Research article

Model thresholds are more important than presence location type: Understanding the distribution of lowland tapir (*Tapirus terrestris*) in a continuous Atlantic forest of southeast Brazil

Darren Norris[1,2]

[1] Programa de Pos-Graduacao em Biodiversidade Tropical, Universidade Federal do Amapa, Rod. Juscelino Kubitschek, KM-02 Jardim Marco Zero Macapá - AP CEP 68.903-419, Brazil

[2] Instituto Nacional de Pesquisas da Amazônia, Coordenação de Pesquisas Em Ecologia, Avenida André Araújo, 2936, Cx.Postal 478, Aleixo, Manaus, 69060-001, AM, Brazil

Email: dnorris75@gmail.com

Abstract

Modeling the distribution of rare and endangered species is challenging, and there is substantial debate regarding what species distribution models (SDMs) actually represent. Here I investigated whether locations of different lowland tapir signs (feces, trails and tracks) generated different distributions of suitable habitat using a presence-only species distribution modeling technique. Comparison of the equivalence and overlap of the predicted distributions showed no significant differences between the different signs. The contribution of the 11 variables used to build the distribution models was also similar between signs. Although predictions from different signs were similar, the use of different threshold selection methods generated substantially different suitable areas and omission errors. These results highlight the importance of a fundamental understanding of species natural history to determine not only appropriate model parameters, but also the biological relevance of SDMs. My findings also support the need for healthy skepticism regarding what is represented by presence-only species distributions. To help address this skepticism I conclude by providing guidelines for generating reliable local-scale distribution models.

Key words: Atlantic Forest, habitat suitability, MaxEnt, species distribution modeling, *Tapirus terrestris*

Resumo

A modelagem da distribuição de espécies raras é um desafio e há um debate substancial sobre o que os modelos de distribuição de espécies (MDEs) realmente representam. Aqui, através de nichos ecológicos modelados utilizando apenas dados de presença eu investiguei se sinais indiretos diferentes (fezes, pegadas/carreiros) de anta (*Tapirus terrestris*) resultaram em diferentes distribuições de habitat. Comparações das equivalências e sobreposições das distribuições não apresentaram diferenças significativas entre os diferentes sinais. A contribuição de 11 variáveis utilizadas para construir os modelos de distribuição também foi semelhante entre os tipos de sinais. Embora as distribuições fossem semelhantes, a utilização de diferentes métodos de seleção de limiares gerou diferenças expressivas nas áreas adequadas e erros de omissão. Estes resultados destacam a importância de uma compreensão de história natural da espécie para determinar não só os parâmetros adequados do modelo, mas também a relevância biológica do MDEs. Os resultados também apoiam a necessidade de ceticismo cauteloso em relação ao que é representado pela distribuição das espécies com apenas pontos de presença; e que esse ceticismo deve ser tratado antes de informar as políticas e programas da conservação.

Keywords: Mata Atlântica, adequacão do habitat, MaxEnt, modelagem da distribução de espécies, *Tapirus terrestris*

Received: 21 January 2014; Accepted 1 June 2014; Published: 22 September 2014

Cite this paper as: Norris, D. 2014. Model thresholds are more important than presence location type: Understanding the distribution of lowland tapir (*Tapirus terrestris*) in a continuous Atlantic forest of southeast Brazil. *Tropical Conservation Science* Vol.7 (3):529-547. Available online: www.tropicalconservationscience.org

Introduction

Predicting the geographic distributions of species is a growing field in conservation science [1-4]. Species distribution models (SDMs) permit the analysis of a wide variety of biodiversity phenomena, including future potential distributions under scenarios of climate change, species' invasions, and priorities for biodiversity conservation [3-4]. Yet there is uncertainty about the inferences possible from novel prediction methods [3].

Modeling the distribution of rare and endangered species is challenging not only because acquiring robust empirical field data is often prohibitively expensive (in terms of time and money), but also because many techniques available for modeling species distributions are not appropriate for data that are typically sparse and clustered [3-4]. Novel models can generate accurate and informative predictions from presence-only locations for a variety of faunal and floral species [1-2,5-6], but studies also highlight that model predictions are sensitive to a number of analytic and sampling biases [3,7].

Tapirs (*Tapirus* spp.) characterize the challenges of modeling species distributions in the tropics. Due to tapirs' relatively low densities and secretive nature, indirect signs (such as feces, tracks and trails) have been frequently used to estimate the distribution and abundance of tapir species in numerous tropical biomes [8-13]. Often a combination of different sign types is used for generating tapir distribution models (e.g. [8,14]). Yet it is unclear what can be inferred from the use of these different indirect signs when modeling the species distribution. For example, when predicting distributions is it appropriate to model different signs together (to increase sample size and analytic power)? Do different signs generate different niches and therefore different distribution maps? Answering such questions is vital for understanding whether the inferences made from the predicted distributions are robust and reliable for the study species [3].

