way, the environmental covariates are split into knots, and data in each knot section are fit independently, furthermore adding functions of knots to predict the link function. GAMs are frequently used in species distribution models [38]. The data were fit with individual id strata and case-control utilizing the Cox Proportional Hazard function, where time events were all set to 1, and the cases (use-available) added as weights [39]. This method is shown to be comparative to a conditional logistic model and with the additive effects of smooths for the GAM. All GAM models used the gam package in program R [39].

2.6.3. Random Forests Classification

Decision trees, like Random Forest (RF), are used to create partitions or splits between the predictors, forming them into regression trees. These models are ensemble models that allow for multiple models to be fit, combining the results with the rationale that this will produce a better result than a single model. It does this through the model fitting with training data and then using testing data to estimate error and the importance of each variable. RF is one approach to classify data into decision trees by generating B different bootstrapped training sets in a technique called "bagging" (Equation (4)). Bagging is a non-parametric modeling technique that is useful for high-variance predictors. Bagging averages the observations, and can significantly lower the variance compared to traditional classification trees [40].

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum\nolimits_{b=1}^{B} \hat{f}^{*b}(x) \tag{4}$$

Models generated by bagging alone have issues with becoming correlated due to selecting the strong predictor in the topmost split for all of the trees generated. Random forest is an extension of bagging, such that trees generated by bagging by using a subset of randomly chosen p predictors ($\sqrt{p}$), to decorrelate the trees from having dominant predictors in any of the models [41].

Further, model-averaging with classification trees that have low pairwise correlations, due to this variable separation among trees, reduces model bias and improves model accuracy [42]. Out-of-bag samples are then used for accuracy and error rates then averaged for the tree prediction. The randomForest package in program R allows for proper tuning of variability in the tree, which can allow for the selection of the number of variables to be split by each node. The package also allows for the individual id of the animal to be added as strata. Random forest models used program R with package randomForest.

2.6.4. Gradient Boosted Method

Gradient Boosted Method (GBM) works similarly to RF except that it does not use a bootstrapped dataset. In GBM, trees are built using the residuals from previously grown decision trees to improve the function. Boosting works as an optimization algorithm, gradient descent method. Boosting minimizes the loss function at each step, reducing the residuals through shrinkage methods whereby irrelevant predictors are made to have minimal effect on predictions [38].

$$\hat{f}(x)=\sum_{b=1}^{B}\lambda\hat{f}^{b}(x) \tag{5}$$

The shrinkage parameter λ works out inconsistencies in the residuals even further by forming new arrangements of trees (5) [43]. These methods reduce bias and variance compared to RF by using forward stepwise selection and model averaging techniques in fitting tree sequentially in contrast to bootstrapping the data [38]. GBM models "learn" slowly, *i.e.* the regression tree grows each split. Training data are used to generate an initial decision tree. Then the residuals are fit to new trees using a shrinkage parameter repeatedly and additively to update the final model. Boosting avoids overfitting that is a limitation of other classification tree methods [41]. The GBM models used package gbm.step in program R [44]. This package lacks an argument to specify the strata of the individual animal ids, thus making all presences and absences within the entire generated set the response. Group level is a much larger level than if they could be subset by individual id or as a case-control.

2.7. Model Evaluation

In many wildlife studies, AIC (Akaike's Information Criterion) is normally used to select models and select for the best set of predictors. AIC does not evaluate the ability of models' predictive functions. Learning methods that subset the data into training and testing observations are validating the efficacy of the model. Model evaluation using k-fold cross-validation subset data into training data subsets, or folds (k), and then fit models using $K-1$ folds for the model training. Five folds were used and divided evenly the presence and background data. The step selection data was a subset in 5 folds (n = 8539). The path selection data was a subset in 5-folds (n = 881). Model evaluation metrics were used to evaluate model accuracy. This study used Area Under the Curve (AUC), Cohen's kappa (Kappa), and the True Skill Statistic (TSS). AUC is a graphic method for specificity and sensitivity, with AUC values greater than 0.5 known to perform better than completely random noise. The Kappa statistic is based on thresholds derived from a confusion matrix, looking for the maximum Kappa value between 0 and 1 to determine model efficacy. The True Skill Statistic uses sensitivity and specificity for a confusion matrix and ranges from −1 to +1, and any values 0 or less indicate random models [45]. All models (RF, GBM, GAM, and conditional logistic regression) were compared using model evaluation metrics (AUC, ROC, cor, KAPPA, TSS) achieved by running 5-fold cross-validation.

