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

Eve Bohnett (evebohnett@ufl.edu)
University of Florida https://orcid.org/0000-0002-1870-8897
Thomas Hoctor
University of Florida
David Hulse
University of Florida
Bilal Ahmad
University of Swat
Bernardo Niebuhr
Instituto Chico Mendes de Conservacao da Biodiversidade
Ronaldo Morato
Instituto Chico Mendes de Conservacao da Biodiversidade

Research

Keywords: gradient boosting method, machine learning, movement ecology, multi-level, multi-scale, path selection, predictive mapping, random forest, step selection

Posted Date: February 11th, 2020

DOI: https://doi.org/10.21203/rs.2.23193/v1

MULTI-LEVEL, MULTI-SCALE MODELING AND PREDICTIVE MAPPING FOR JAGUARS IN THE BRAZILIAN PANTANAL

BOHNETT, E.[1], HOCTOR, T.[1], HULSE, D.[1], AHMAD, B.[2], NIEBUHR, B.[3], MORATO, R.[3]

1. Department of Landscape Architecture, University of Florida, Gainesville, United States of America 2. Institute of Agriculture Sciences and Forestry, University of Swat, Swat, Pakistan. 3. Centro Nacional de Pesquisa e Conservação de Mamíferos Carnívoros, Instituto Chico Mendes de Conservação da Biodiversidade, Atibaia, SP, Brazil

*Corresponding Author: Eve Bohnett, evebohnett@ufl.edu

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

Abstract

• Background

Machine learning (ML) methods and remote sensing data were used to build multi-level multi-scale resource selection models and predictive maps onto the extended landscape for jaguars (Panthera onca) in the Brazilian Pantanal. Objectives were to compare multiple predictive modeling and exploratory modeling approaches. Included in the analysis, multi-scale raster grains (30m, 90m, 180m, 360m, 720m, 1440m), GPS collaring temporal levels (point, path, and step) and model data structural levels (group, individual, case-control).

•Methods

Multi-scale multi-level data subsets were fit with explanatory and predictive statistical methods. Conditional logistic regression, generalized additive modeling (GAM), and classification regression trees, such as random forests (RF) and gradient boosted regression tree (GBM) were compared for their utility to the study. Model evaluation, using training and testing data in a k-fold cross-validation approach, determined the AUC, Kappa, and TSS for model evaluation and comparison.

• Results

Results indicated that the multi-level, multi-scale techniques improved model outputs. Overall, larger level models and those that used multi-scale raster grains showed the best model evaluation. The highest-ranked model was the multi-scale path selection function GBM and was one of the broadest levels of data.

• Conclusions

Results indicated that multi-level, multi-scale models produced mixed results of applicability across models and levels. The identification of the appropriate temporal scale and statistical model needs careful consideration in predictive mapping efforts.

•Keywords

gradient boosting method, machine learning, movement ecology, multi-level, multi-scale, path selection, predictive mapping, random forest, step selection

Background

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. Models can respond to changes in spatial extent, spatial grain, temporal duration, temporal grain, entities measured, variables measured, and processes

linking entities and variables (1–3). 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 the resource selection modeling frameworks, one of the approaches researchers apply to integrate scale are buffers of various sizes around the data points to average environmental covariates within a given area. Therefore, integrating the effective scale of environmental predictors (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 applies to studies on various organisms by performing scale optimized multi-level modeling. These methods determine how scales and levels can perform within each level of resource selection trying to optimize the scale of the environmental covariates and any temporal levels within the data (6,7). By quantifying the patterns and processes that naturally occur at different scales in time and space, we can alter conclusions regarding the key ecological and evolutionary processes that compose landscapes.