The objective of this study was to evaluate whether different lowland tapir (*Tapirus terrestris*) signs (Feces, Tracks and Trails) generated different distributions of suitable habitat from ecological niches modeled using presence-only data. Specifically, as fecal samples occur where lowland tapir have walked, I predicted that the distribution from Feces locations should represent a subset of that obtained from Tracks & Trails. To test this prediction, I used a maximum-entropy algorithm (MaxEnt [2]) to compare the distribution of

suitable habitat derived from locations of *Tapirus terrestris* in a protected area of the Brazilian Atlantic Forest.

Methods

Study area

Surveys took place in Núcleo Caraguatatuba (hereafter Caragua). Caragua is a ≈49,953ha administrative unit of the Serra-do-Mar State Park (Fig. 1, [15]), which protects ≈315,390ha of Atlantic Forest in the Brazilian State of Sao Paulo. The Serra-do-Mar State Park is located along the pre-Cambrian Serra do Mar mountain chain [16]. Caragua is located in the center of the coastal tourist region of Sao Paulo, and receives approximately 5,000 visitors annually [15]. Caragua is bisected by the Tamoios road, a state highway that leads to the town of Caraguatatuba (45° 25' 57'' W and 23° 35' 52'' S). The western portion of Caragua is also traversed by one of the main pipelines of the Brazilian petroleum company "Petrobras." The poorly monitored access provided by the Tamoios highway and the pipeline are the two principal vectors of anthropogenic pressure (including illegal hunting and palm-heart harvesting) in Caragua ([15], p.119-143).

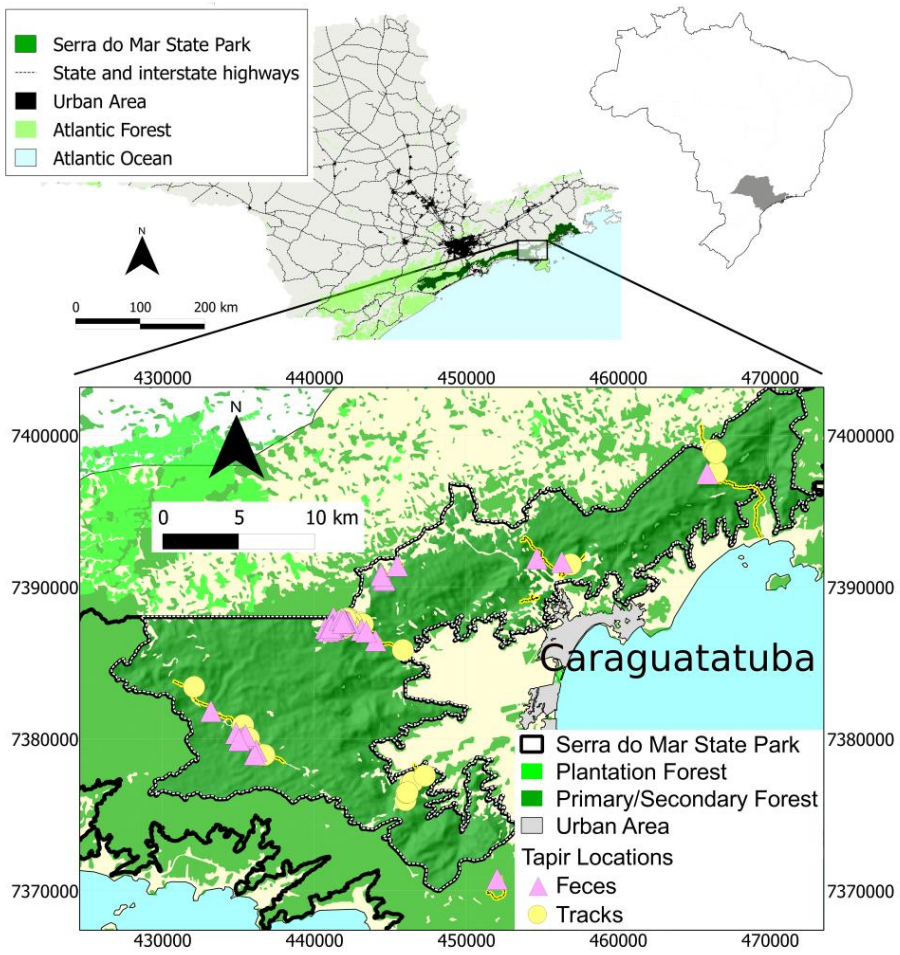


Fig. 1 Location of Núcleo Caraguatatuba in the Serra-do-Mar State Park. Insets show study area in the Brazilian State of Sao Paulo. The detail map shows survey trails (yellow lines) and 141 presence locations used to model the habitat suitability of *Tapirus terrestris* (note presence locations enlarged for clarity, each point spanning 2km). Spatial distribution of locations shown in relation to the survey area (Núcleo Caraguatatuba - shaded relief area with dashed border), prediction area (Serra do Mar State Park – solid black lines) and modelling area (10 km shaded buffer surrounding Núcleo Caraguatatuba).