3. Results

In this study, we assessed predictive modeling approaches for a multi-level multi-scale GPS-collaring study. Our approach allowed us to determine which of the levels and scales might perform adequately for predictive mapping of the landscape. In this case, the multi-scale path selection function GBM model performed the best (AUC = 1, cor = 0.989, TSS = 1, Kappa = 1). There were comparatively good single-scale path selection function RF model results. Additionally, the single and multi-scale point selection function RF also performed similarly well. Predictive maps were generated for the machine learning outputs (RF and GMB) and habitat suitability results were scaled at equal intervals for comparison of the landscape predictions (Figure 2).

It became imperative to identify the best fitting model for the smallest scale and level that could generate predictive maps. In this case, the smallest level of

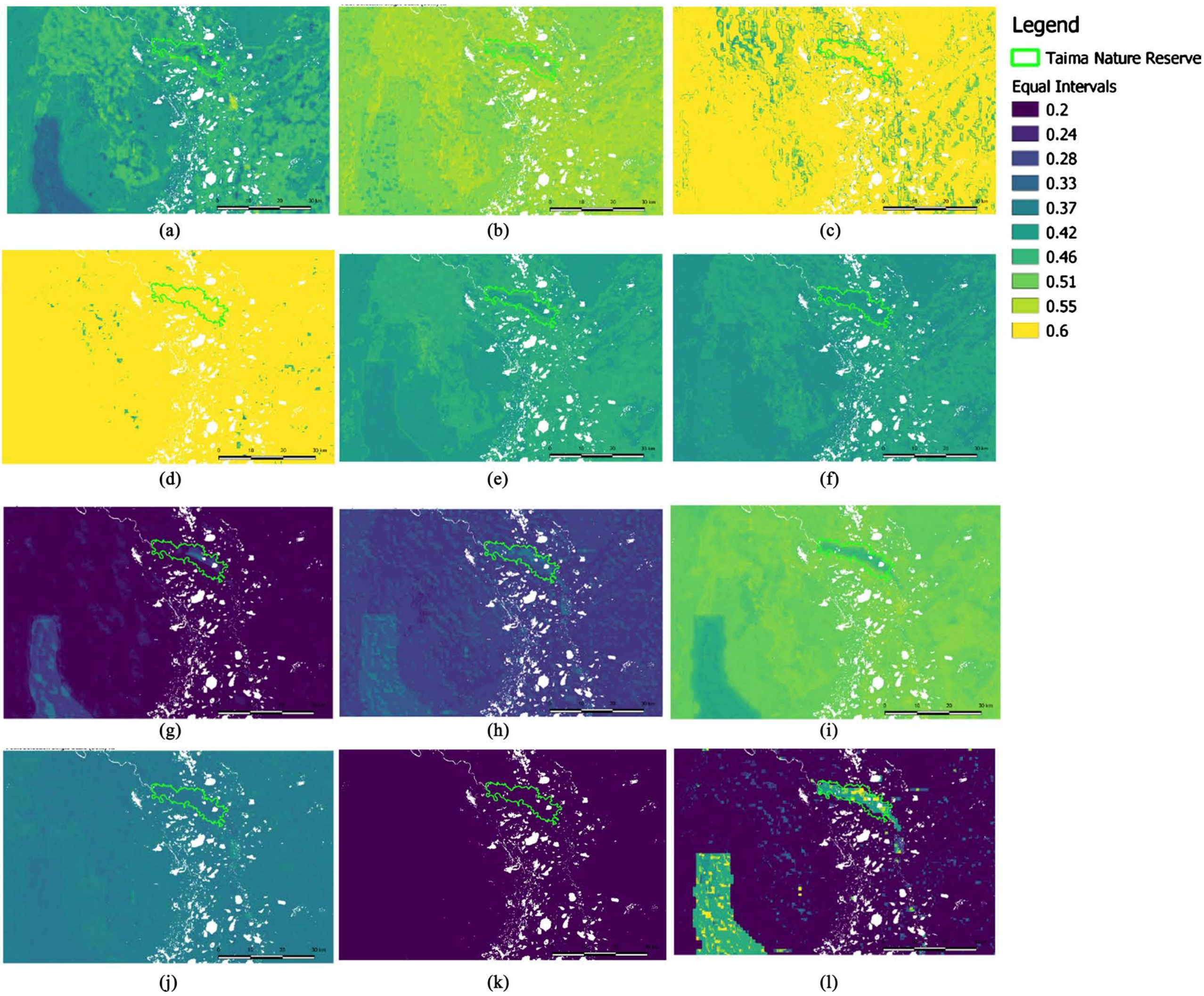

Figure 2. Predictive Resource Selection Maps for Random Forest (RF) and Gradient Boosting Method (GBM) for each of the single scale (30 m) and multi-scale outputs within each level (point selection, step selection, and path selection functions). Maps were scaled to the same 10 bins of habitat suitability equal intervals (0.2 - 0.6). (a) path selection multiscale RF; (b) path selection single scale RF; (c) path selection multiscale GMB; (d) path selection multiscale GBM; (e) step selection multiscale RF; (f) step selection single scale RF; (g) step selection multiscale GBM; (h) step selection single scale GBM; (i) point selection multiscale RF; (j) point selection single scale RF; (k) point selection multiscale GBM; (l) point selection single scale GBM.

data, single-scale and multi-scale SSFs, did not perform adequately on any models. The PSF case-control models were the next smallest temporal level. However, the model did not evaluate well, nor fit the data. The multi-scale PSF, with an individual strata level, did fit the data well and can be considered the best fitting smallest scale model. Through examining the results, it became clear that the largest level of models fit the data better, although were also the most broadly applied and less specific to the data themselves, so we thought to search for the most specific model to the data as possible.