Temporal Levels of Resource Selection for GPS Collaring

Resource selection may also depend on the temporal scale of the process of interest. Resource selection functions have been parsed out into various temporal levels (point, step, and path selection functions). For GPS data, these levels are the sampling units for classifying animal movement. Binary response variables include (1) animal movement data and (0) pseudoabsences generated on the landscape within several levels by using a “used vs. available” sampling design broadly referred to as resource selection function (RSF) (Johnson et al., 2008). Fine-scale GPS collaring trajectories allow the extension of traditional RSF’s to point, step, and path selection functions. Studies with these levels also should asses 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 (Johnson, 1980). These levels are different for step and path selection functions. Animal steps and paths are sectioned into hourly or daily sections respectively. Then, pseudoabsences are generated for steps or paths *not* traveled, yet were available for animal movement and not chosen, (Zeller et al., 2016). Steps and paths are specific to third-order habitat selection at the level of the resource patches. Temporal differences in the generation of pseudoabsences for path selection have been shown to optimize the scale of effect for large carnivore dispersal studies (10,11). Meaning, paths can be generated for 12 hours, 24 hours, or 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.

Analysis and Modeling Options

Explanatory and predictive models are both important options for analysis for movement ecology. Explanatory models obtain estimates for an organism's movement with environmental covariates. For example, conditional logistic regression is often using a paired case-control framework assigning presences (case=1) and absences (control =0). For example, researchers use conditional logistic regression for pairing step selection presences with pseudoabsences in a case-control framework. Researchers thus obtain specific estimates of an organism's relationship to the covariates. However, these explanatory modeling methods are not recommended for predictive mapping or predicting results on new data. Predictive mapping efforts that utilize machine learning, for example, are most suitable (12,13). Machine learning methods or other predictive modeling approaches are better suited for predicting the model on new data.

Emerging studies on machine learning (ML) methods for landscape ecology have used classification regression trees, showing they outperform other methods for multi-scale modeling (Cushman, 2018). At continental scales where GPS collaring data may be sparse, random forest (RF)

models performed adequately (14). 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 (15).

Previous research has also applied ML methods like RF for movement studies for point selection functions for mule deer (16) and Florida panther (17). Zeller (2018) explored non-parametric and semi-parametric statistical modeling approaches for use in resource selection functions to derive predictive distribution maps for large carnivore conservation (18). They compared RSFs generated from points, steps, and paths using conditional logistic regression, to point selection functions using machine learning methods (18). This study extends this research by applying non-parametric and semi-parametric techniques to multi-scale multi-level methods for GPS collaring data using path and step selection functions.

How predictive modeling, like machine learning methods, may apply to step and path selection functions is unknown. Step and path selection functions traditionally use conditional logistic regression. 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 (Google Earth Engine) and animal movement data (MOVEBANK) repositories are making data more available, and also processing and modeling more sophisticated (19–22).

Here we aimed at understanding how explanatory data models and predictive models can contribute to landscape estimates and predictive mapping. In this study, multi-scale, multi-level modeling of habitat selection is explored using jaguars in the Brazilian Pantanal as a case study.

Methods

Study Area

The study area is the Brazilian Pantanal region surrounding the Taiama Ecological Station. 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 (23). 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.

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, namely elevation, slope, aspect, LULC, forest non-forest, total canopy cover, roads, water and hydrology, human density, and cattle density (Table 2). Collinearity was checked using pairwise comparison for those models where this may be an issue. Machine learning models do not consider collinearity to be an issue so all data layers can be included.

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 (1hr), and path (24 hour) GPS collaring temporal levels (Table 1a) analyzed at different model levels (group, individual ID strata, case-control) (Table 1b), furthermore incorporating multi-scale raster grain data to compare single grain (30m) and multi-grain raster data (30m,90m,180m,360m,720m,1440m) (Table 1c). 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 (24).

Raster Grain

To represent the raster layers at multiple scales, then a Gaussian kernel smoother was used to average the layers at multiple extents (90m, 180m, 360m, 720m, and 1440m), resulting in a total of 84 variables. 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 (25). 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 (26). Although similar in technique to expanding distance buffers and averaging the pixels around a point or line.

The conditional regression model and GAM models are fit univariately to determine the adequate scale. 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 (Fig 2).

Study Species

Modeling methods rely on niche habitat concepts for environmental covariates to construct accurate models of jaguar distributions on the landscape (27). 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 (28) and roads (29). Jaguars prefer areas having topography with moderate slopes. They also prefer riparian areas with high amounts of water (30,31).

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 42741 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 (29).