(full quality image: https://dl.dropboxusercontent.com/u/103821739/fig1v4.tif)

The regional climate is subtropical, with a mean annual temperature of 23.2 ºC (daily means ranging from 4.6 to 36.1 ºC, data from 2010 downloaded from the Brazilian weather center http://www.cptec.inpe.br/, station ID: 83671, Lat -21.98 , Long: -47.35, masl = 598), and annual rainfall from 1,400 to 4,000 mm [16]. Forests range from coastal (≈20 m) to elevations > 900m, with stark floristic gradients from shrubs to well-developed montane forest [15,17].

Tapir locations

Between April and November 2011 (for a total of 61 days), line-transect surveys were used to sample the distribution of *T. terrestris* across Caragua. Diurnal surveys were conducted by two observers, who recorded all indirect signs (feces, tracks and trails) encountered. The locations of these signs were recorded using a GPS (Garmin 60x, horizontal error <9m). Further details of the study area and survey methodology are presented in [18].

To provide a representative sample, pre-existing trails were walked by two observers (n=14, total trail length = 68.8 km, length range = 2.1 – 15.7 km). Trails had been established for at least five years prior to surveys and were distributed throughout the area (Fig. 1), encompassing both the full altitudinal range (20 – 874 masl) and the variety of secondary and primary forest habitats found within the survey area. As rainfall occurs throughout the year in the study region I assume that climatic conditions did not influence sign detectability. As the survey effort was distributed evenly between the austral winter (May to July) and summer months (September to November), I also assume that surveys accounted for any possible seasonal variation in *T. terrestris* distribution.

Environmental data

I considered 11 environmental and anthropogenic predictors (Appendix 1) that (i) based on previous studies [10,19-22] could be important determinants of *T. terrestris* distribution within the Atlantic Forest study area, and (ii) for which the pairwise correlations were less than 0.85 [1] (Pearson correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables [23]). All variable layers were resampled using a common origin to a 1km^2 cell size and projected to the same coordinate system (SAD69, UTM zone 23S). Following [24], this cell size was chosen based on a combination of: (i) the question being asked (i.e., conservation / management requirements of the relatively large study area); (ii) our knowledge of the spatial response (i.e., species that ranges widely across a variety of habitats [10,20-21,25]); and (iii) the spatial properties of the available occurrence data (Appendix 2). All GIS processing was carried out using SAGA GIS (http://www.saga-gis.org/en/index.html) and QGIS (http://www.qgis.org/en/site/).

Comparison of predicted distributions

The distribution of different signs (Feces, Tracks & Trails and All) was modeled using a maximum-entropy approach (MaxEnt version 3.3.3k, download URL: http://www.cs.princeton.edu/~schapire/maxent/; [2,26]). Although a number of different modeling approaches are available for presence-only data [1], I selected MaxEnt as it has been shown to perform relatively well compared to alternative approaches for modeling species such as *T. terrestris* that are widely distributed and represented by a low (5-21) to moderate (38-94) number of presence locations [1,6]. Full details of the MaxEnt modeling are provided in Appendix 3. To facilitate reproduction and validation, the complete MaxEnt output files can be obtained

from the corresponding author or downloaded from: http://sdrv.ms/1dNPVhf . To ensure that modeled differences between signs were not biased by pre-existing differences in the spatial distribution of locations, I compared the spatial scale, intensity and autocorrelation of the different sign locations using diagnostic functions available in the R [27] package "spatstat" [28] prior to MaxEnt modeling (Appendix 2).

The logistic output of MaxEnt generates a map with values ranging from 0 to 1. I interpreted this map as representing the distribution of suitable habitat (i.e. a habitat suitability index (HSI), Appendix 3). To compare the predicted distribution from the different types of sign, I examined equivalence and overlap [29] of suitable habitat using functions available in ENMTools [30] and/or the R packages SDMtools [31] and dismo [32]. To examine the relative ranking of variables used to generate model predictions and their importance in the models, I compared the ranked order of variable contributions (standard MaxEnt output) using the Chi-squared test.