For the largest level model, the "group" level, the gradient boosted tree (GBM), gave the highest model accuracy for the PSF. The level of this analysis is comparatively larger than other models. The GBM model only slightly outperformed the RF algorithm; however, the RF performs at a slightly more specific level of the individual id strata. We assume that larger level RSFs will likely evaluate as being better models because of the broader level of analysis. Models that are more specific, *i.e.* specifying strata (individual id, case-control), are more specific to pairing the data generated for each path or step not traversed. Additionally, when we consider the point selection function as having no case-controls or anything to pair due to the inherent randomness of background point generation within the MCP, pairing data points is non-consequential and thus at a higher level than the steps or paths. Usually, point selection outperforms step and path selection; although step and path are more specific to an animals' hourly or daily movements. Therefore, taking methods at different temporal and levels of analysis can mean that the researcher should attempt to classify the best fitting and smallest levels of analysis as possible during model evaluation to generate the most accurate predictions.

For the explanatory models, the conditional logistic regression only performed on the 30 m^2 raster data point selection function (AUC = 0.821, cor = 0.54, TSS = 0.50, and Kappa = 0.50), and GAM performed on the multi-scale point selection function (AUC = 0.90, cor = 0.654, TSS = 0.657, Kappa = 0.656). These results can be interpreted as better than random. All other models were unable to fit GAM or conditional logistic regression. In this case, the case-control framework did not improve the analysis, as point selection functions generate background points randomly without any temporal or spatial "pairing" in the data, rendering the case-control functionality completely random although the level of this analysis could be considered "individual strata" within the home range MCP that the background points were generated. GAM operated at the level of individual id, and was able to outperform the conditional logistic regression for single and multi-grain point selection.

Model Results

Single Grain (30 m^2) Point Selection Function

Using the smallest available grain for all layers, the point selection function showed the RF model performed very well, followed by GBM (**Table 4(a)**). In this case, the point "temporal level", individual ID strata "level" was the most

Table 4. Model evaluation for the area under the curve (AUC), correlation, true skills statistic (TSS) and Cohen's Kappa. (a) Point selection with 30 m^2 scale rasters; (b) Step selection with 30 m^2 scale rasters; (c) Path selection with 30 m^2 scale rasters; (d) Point selection multi-scale rasters; (e) Step selection multi-scale rasters; (f) Path selection with multi-scale rasters.