Point Selection Functions

Minimum convex polygons have been the traditional method for home range estimation, drawing a polygon around the point locations for the animal. This technique provided a crude estimate of the home range (Fig 1a), most commonly reported with 95 percent of data points (32). Here minimum convex polygons 95% (MCP95) were determined using the MCP function in the *adehabitatHR* package in R (Callenge 2006). Random points within the MCP were generated using the *dismo* package in R (Hijmans, 2017).

Step Selection Functions

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. Pseudoabsence steps had the same original step distance projected into a different angle around the starting point (Fig 1b). Steps from hourly data had a total of 85388 presences and absences, generating one pseudoabsence

191 step for each step. Step selection was performed in program R, using the package AdeHabitatLT
192 (Calenge, 2016).

193 **Path Selection Functions**

194 Animal trajectories were subset into 24-hour time sequences of steps, generating longer paths for daily
195 intervals, with a total of 4409 present and absent paths (Fig 1c). Absences were generated using a
196 correlated random walk (CRW) method. The CRW began at the starting point, simulating a trajectory of
197 a similar length at an alternate angle. CRWs are a completely randomized simulation of blind jaguar
198 movement at any chance direction. CRW does not account for any decisions the jaguars normally
199 encounter in time or space.

200 Nonetheless, the correlated random walk provided a randomized path to understand the simplest
201 baseline from which to compare the actual jaguar movement. This study is the first to simulate path
202 movement using a correlated random walk model to simulate absences on the landscape for analysis in
203 a path framework. The mean of all steps values in the path aggregated to one single value to represent
204 each path extracted covariate. Paths were generated in program R, package SiMRiv (Porto & Quaglietta,
205 2018).

206 **Model Levels**

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

211 However, the RF model used a higher order level at the individual ID strata and was not case-control.
212 Therefore, RF performs the analysis for all steps and all pseudosteps 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(Fig 2).

Conditional Logistic Regression

Traditional resource selection functions for GPS collaring studies use an exploratory 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 (35).

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

The likelihood is maximized for resource coefficients β.

$$\prod_{t=1}^{T}[x(s_t)|\beta]_u$$

A vector of resource covariates, β is a set of regression coefficients, x, a normalizing constant, and $g(x,\beta)$ is a resource selection function (35). 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 (Elff, 2018). The mclogit package includes case-control and individual level strata for model fitting.

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 explain relationships between variables.

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

The smoothing splines, thin-plate splines, cubic splines, and splines with variable knots (k =3, k=8) applied to applicable covariates. 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 (Elith et al., 2006). The data were fit with individual id strata. Then, case-control data were fit with the a Cox Proportional Hazard function, where time events were all set to 1, and the cases (use-available) added as weights (Hastie, 2018). 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 (Hastie, 2019).

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.' 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 (36).

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

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 (James, Witten, Hastie, Tibshirani, 2013).

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 (37). 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 (Breiman, 2018).

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).

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

The shrinkage parameter λ works out inconsistencies in the residuals even further by forming new arrangements of trees (James et al. 2014). 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 (James, Witten, Hastie, Tibshirani, 2013).The GBM models used package gbm in program R (39). 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.

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. 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 (40). 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.

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.

In the goals of this work, it was 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 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 RF, with an individual strata level, did fit the data well and can be considered the best fitting smallest scale model.

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 pseudoabsence 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. In this case, the multi-scale path selection GBM performed the best, and also predicted most of the landscape as having good habitat for jaguars.

For the explanatory models, the conditional logistic regression only performed on the $30m^2$ 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 pseudo-absences 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 pseudoabsences 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.

Single Grain ($30m^2$) Point Selection Function

Using the smallest available grain available for all layers, the point selection function showed the RF model performed very well, followed by GBM (Table 3a). In this case, the point "temporal level", individual ID strata "level" was the most successful model out of all of the various levels (step or path) at this $30m^2$ raster scale. This scale and level also had the best performing conditional logistic regression models.

Single Grain ($30m^2$) Step Selection Function

The best model for the 30m^2 grain step selection function was RF, followed by GBM (Table 3b). 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 (30m^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 3c). 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 30m^2 grain (Table 3d). 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 30m^2, with negligible increases in the AUC, cor, Kappa, and TSS (Table 3e). From this we can understand that using the smallest grain possible or resampled to 30m^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 30m^2 and multi-scale point selection RF models. (Table 3f).