Finally, I compared the area of suitable habitat obtained for each sign using seven of the threshold selection methods available in MaxEnt: Minimum training presence, Fixed cumulative value 1, Fixed cumulative value 5, Fixed cumulative value 10, 10 percentile training presence, Equal training sensitivity and specificity, and Maximum training sensitivity plus specificity. For each selection method, I used the mean logistic threshold value from the 50 runs and calculated the omission error (proportion of all locations with values below the threshold) and the proportion of the prediction area classified as "absent" (i.e. unsuitable areas below the threshold) for each sign. As the thresholds were used to compare the final mean prediction maps considering all locations, the results should not be compared with those reported by MaxEnt.

Results

From 354.5 survey km a total of 141 *T. terrestris* presence locations were obtained (Table 1) from Tracks & Trails (63) and Feces (78). Excluding duplicates within the same 1km^2 pixel reduced the number of locations to 80, 50, and 39 (All, Tracks & Trails, and Feces respectively). Although spatial diagnostics showed that the spatial scale, intensity and autocorrelation of Tracks & Trails and Feces were similar (Appendix 2), examination of nearest neighbor distances (within a radius of 3km) showed that locations of feces tended to be more clustered (mean distances = 81.2, 115.4 m, Feces and Tracks & Trails respectively, Mann-Whitney test, $P<0.0001$). When duplicates within the same 1km cell were excluded, there was no significant difference between mean nearest neighbor distances (mean distances within a 3km radius = 1,233.4, 1,158.7 m, Feces and Tracks & Trails respectively, Mann-Whitney test, $P=0.466$).

Table 1. MaxEnt model summary. Summary of MaxEnt models used to predict the distribution of *Tapirus terrestris* within an Atlantic forest protected area using presence locations from different sign types.

Sign type	[a] Locations	[b] AUC
All	141 (80)	0.92 (0.022)
Tracks & Trails	63 (50)	0.90 (0.032)
Feces	78 (39)	0.92 (0.029)

[a] Number of *T. terrestris* locations and in parenthesis the number of locations excluding duplicates within the same 1km^2 grid cell.

[b] Mean and SD (in parenthesis) of AUC values from the test data of 50 MaxEnt model replicates.

MaxEnt model predictions appeared to be accurate, with mean test data AUC > 0.90 for the three types of locations (Table 1), which is a good score for the model validation [26]. Not only were AUC means similar but the standard deviation of AUC values was also similar among the three types of sign (Table 1), which suggests that differences in sample size did not influence the appropriateness of the MaxEnt models.