(a)

Model	AUC	cor	TSS	Kappa
RF	1	0.989337	1	1
GMB	0.988558	0.903022	0.896347	0.895994
GAM	0.879735	0.625121	0.605795	0.605548
Conditional Logistic	0.821929	0.540328	0.507053	0.50537

(b)

Model	AUC	cor	TSS	Kappa
RF	0.739253	0.46462	0.393011	0.393657
GBM	0.703558	0.372691	0.305952	0.306359
GAM	0.558275	0.102469	0.085603	0.085621
Conditional Logistic	0.549032	0.081891	0.076443	0.076415

(c)

Model	AUC	cor	TSS	Kappa
RF	1	0.979821	1	1
GBM	0.978913	0.8614446	0.8516936	0.8516711
GAM	0.746084	0.416801	0.383126	0.383275
Conditional Logistic	0.630573	0.216279	0.209406	0.209485

(d)

Model	AUC	cor	TSS	Kappa
RF	1	0.989594	1	1
GBM	0.993498	0.927041	0.921899	0.9216746
GAM	0.902305	0.654069	0.657657	0.6568519
Conditional Logistic	0.746150	0.248888	0.000709	0.0006918

(e)

Model	AUC	cor	TSS	Kappa
RF	0.739331	0.464646	0.394652	0.395227
GBM	0.712909	0.391378	0.319602	0.320075
GAM	0.559197	0.104748	0.081383	0.081383
Conditional Logistic	0.5028	0.000508	0	0

(f)

Model	AUC	cor	TSS	Kappa
GBM	1	0.997403	1	1
RF	1	0.98323	1	1
Conditional Logistic	0.630938	0.130467	0.187526	0.187623
GAM	0.842983	−0.59011	−0.1051	−0.1048

successful model out of all of the various levels (step or path) at this 30 m^2 raster scale. This scale and level also had the best performing conditional logistic regression models.

Single Grain (30 m^2) Step Selection Function

The best model for the 30 m^2 grain step selection function was RF, followed by GBM (Table 4(b)). The regression tree models (GBM, RF) performed much better than random. All of the GAM models showed a slightly better than random score for AUC. The conditional logistic regression models both fit no better than random and can be considered to have not fit the data sufficiently.

Single Grain (30 m^2) Path Selection Function

The results from the single grain path selection function showed better model evaluation metrics for all models than step selection (Table 4(c)). The RF performed the best out of all other models, and GAM and conditional logistic models path selection models showed significant improvement when compared with the step selection.

Multi-Scale Point Selection

The multi-scale data improved all model results, improving all of the model estimates when compared with the 30 m^2 grain (Table 4(d)). The RF and GBM models were overall the best fitting models. This scale and level also showed the best GAM model for all models where adding a univariate model fitting approach improved the model estimates.

Multi-Scale Step Selection

The multi-scaled step selection function produced similar results to those found in the 30 m^2, with negligible increases in the AUC, cor, Kappa, and TSS (Table 4(e)). From this we can understand that using the smallest grain possible or resampled to 30 m^2 performed similarly to that of using multiple scales.

Multi-Scale Path Selection

The multi-grain path selection GBM was the best fit model and was fit with 84 multi-scale environmental predictor variables, making only slight improvements over the 30 m^2 and multi-scale point selection RF models (Table 4(f)).

4. Discussion

This research explored methods for modeling the predictive habitat suitability and jaguar resource selection in the area surrounding the Taiama NR in the Brazilian Pantanal. The applicability of parametric models like conditional logistic regression, and non-parametric models that use machine learning (ML) methods, were applied GPS collaring data for various movement sampling units, study designs, scales, and grains available for habitat selection mapping. This study was able to apply multi-level multi-scale modeling similar to other studies[5] [6] [7], with results indicating that temporal and model levels were able to influence the interpretation of the models thus qualifying multi-level, multi-scale resource selection studies as producing better models.

This study also assessed the applicability of ML methods operating at RSF orders such as individual strata or group level, not strictly within a case-control framework like conditional logistic regression or GAM. Similar to other studies that have sought to compare ML methods to conditional logistic regression [14] [19], this study also found that machine learning methods perform better in general. However, the caveat being that the conditional logistic regression provides interpretable model outputs that enable ecologists to determine exact relationships between the study species and the environmental covariates, whereas machine learning methods provide better predictive models for the landscape that have non-interpretable environmental relationships.