Results Summary

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.

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 exploratory data models, like conditional logistic regression, and predictive models, like machine learning (ML) methods, were applied GPS collaring data for various levels, scales, and grains available for habitat selection mapping. This study was able to apply multi-level multi-scale modeling similar to other studies (Bauder et al., 2018; DeCesare et al., 2012; Zeller et al., 2017b), 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.

To further this research, this study 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 (18,42), this study also found that machine learning methods perform

390 better in general. However, the caveat being that the conditional logistic regression provides
391 interpretable model outputs that enable ecologists to determine exact relationships between the study
392 species and the environmental covariates, whereas machine learning methods provide better predictive
393 models for the landscape that have non-interpretable environmental relationships.

394 Previous studies have demonstrated the importance of choice of raster grain for producing resistance
395 surfaces that are then used with movement simulations such as least cost paths (43). Furthermore,
396 multi-level multi-scale models and resistance surfaces have also been used for connectivity estimates
397 (7). This study revealed that the increase in multiple-scales only had a improvement for some models,
398 similar to other studies that have shown for some organisms that a multi-scale approach has no
399 improvement over single scale (15), which here is demonstrated that various models do not necessarily
400 perform better with multi-scale inputs.

401 These results are generally consistent with other findings that random forest and other machine
402 learning algorithms perform "better" than logistic or conditional logistic regression. In this case, the
403 conditional logistic regression fit only slightly better than random except for the point selection function
404 at a single $30m^2$ grain. This study demonstrates a direct comparison between a semi-parametric GAM
405 and conditional logistic regression using case-control data. The GAM improved model estimates
406 compared to conditional logistic regression for all level models, and on the right dataset for these
407 techniques a large improvement in model fit could be shown to improve model evaluation.
408 Advancements in machine learning may include developing specific tools to accommodate case-
409 controls, where specific random forest algorithms developed with conditional logistic regression could
410 be developed specifically for this purpose. Models would have to all be fit using a case-control
411 framework to be directly comparable at one level. Otherwise, levels for individual strata, and group level
412 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.

Limitations of the Study

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 (exploratory and predictive) polarized types of modeling framework (44). Here we attempt both, and see how we can gain information about jaguar distribution for predicting a larger landscape region.

Conclusion

436 This study analyzed non-parametric, semi-parametric, and parametric methods for a multiple temporal
437 levels (point, step, and path selection), model levels (group, individual, case-control), and raster grains
438 to compare the applicability of predictive statistical methods in comparison to explanatory methods
439 such as conditional logistic regression for predicting large areas of the landscape. We compared the
440 parametric statistical approach using conditional logistic regression, with non-parametric and semi-
441 parametric models that accounted for non-linearities such as generalized additive models and
442 classification trees (RF and GBM), comparing the results using model selection methods (AUC, cor,
443 KAPPA, TSS) derived from k-fold cross validation. The results revealed differences in predicting
444 landscape resource selection using non-linear modeling approaches.

445

446 References

447 1. Cushman SA, Huettmann F, editors. Spatial Complexity, Informatics, and Wildlife Conservation
448 [Internet]. Tokyo: Springer Japan; 2010 [cited 2018 Mar 13]. Available from:
449 http://link.springer.com/10.1007/978-4-431-87771-4

450 2. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture.
451 Ecology. 1992 Dec;73(6):1943–67.

452 3. Turner M. Landscape Ecology: The Effect Of Pattern On Process. Annu Rev Ecol Syst. 1989 Jan
453 1;20(1):171–97.

454 4. Johnson DH. The Comparison of Usage and Availability Measurements for Evaluating Resource
455 Preference. Ecology. 1980 Feb;61(1):65–71.

456 5. DeCesare NJ, Hebblewhite M, Schmiegelow F, Hervieux D, McDermid GJ, Neufeld L, et al.
457 Transcending scale dependence in identifying habitat with resource selection functions. Ecol Appl.
458 2012 Jun;22(4):1068–83.

459 6. Bauder JM, Breininger DR, Bolt MR, Legare ML, Jenkins CL, Rothermel BB, et al. Multi-level, multi-
460 scale habitat selection by a wide-ranging, federally threatened snake. Landsc Ecol. 2018
461 May;33(5):743–63.