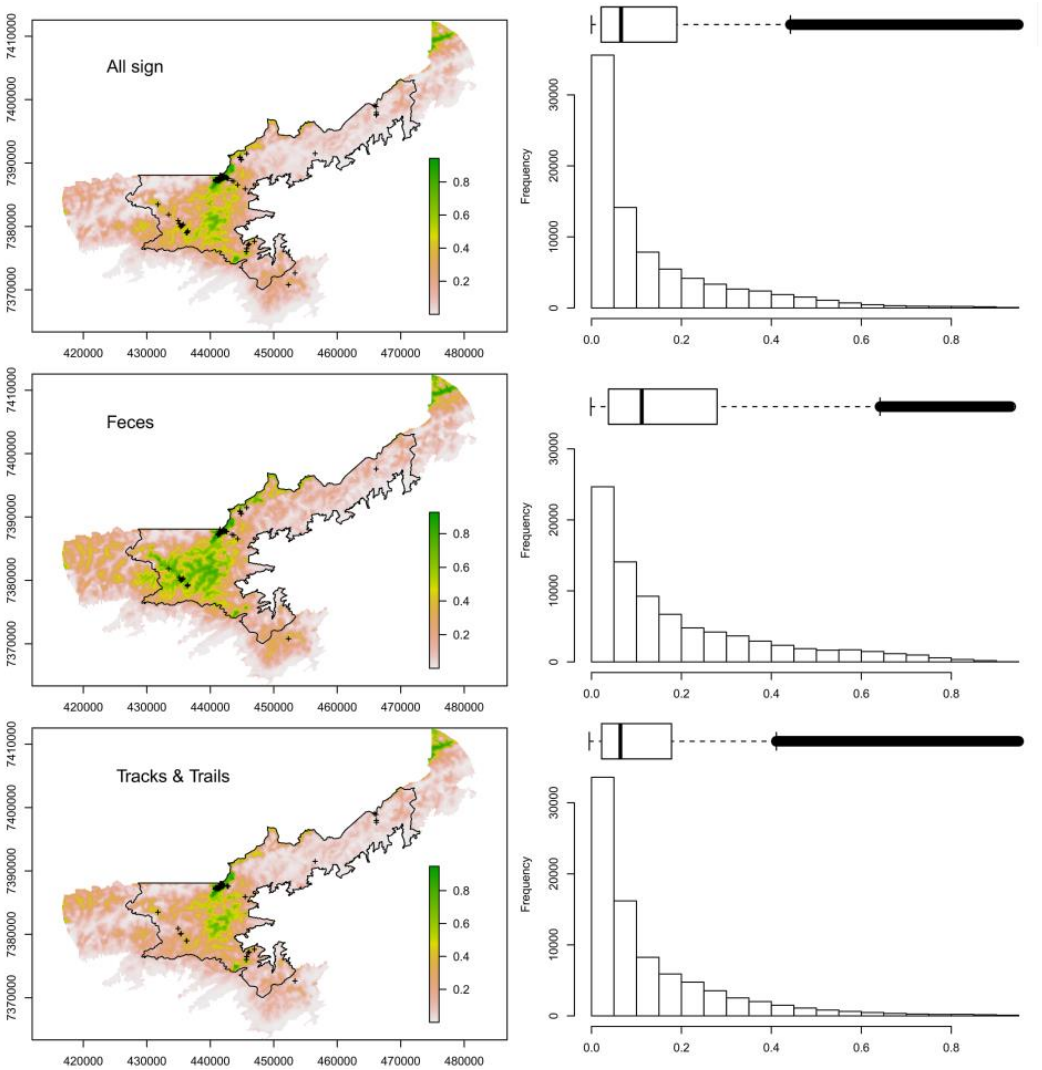


Fig. 2 Predicted distribution of *Tapirus terrestris.* Maps show the spatial distribution of suitable habitat (least (0) to most (1) suitable) and presence locations (crosses) within the prediction area obtained from three sign types (All, Feces, Tracks & Trails). Solid black outline shows Núcleo Caraguatatuba ("survey area"). Associated histograms show frequency distribution of HSI values within the prediction area, with boxplots showing median (bold vertical line) and the first and third quartiles (hinges).
(full quality image: https://dl.dropboxusercontent.com/u/103821739/fig2USA.tif)

The predicted distribution of suitable habitat for *T. terrestris* was not homogeneous in the study area (Fig. 2). Visual comparison suggested that the distribution of suitable habitat was similar in the maps generated from the three sign types (Fig. 2). This visual assessment was confirmed by the similarity in the frequency distributions (Fig. 2, Kolmogorov-Smirnov, P>0.537 for all three pairwise comparisons) and strong correlations between the mapped habitat suitability values derived from the three sign types (Pearson's correlation r>0.85, P<0.0001 for all three pairwise comparisons). This descriptive analysis of MaxEnt predictions was also supported by hypothesis tests that showed the predicted distributions overlapped and were ecologically equivalent (Table 2).

Table 2. Overlap and equivalence of predicted distributions. Comparison of observed overlap (a) and equivalence (b) between the distributions from different sign, using both I (lower diagonal) and D (upper diagonal) similarity metrics. Significance of equivalence tested by the randomization test of [29], where a significant value denotes a pair of sign that are ecologically distinct (ns = not significant i.e. ecologically similar).

(a)	All	Tracks & Trails	Feces
All		0.88 ^{ns}	0.83 ^{ns}
Tracks & Trails	0.99 ^{ns}		0.82 ^{ns}
Feces	0.98 ^{ns}	0.97 ^{ns}	

(b)	All	Tracks & Trails	Feces
All		0.81 ^{ns}	0.74 ^{ns}
Tracks & Trails	0.97 ^{ns}		0.65 ^{ns}
Feces	0.93 ^{ns}	0.89 ^{ns}	

Pearson's chi-squared test showed that the type of sign did not significantly influence the observed importance (ranked contribution) of the different variables (X-squared=13.557, P=0.921). Although the mean contribution of some variables differed slightly among sign types, variable importance was generally consistent among the three sign types (Fig. 3). Generally there was a clear separation with only five variables showing important contributions: distance to park border, distance to road, distance to river, altitude, and proportion of forest within a 5km radius (summed variable contribution % = 94.3, 85.9, 92.5, for All, Feces, and Tracks & Trails respectively). The other six variables contributed little on average (<5% each, Fig. 3).