Previous studies have demonstrated the importance of choice of raster grain for producing resistance surfaces that are then used with movement simulations such as least cost paths [46]. Furthermore, multi-level multi-scale models and resistance surfaces have also been used for connectivity estimates [7]. This study revealed that the increase in multiple-scales only had a improvement for some models, similar to other studies that have shown for some organisms that a multi-scale approach has no improvement over single scale [16], which here is demonstrated that various models do not necessarily perform better with multi-scale inputs.

These results are generally consistent with other findings that random forest and other machine learning algorithms perform "better" than logistic or conditional logistic regression. In this case, the conditional logistic regression fit only slightly better than random except for the point selection function at a single 30 m^2 grain. This study demonstrates a direct comparison between a semi-parametric GAM and conditional logistic regression using case-control data. The GAM improved model estimates compared to conditional logistic regression for all level models, and on the right dataset for these techniques a large improvement in model fit could be shown to improve model evaluation. Advancements in machine learning may include developing specific tools to accommodate case-controls, where specific random forest algorithms developed with conditional logistic regression could be developed specifically for this purpose. Models would have to all be fit using a case-control framework to be directly comparable at one level. Otherwise, levels for individual strata, and group level can be seen as having this overarching discrepancy.

This multi-level approach allowed us to further understand the effects of level on predictive modeling approaches. From these results, it may be assumed that larger temporal and model levels fit these particular data better, and that looking at smaller levels like case-control for the most specific step level was not able to perform adequate enough to use the results. It was our objective to identify the levels and scale to suit the analysis and the results indicated that larger scale models fit the data best, although are not as specific to movements of the organism.

5. Limitations of the Study

Due to the computational package functionality of each statistical model to operate at the levels for the entire group, or strata for individual id, or work within a case-control framework, this difference in model level potential within the analysis became a subject of concern. For example, GBM operates at the group level, and random forest at the individual id strata, GAM and conditional logistic at the individual id strata and case-control. Furthermore, this study explores the inherent limitations of current machine learning packages for working within one level of analysis, suggesting a supervised learning approach that can either be specific to one level of RSF for direct comparison, or comparison at higher order levels if necessary.

The machine learning derived predictive maps and models used in this study are not meant for direct interpretation of variable importance such as giving direct estimates of preferred canopy cover or distance to water. In predictive modeling, the goal is to accurately predict and project something new and optimize accuracy of making predictions, in contrast to understanding why these models predicted in the way they do [13]. The Gini Index provided very different variable importance plots between RF and GBM. Differences are likely based in the methods the algorithms use to make the trees, as well as the level of the RSF inherent within the tree building process. In this case, model interpretation becomes less important, as the real interest is in generating accurate predictions on new sets of data, which is one major benefits of choosing machine learning algorithms over statistical data models. In the process of comparing conditional logistic regression, an explanatory model, with predictive models such as GAM, RF, and GBM, we are trying to use both tools to understand how the jaguars respond to the landscape, and also predict onto the wider landscape. Utilizing tools from data modeling and machine learning algorithms may be the best way to bridge gaps in methodological development between the two (explanatory and predictive) polarized types of modeling framework [47]. Here we attempt both, and see how we can gain information about jaguar distribution for predicting a larger landscape region.

6. Conclusion

This study analyzed non-parametric, semi-parametric, and parametric methods for multiple temporal levels (point, step, and path selection), model levels (group, individual, case-control), and raster grains to compare the applicability of predictive statistical methods in comparison to explanatory methods, such as conditional logistic regression for predicting large areas of the landscape. We compared the parametric statistical approach using conditional logistic regression, with non-parametric and semi-parametric models that accounted for non-linearities such as generalized additive models and classification trees (RF and GBM), comparing the results using model selection methods (AUC, cor, KAPPA, TSS) derived from k-fold cross validation. The results revealed differences in predicting landscape resource selection using non-linear modeling approaches. This case study illustrates inability for a direct comparison of machine learning and explanatory modeling approaches for step and path selection functions due to the inherent data levels (case-control, individual stratum, group) that the models will accept.