462 7. Zeller KA, Vickers TW, Ernest HB, Boyce WM. Multi-level, multi-scale resource selection functions
463 and resistance surfaces for conservation planning: Pumas as a case study. Boyce MS, editor. PLOS
464 ONE. 2017 Jun 13;12(6):e0179570.

8. Johnson DS, Thomas DL, Ver Hoef JM, Christ A. A General Framework for the Analysis of Animal Resource Selection from Telemetry Data. Biometrics. 2008 Sep;64(3):968–76.

9. Zeller KA, McGarigal K, Cushman SA, Beier P, Vickers TW, Boyce WM. Using step and path selection functions for estimating resistance to movement: pumas as a case study. Landsc Ecol. 2016 Aug;31(6):1319–35.

10. Elliot NB, Cushman SA, Macdonald DW, Loveridge AJ. The devil is in the dispersers: predictions of landscape connectivity change with demography. Pettorelli N, editor. J Appl Ecol. 2014 Oct;51(5):1169–78.

11. Krishnamurthy R, Cushman SA, Sarkar MS, Malviya M, Naveen M, Johnson JA, et al. Multi-scale prediction of landscape resistance for tiger dispersal in central India. Landsc Ecol. 2016 Aug;31(6):1355–68.

12. Drew CA, Wiersma YF, Huettmann F, editors. Predictive Species and Habitat Modeling in Landscape Ecology [Internet]. New York, NY: Springer New York; 2011 [cited 2019 Mar 31]. Available from: http://link.springer.com/10.1007/978-1-4419-7390-0

13. Kuhn M, Johnson K. Applied Predictive Modeling [Internet]. New York, NY: Springer New York; 2013 [cited 2019 Apr 23]. Available from: http://link.springer.com/10.1007/978-1-4614-6849-3

14. Mi C, Huettmann F, Guo Y, Han X, Wen L. 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. 2017 Jan 12;5:e2849.

15. Martin AE, Fahrig L. Measuring and selecting scales of effect for landscape predictors in species–habitat models. Ecol Appl. 2012 Dec;22(8):2277–92.

16. Shoemaker KT, Heffelfinger LJ, Jackson NJ, Blum ME, Wasley T, Stewart KM. A machine-learning approach for extending classical wildlife resource selection analyses. Ecol Evol. 2018 Mar;8(6):3556–69.

17. Frakes RA, Belden RC, Wood BE, James FE. Landscape Analysis of Adult Florida Panther Habitat. Mortelliti A, editor. PLOS ONE. 2015 Jul 29;10(7):e0133044.

18. Zeller KA, Jennings MK, Vickers TW, Ernest HB, Cushman SA, Boyce WM. Are all data types and connectivity models created equal? Validating common connectivity approaches with dispersal data. Bolliger J, editor. Divers Distrib. 2018 Jul;24(7):868–79.

19. Duhart C, Dublon G, Mayton B, Davenport G, Paradiso JA. Deep Learning for Wildlife Conservation and Restoration Efforts. :5.

20. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017 Dec;202:18–27.

21. Miller HJ, Dodge S, Miller J, Bohrer G. Towards an integrated science of movement: converging research on animal movement ecology and human mobility science. Int J Geogr Inf Sci. 2019 May 4;33(5):855–76.

22. Wearn OR, Freeman R, Jacoby DMP. Responsible AI for conservation. Nat Mach Intell. 2019 Feb;1(2):72–3.

23. Evans TL, Costa M, Tomas WM, Camilo AR. Large-scale habitat mapping of the Brazilian Pantanal wetland: A synthetic aperture radar approach. Remote Sens Environ. 2014 Dec;155:89–108.

24. McGarigal K, Wan HY, Zeller KA, Timm BC, Cushman SA. Multi-scale habitat selection modeling: a review and outlook. Landsc Ecol. 2016 Aug;31(6):1161–75.

25. Galpern P, Manseau M. Finding the functional grain: comparing methods for scaling resistance surfaces. Landsc Ecol. 2013 Aug;28(7):1269–81.