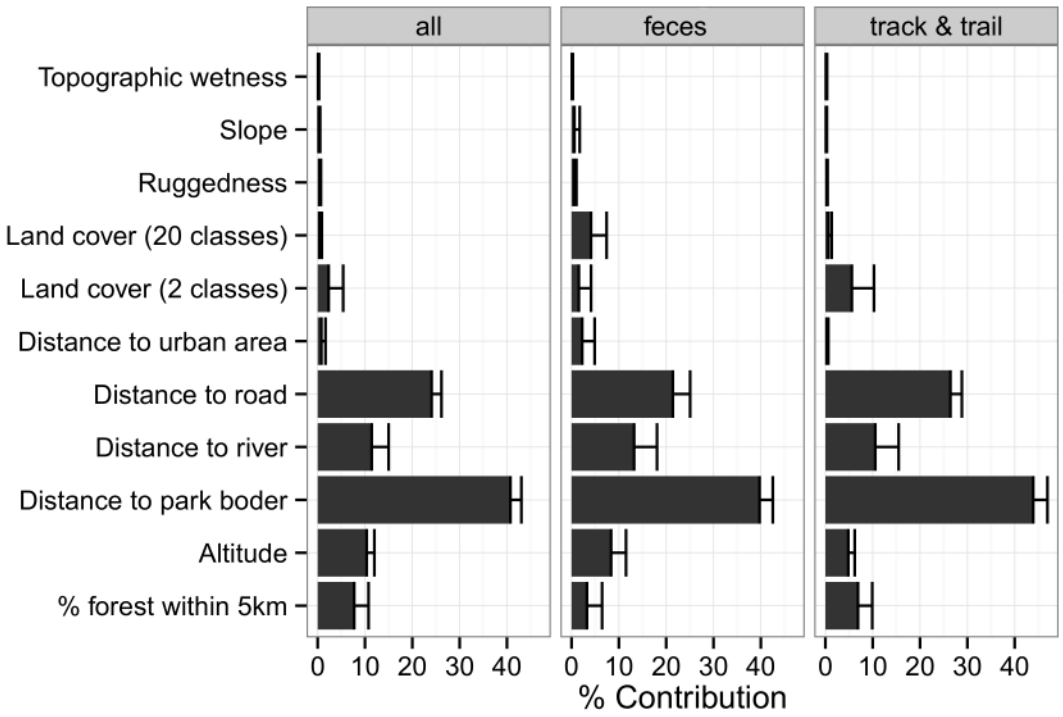
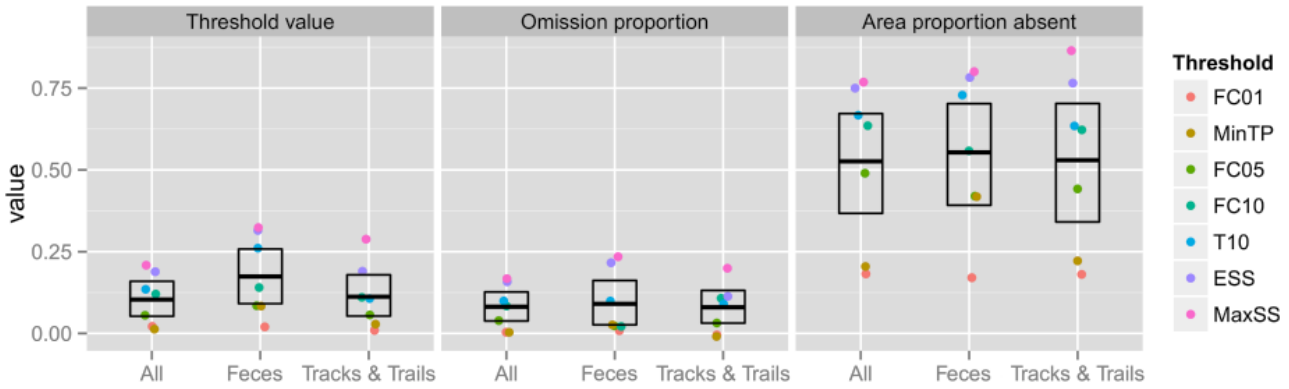


Fig. 3. Variable contribution of *Tapirus terrestris* predicted distributions. Percentage contribution (mean plus 1SD) of 11 variables used to model the distribution of three sign types (All, Feces, Tracks & Trails).
(full quality image: https://dl.dropboxusercontent.com/u/103821739/fi3bUSAflat.tif)

Using different threshold methods resulted in substantial differences in omission errors and unsuitable areas (Fig. 4), yet results from the three different sign types showed small and insignificant differences between the average threshold values (Kruskal-Wallis Rank Sum Test, $P>0.1$, for the comparison of threshold, omission and unsuitable area values, Fig. 4). Only Minimum Training presence and Fixed cumulative value 1 had zero omission error, all presence locations being correctly retained with values above the threshold. On the other hand, Maximum training sensitivity plus specificity and Equal training sensitivity and specificity generated the most omission errors (between 15 and 21%). The different threshold selection methods also resulted in substantial differences in the predicted distributions, with the unsuitable area ranging from 18 to 85% of the prediction area depending on the method adopted (Fig. 4). Again, the greatest contrast was between values from Minimum Training presence and Fixed cumulative value 1 (smallest unsuitable areas) and those obtained from Maximum training sensitivity plus specificity and Equal training sensitivity and specificity (largest unsuitable areas) (Fig. 4).