Author Contributions

E.B. conceived of the idea. E.B. developed the analytical methods and performed the computations, wrote the manuscript. D.H. and T.H. supervised the final draft and finding of the work. D.H. suggested major revisions to conceptual structures and language. B.A. discussed the results and contributed comments to the manuscript.

Acknowledgements

Special thanks to the field team at Centro Nacional de Pesquisa e Conservação de Mamíferos Carnívoros, Instituto Chico Mendes de Conservação da Biodiversidade, Atibaia, SP, Brazil who were involved in the fieldwork and project discussion for this project. Espcially Dr. Morato and Dr. Santos who made extensive comments throughout manuscript preparation. Special thanks to the GIS preparation team that helped assemble remote sensing data layers for the analysis. Special thanks to Kathy Zeller for making initial comments.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cushman, S.A. and Huettmann, F. (2010) Spatial Complexity, Informatics, and Wildlife Conservation. Springer Japan, Tokyo. https://doi.org/10.1007/978-4-431-87771-4

[2] Levin, S.A. (1992) The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, **73**, 1943-1967. https://doi.org/10.2307/1941447

[3] Turner, M. (1989) Landscape Ecology: The Effect of Pattern on Process. *Annual Review of Ecology and Systematics*, **20**, 171-197. https://doi.org/10.1146/annurev.es.20.110189.001131

[4] Johnson, D.H. (1980) The Comparison of Usage and Availability Measurements for Evaluating Resource Preference. *Ecology*, **61**, 65-71. https://doi.org/10.2307/1937156

[5] DeCesare, N.J., Hebblewhite, M., Schmiegelow, F., Hervieux, D., McDermid, G.J., Neufeld, L., *et al.* (2012) Transcending Scale Dependence in Identifying Habitat with Resource Selection Functions. *Ecological Applications*, **22**, 1068-1083. https://doi.org/10.1890/11-1610.1

[6] Bauder, J.M., Breininger, D.R., Bolt, M.R., Legare, M.L., Jenkins, C.L., Rothermel, B.B., *et al.* (2018) Multi-Level, Multi-Scale Habitat Selection by a Wide-Ranging,

Federally Threatened Snake. *Landscape Ecology*, **33**, 743-763. https://doi.org/10.1007/s10980-018-0631-2

[7] Zeller, K.A., Vickers, T.W., Ernest, H.B. and Boyce, W.M. (2017) Multi-Level, Multi-Scale Resource Selection Functions and Resistance Surfaces for Conservation Planning: Pumas as a Case Study. *PLoS ONE*, **12**, e0179570. https://doi.org/10.1371/journal.pone.0179570

[8] Johnson, D.S., Thomas, D.L., Ver Hoef, J.M. and Christ, A. (2008) A General Framework for the Analysis of Animal Resource Selection from Telemetry Data. *Biometrics*, **64**, 968-976. https://doi.org/10.1111/j.1541-0420.2007.00943.x

[9] Zeller, K.A., McGarigal, K., Cushman, S.A., Beier, P., Vickers, T.W. and Boyce, W.M. (2016) Using Step and Path Selection Functions for Estimating Resistance to Movement: Pumas as a Case Study. *Landscape Ecology*, **31**, 1319-1335. https://doi.org/10.1007/s10980-015-0301-6

[10] Elliot, N.B., Cushman, S.A., Macdonald, D.W. and Loveridge, A.J. (2014) The Devil Is in the Dispersers: Predictions of Landscape Connectivity Change with Demography. *Journal of Applied Ecology*, **51**, 1169-1178. https://doi.org/10.1111/1365-2664.12282

[11] Krishnamurthy, R., Cushman, S.A., Sarkar, M.S., Malviya, M., Naveen, M., Johnson, J.A., *et al.* (2016) Multi-Scale Prediction of Landscape Resistance for Tiger Dispersal in Central India. *Landscape Ecology*, **31**, 1355-1368. https://doi.org/10.1007/s10980-016-0363-0

[12] Drew, C.A., Wiersma, Y.F. and Huettmann, F. (2011) Predictive Species and Habitat Modeling in Landscape Ecology. Springer, New York. https://doi.org/10.1007/978-1-4419-7390-0