26. McGarigal K, Zeller KA, Cushman SA. Multi-scale habitat selection modeling: introduction to the special issue. Landsc Ecol. 2016 Aug;31(6):1157–60.

27. Morato RG, Connette GM, Stabach JA, De Paula RC, Ferraz KMPM, Kantek DLZ, et al. Resource selection in an apex predator and variation in response to local landscape characteristics. Biol Conserv. 2018 Dec;228:233–40.

28. Jędrzejewski W, Robinson HS, Abarca M, Zeller KA, Velasquez G, Paemelaere EAD, et al. Estimating large carnivore populations at global scale based on spatial predictions of density and distribution – Application to the jaguar (Panthera onca). Hagen CA, editor. PLOS ONE. 2018 Mar 26;13(3):e0194719.

29. Morato RG, Stabach JA, Fleming CH, Calabrese JM, De Paula RC, Ferraz KMPM, et al. Space Use and Movement of a Neotropical Top Predator: The Endangered Jaguar. Stöck M, editor. PLOS ONE. 2016 Dec 28;11(12):e0168176.

30. Cullen Junior L, Sana DA, Lima F, Abreu KC de, Uezu A. Selection of habitat by the jaguar, Panthera onca (Carnivora: Felidae), in the upper Paraná River, Brazil. Zool Curitiba. 2013 Aug;30(4):379–87.

31. de la Torre JA, Núñez JM, Medellín RA. Habitat availability and connectivity for jaguars (Panthera onca) in the Southern Mayan Forest: Conservation priorities for a fragmented landscape. Biol Conserv. 2017 Feb;206:270–82.

32. Boitani L, Fuller TK, editors. Research techniques in animal ecology: controversies and consequences. New York: Columbia University Press; 2000. 442 p. (Methods and cases in conservation science).

33. Calenge C. Analysis of Animal Movements in R: the adehabitatLT Package. :85.

34. Porto M, Quaglietta L. 'SiMRiv' (version 1.0.3) An R package for simulation and analysis of spatially-explicit individual multistate (animal) movements in any landscape. :15.

35. Hooten MB, Hanks EM, Johnson DS, Alldredge MW. Temporal variation and scale in movement-based resource selection functions. Stat Methodol. 2014 Mar;17:82–98.

36. De'ath G. BOOSTED TREES FOR ECOLOGICAL MODELING AND PREDICTION. Ecology. 2007 Jan;88(1):243–51.

37. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. Ecology. 2007 Nov;88(11):2783–92.

38. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008 Jul;77(4):802–13.

39. Ridgeway G. Generalized Boosted Models: A guide to the gbm package. :15.

40. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. J Appl Ecol. 2006 Sep 12;43(6):1223–32.

41. Zeller KA, Vickers TW, Ernest HB, Boyce WM. Multi-level, multi-scale resource selection functions and resistance surfaces for conservation planning: Pumas as a case study. Boyce MS, editor. PLOS ONE. 2017 Jun 13;12(6):e0179570.

42. Cushman SA, Wasserman TN. 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 DR, Huettmann F, editors. Machine Learning for Ecology and Sustainable Natural Resource Management [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 Mar 31]. p. 185–203. Available from: http://link.springer.com/10.1007/978-3-319-96978-7_9

43. Etherington TR. Least-Cost Modelling and Landscape Ecology: Concepts, Applications, and Opportunities. Curr Landsc Ecol Rep. 2016 Mar;1(1):40–53.

44. Shmueli G. To Explain or to Predict? Stat Sci. 2010 Aug;25(3):289–310.

Abbreviations

Machine Learning (ML)
Random Forest (RF)
Gradient Boosting Method (GBM)
Generalized Additive Models (GAM)
Resource Selection Function (RSF)
Path Selection Function (PSF)
Step Selection Function (SSF)
Correlated random walk (CRW)
Area Under the Curve (AUC)
Cohen's kappa (Kappa)
True Skill Statistic (TSS)
Aikake's Information Criterion (AIC)

Declarations

- Ethical approval and consent to participate

-Not applicable

- Consent for publication

-Not applicable

- Availability of data and materials

-Jaguar GPS collaring data is freely available through MOVEBANK. All remote sensing data layers were compiled from original sources using internet downloads.