(Fig. 4 Threshold comparison. Comparison of the threshold values, omission proportion and proportion of area absent obtained from seven of the thresholds available in MaxEnt. Colored dots show values from the different thresholds: Minimum training presence (MinTP), Fixed cumulative value 1 (FC01), Fixed cumulative value 5 (FC05), Fixed cumulative value 10 (FC10), 10 percentile training presence (T10), Equal training sensitivity and specificity (ESS) and Maximum training sensitivity plus specificity (MaxSS). Boxplots show median (bold horizontal line) and the first and third quartiles (hinges). (full quality image: https://dl.dropboxusercontent.com/u/103821739/fig4USAflat.tif)

Discussion

There is substantial debate about what species distribution and ecological niche models actually represent, and perhaps more importantly, how these two concepts are related [3]. My findings support the necessity of "a healthy skepticism about which components of the niche are represented by predictions from an SDM" [3].

Theoretically, the *T. terrestris* presence locations are "natural distribution data" representing a "realized niche" [33-34]. The habitat suitability from different signs therefore reflects the use of the available habitats by *T. terrestris* [33-34]. If this were true, the maps of habitat suitability for *T. terrestris* would be extremely informative for researchers and park managers in the study area, particularly as the Serra-do-Mar continues to be intensely threatened by anthropogenic perturbations. However, there are substantial challenges to modeling highly mobile species, which also typically (through sample bias and/or natural history) exhibit spatial autocorrelation (but see [35] for an alternative approach for generating a "realized" distribution).

What the presence locations from different signs actually represent (i.e. their ecological meaning) is undeniably contentious and perhaps irrelevant from the perspective of modeling species distributions [36]. However, what the predicted distributions actually represent does have important implications from the perspective of species management and conservation. For example, feces are an important source for genetic studies, and improving fecal sampling efficiency is an active research area in the tropics e.g. Neotropical carnivores [37] and deer [38]. Predicting the likely location of fecal samples would improve sampling efficiency and reduce costs. However, the similarity found between the predicted distributions of Feces and Tracks & Trails suggests that the approach adopted, which evaluates correlative rather than

mechanistic relationships [2-3], is not adequate for such purposes. In other words, the maps derived from these signs may show a substantial portion of the confirmed distribution of the species, but are not necessarily suitable for establishing ecological associations.

From another perspective, the similarity between the distributions of the different sign types can be useful for conservation and management activities. The similarity of the predicted distributions means that it should be possible to combine different signs to increase the statistical robustness of *T. terrestris* distribution models. For example, increasing the sample size enables the adoption of additional modeling parameters/features and improved predictions [6-7,39-40]. Additionally, this similarity suggests that results from studies using different combinations of feces and/or tracks and trails should also be comparable. However, the findings also highlight that the ability to compare results will strongly depend on the model parameters and threshold selection methods applied.

Threshold selection is one of the many possible biases in species distribution modeling [2,7,40-41], yet few studies have evaluated the influence of threshold selection for presence-only data. In a recent study, Liu et al. [42] evaluated the suitability of 13 threshold selection methods for presence-only data using simulated species. These authors found Maximum training sensitivity plus specificity to be a promising selection method for presence-only data. This result contrasts with my findings (using a "real" species to represent the challenges typical of tropical species) that Maximum training sensitivity plus specificity resulted in both the greatest omission error and increased loss of suitable areas among all the predicted distributions of *T. terrestris*.

The determination of thresholds should not be arbitrary and should consider the relative importance of omission and commission errors [6,41-42]. For the case of *T. terrestris* in Caragua, reducing omission error is the most important determinant of threshold selection method, because this wide-ranging and long-lived species is likely to find suitable conditions throughout the prediction area. In the present study, both sample sizes and spatial arrangement were similar and MaxEnt model parameters standardized such that the differences in suitable areas and omission rates can be directly attributable to the threshold selection method adopted. But which threshold should be chosen? Although *T. terrestris* are rare, the natural history of this charismatic and widespread species is well studied [20-21]. *T. terrestris* occur in dry forests (e.g. Bolivian tropical and subtropical dry Chaco forests [20,43]) to tropical regions, and lowland to upland habitats [20]. Therefore, based on such knowledge, it is possible to conclude that the threshold selection methods which resulted in lower threshold values, i.e. with a wider distribution of suitable habitat and close to zero omission error (Minimum training presence or Fixed cumulative value 1) should be the most appropriate to identify suitable and unsuitable areas for *T. terrestris* in the Serra do Mar protected area.