[13] Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3

[14] Cushman, S.A. and Wasserman, T.N. (2018) Landscape Applications of Machine Learning: Comparing Random Forests and Logistic Regression in Multi-Scale Optimized Predictive Modeling of American Marten Occurrence in Northern Idaho, USA. In: Humphries, G., Magness, D.R. and Huettmann, F., Eds., *Machine Learning for Ecology and Sustainable Natural Resource Management*, Springer International Publishing, Cham, 185-203. https://doi.org/10.1007/978-3-319-96978-7_9

[15] Mi, C., Huettmann, F., Guo, Y., Han, X. and Wen, L. (2017) Why Choose Random Forest to Predict Rare Species Distribution with Few Samples in Large Undersampled Areas? Three Asian Crane Species Models Provide Supporting Evidence. *PeerJ*, **5**, e2849. https://doi.org/10.7717/peerj.2849

[16] Martin, A.E. and Fahrig, L. (2012) Measuring and Selecting Scales of Effect for Landscape Predictors in Species-Habitat Models. *Ecological Applications*, **22**, 2277-2292. https://doi.org/10.1890/11-2224.1

[17] Shoemaker, K.T., Heffelfinger, L.J., Jackson, N.J., Blum, M.E., Wasley, T. and Stewart, K.M. (2018) A Machine-Learning Approach for Extending Classical Wildlife Resource Selection Analyses. *Ecology and Evolution*, **8**, 3556-3569. https://doi.org/10.1002/ece3.3936

[18] Frakes, R.A., Belden, R.C., Wood, B.E. and James, F.E. (2015) Landscape Analysis of Adult Florida Panther Habitat. *PLoS ONE*, **10**, e0133044. https://doi.org/10.1371/journal.pone.0133044

[19] Zeller, K.A., Jennings, M.K., Vickers, T.W., Ernest, H.B., Cushman, S.A. and Boyce, W.M. (2018) Are All Data Types and Connectivity Models Created Equal? Validating Common Connectivity Approaches with Dispersal Data. *Diversity and Distri-*

butions, **24**, 868-879. https://doi.org/10.1111/ddi.12742

[20] Duhart, C., Dublon, G., Mayton, B., Davenport, G. and Paradiso, J.A. (2019) Deep Learning for Wildlife Conservation and Restoration Efforts. 36*th International Conference on Machine Learning*, Long Beach, 5.

[21] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017) Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sensing of Environment*, **202**, 18-27. https://doi.org/10.1016/j.rse.2017.06.031

[22] Miller, H.J., Dodge, S., Miller, J. and Bohrer, G. (2019) Towards an Integrated Science of Movement: Converging Research on Animal Movement Ecology and Human Mobility Science. *International Journal of Geographical Information Science*, **33**, 855-876. https://doi.org/10.1080/13658816.2018.1564317

[23] Wearn, O.R., Freeman, R. and Jacoby, D.M.P. (2019) Responsible AI for Conservation. *Nature Machine Intelligence*, **1**, 72-73. https://doi.org/10.1038/s42256-019-0022-7

[24] Evans, T.L., Costa, M., Tomas, W.M. and Camilo, A.R. (2014) Large-Scale Habitat Mapping of the Brazilian Pantanal Wetland: A Synthetic Aperture Radar Approach. *Remote Sensing of Environment*, **155**, 89-108. https://doi.org/10.1016/j.rse.2013.08.051

[25] McGarigal, K., Wan, H.Y., Zeller, K.A., Timm, B.C. and Cushman, S.A. (2016) Multi-Scale Habitat Selection Modeling: A Review and Outlook. *Landscape Ecology*, **31**, 1161-1175. https://doi.org/10.1007/s10980-016-0374-x

[26] Galpern, P. and Manseau, M. (2013) Finding the Functional Grain: Comparing Methods for Scaling Resistance Surfaces. *Landscape Ecology*, **28**, 1269-1281. https://doi.org/10.1007/s10980-013-9873-1

[27] McGarigal, K., Zeller, K.A. and Cushman, S.A. (2016) Multi-Scale Habitat Selection Modeling: Introduction to the Special Issue. *Landscape Ecology*, **31**, 1157-1160. https://doi.org/10.1007/s10980-016-0388-4