- Competing interests

-The authors declare that they have no competing interests

- Funding

-Not applicable

- Authors' contributions

-EB prepared data analyses and wrote up the initial draft. RM provided GPS collaring and remote sensing data. All co-authors read the manuscript and provided detailed comments for improvement.

- 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 for this project. Special thanks to the GIS preparation team that helped assemble remote sensing data layers for the analysis. Also, special thanks to Kathy Zeller for making initial comments.

- Authors' information (optional)

-Eve Bohnett is currently a PhD Candidate at University of Florida. She works currently in the Center for Landscape Conservation Planning and the Florida Institute for Built Environmental Resilience. Her research focuses mainly on statistical ecology and predictive species distribution modeling.

Figures

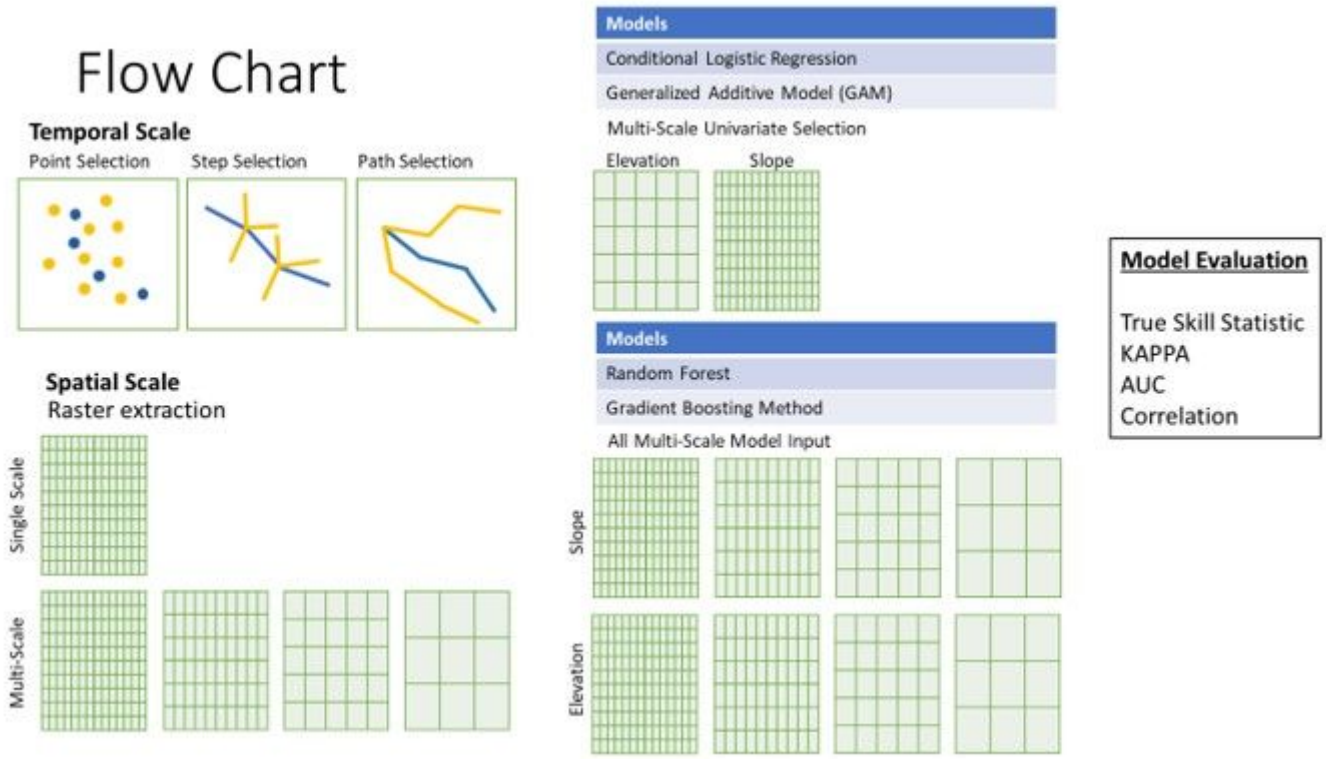

Figure 1

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.