[28] Morato, R.G., Connette, G.M., Stabach, J.A., De Paula, R.C., Ferraz, K.M.P.M., Kantek, D.L.Z., *et al.* (2018) Resource Selection in an Apex Predator and Variation in Response to Local Landscape Characteristics. *Biological Conservation*, **228**, 233-240. https://doi.org/10.1016/j.biocon.2018.10.022

[29] Jędrzejewski, W., Robinson, H.S., Abarca, M., Zeller, K.A., Velasquez, G., Paemelaere, E.A.D., *et al.* (2018) Estimating Large Carnivore Populations at Global Scale Based on Spatial Predictions of Density and Distribution—Application to the Jaguar (*Panthera onca*). *PLoS ONE*, **13**, e0194719. https://doi.org/10.1371/journal.pone.0194719

[30] Morato, R.G., Stabach, J.A., Fleming, C.H., Calabrese, J.M., De Paula, R.C., Ferraz, K.M.P.M., *et al.* (2016) Space Use and Movement of a Neotropical Top Predator: The Endangered Jaguar. *PLoS ONE*, **11**, e0168176. https://doi.org/10.1371/journal.pone.0168176

[31] Cullen Junior, L., Sana, D.A., Lima, F., Abreu, K.C. de and Uezu, A. (2013) Selection of Habitat by the Jaguar, *Panthera onca* (Carnivora: Felidae), in the Upper Paraná River, Brazil. *Zoologia* (*Curitiba*), **30**, 379-387. https://doi.org/10.1590/S1984-46702013000400003

[32] de la Torre, J.A., Núñez, J.M. and Medellín, R.A. (2017) Habitat Availability and Connectivity for Jaguars (*Panthera onca*) in the Southern Mayan Forest: Conservation Priorities for a Fragmented Landscape. *Biological Conservation*, **206**, 270-282. https://doi.org/10.1016/j.biocon.2016.11.034

[33] Boitani, L. and Fuller, T.K. (2000) Research Techniques in Animal Ecology: Controversies and Consequences. Columbia University Press, New York.

[34] Calenge, C. (2015) Analysis of Animal Movements in R: The adehabitatLT Package. 85.

[35] Hijmans, R.J. and Ghosh, A. (2019) Spatial Data Analysis with R. 135.

[36] Porto, M. and Quaglietta, L. (2019) "SiMRiv" (Version 1.0.3): An R Package for Simulation and Analysis of Spatially-Explicit Individual Multistate (Animal) Movements in Any Landscape. 15.

[37] Hooten, M.B., Hanks, E.M., Johnson, D.S. and Alldredge, M.W. (2014) Temporal Variation and Scale in Movement-Based Resource Selection Functions. *Statistical Methodology*, **17**, 82-98. https://doi.org/10.1016/j.stamet.2012.12.001

[38] Elith, J., Leathwick, J.R. and Hastie, T. (2008) A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology*, **77**, 802-813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

[39] Hastie, T. (2019) Generalized Additive Models 1.16.1. CRAN Repository.

[40] De'ath, G. (2007) Boosted Trees for Ecological Modeling and Prediction. *Ecology*, **88**, 243-251. https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2

[41] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: with Applications in R. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

[42] Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., *et al.* (2007) Random Forests for Classification in Ecology. *Ecology*, **88**, 2783-2792. https://doi.org/10.1890/07-0539.1

[43] James, G., Witten, D., Hastie, T. and Robert, T. (2017) An Introduction to Statistical Learning: With Applications in R.

[44] Ridgeway, G. (2019) Generalized Boosted Models: A Guide to the GBM Package. 15.

[45] Allouche, O., Tsoar, A. and Kadmon, R. (2006) Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS): Assessing the Accuracy of Distribution Models. *Journal of Applied Ecology*, **43**, 1223-1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

[46] Etherington, T.R. (2016) Least-Cost Modelling and Landscape Ecology: Concepts, Applications, and Opportunities. *Current Landscape Ecology Reports*, **1**, 40-53. https://doi.org/10.1007/s40823-016-0006-9

[47] Shmueli, G. (2010) To Explain or to Predict? *Statistical Science*, **25**, 289-310. https://doi.org/10.1214/10-STS330