Figure 2

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

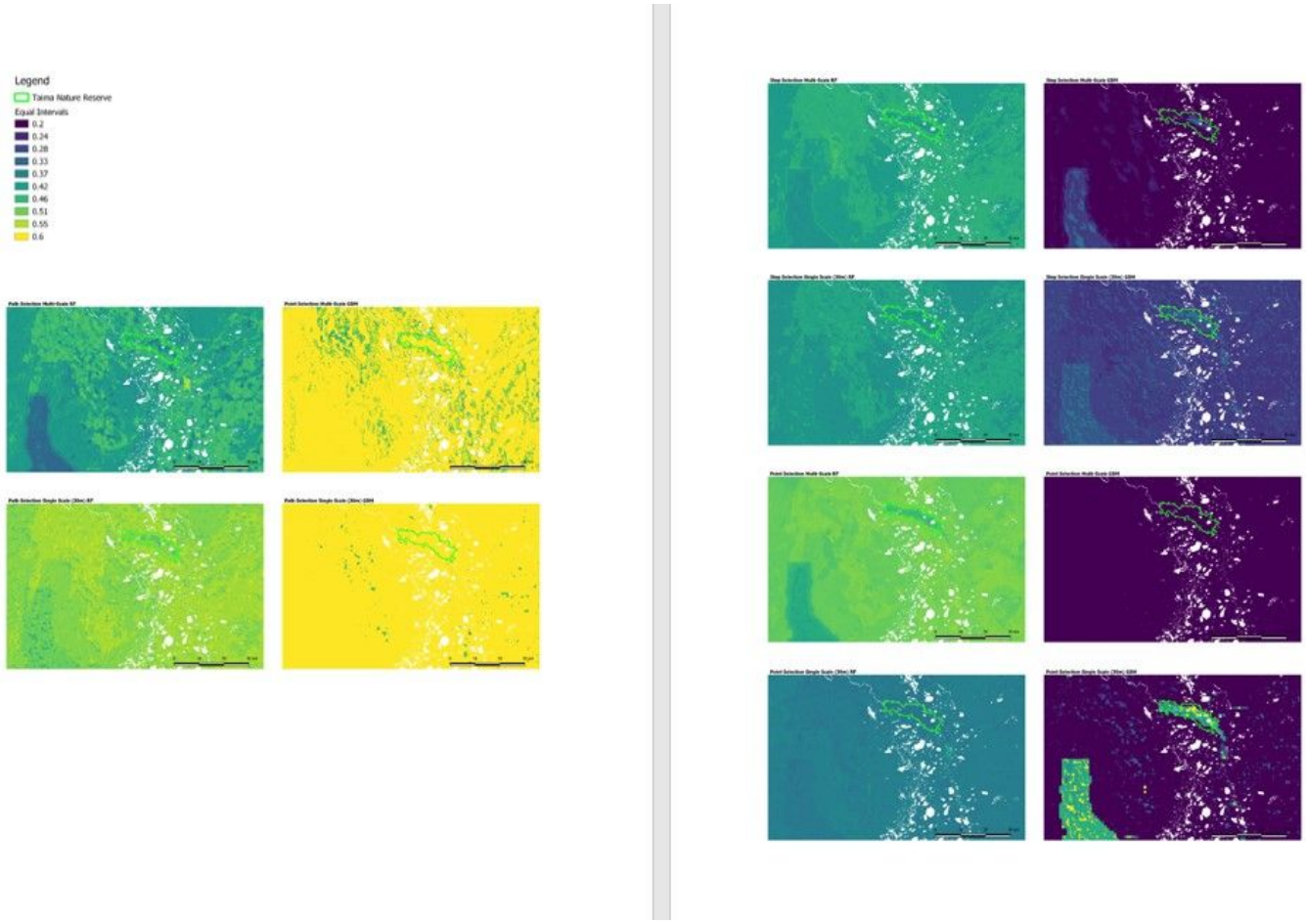

Figure 3

Predictive Resource Selection Maps for Random Forest (RF) and Gradient Boosting Method (GBM) for each of the single scale (30m) and multi-scale outputs within each level (point selection, step selection, and path selection functions). Maps were scaled to the same 10 bins of equal intervals (0.2-0.6)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryInformation.docx