Considering the threshold values from the Minimum training presence and Fixed cumulative value 1, an average of 23% of the prediction area (equivalent to 18,900 ha) is unsuitable for *T. terrestris*. For all predicted distributions, the most important variables were related to anthropogenic access (Distance to park border and Distance to road). These results are unsurprising, given the location of Caragua and the impact of intense anthropogenic pressures (both current and historic). Previous studies have shown the importance of anthropogenic (distance to park border and distance to road) and environmental variables (altitude and forest cover) as determinants of the distributions of other endangered mammals in the region, such as white-lipped peccary (*Tayassu pecari*, [35]) and buffy-tufted-ear marmosets (*Callithrix aurita*, [44]). Whilst anthropogenic factors are obviously important determinants of mammal distributions

in remaining Atlantic Forest areas, the results from *T. terrestris* highlight limitations of correlative distribution models.

Correlative distribution models provide a simple output (distribution map) that indirectly represents many different processes [45-46]. Should the distribution change over time, for example, in response to management actions within the protected area, we are left with an uncomfortable, nagging uncertainty about what has actually changed (individual behavior, population demographics, etc.). Recent studies have started to explore how to develop mechanistic distribution models that incorporate aspects such as physiology [46] and population demographics [47]. However, such approaches are still developmental [45] and the use of such models requires additional data that are not readily available for the majority of tropical species.

Implications for conservation

Modeling presence locations to represent broad scale species distributions undoubtedly provides robust and reliable inferences, but my findings highlight some of the challenges for local scale predictions [3]. With appropriate sampling, model building, and threshold selection, it is possible to gain an understanding of local scale distributions from presence-only locations, yet what these distributions actually represent remains unclear [1-2,7,36].

Theoretical [3,36] and statistical [48] advances in species distribution modeling must be accompanied by the collection of more detailed field data. For example, if we knew that individual *T. terrestris* mark their range limits with fecal samples, the inferences possible from the models would be entirely different from the case that fecal samples represent areas that are more commonly used by individuals. While previous reviews highlight the importance of integrating theory [3], improving modeling methods [36] and improving data quality/availability [4], my findings suggest that integrating complementary data regarding species natural history (understanding the how and why of species distributions) is vital to generate meaningful conservation insight from local scale presence-only species distribution models.

Finally, findings from the analysis conducted and from previous studies [3,24,35,49] enable me to present some practical guidelines for generating reliable local scale distribution models that should be useful for conservation practitioners:

- Spatial resolution must be established prior to distribution modeling, based on (i) the question being addressed, (ii) knowledge of the spatial response and (iii) spatial properties of the available occurrence data.
- Spatial autocorrelation should be quantified and if necessary implicitly addressed within the modeling workflow.
- Ensure surveys are conducted in a representative sample of the environmental gradients in the survey area. Prediction outside of the sampled gradients is not recommended.
- Potential biases caused by the integration of different sources and types of occurrence data should be evaluated as part of the modeling workflow.
- Environmental data must be selected based on (i) previous studies, (ii) availability at the required spatial resolution, and (iii) accessibility (freely available data sources that allow reproduction should be preferred).

- The selection of modeling algorithm (or algorithms in the case of ensemble models) should be based on (i) study objectives, (ii) species natural history and (iii) spatial properties of the available occurrence data.
- Because modeling algorithms are increasingly complex, model inputs and outputs must be available for independent peer review and validation by experts.

Acknowledgements

I am deeply indebted to José F. Moreira-Ramírez for his tireless help and dedication during field work. I thank Milton Ribeiro who assisted with the GIS processing and together with participants of the "Ecologia de Paisagem" course inspired much of the analysis. I also thank Mauro Galetti, Alexandra Sanches, Rafael Loyola, Jose Alexandre F. Diniz-Filho and two anonymous reviewers for comments that improved a previous version. Financial support came from a Rufford Small Grant for Nature Conservation and post-graduate scholarships from CNPq (159806/2012-7 and 164999/2013-2). I also thank UNESP (Rio Claro) for logistical support and the Instituto Florestal de São Paulo for permission to conduct research (COTEC-SMA: 260108014.661/010).

References

[1] Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129-151.

[2] Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31: 161-175.

[3] Elith, J. and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677-697.

[4] Cayuela, L., Golicher, D., Newton, A., Kolb, H., de Alburquerque, F., Arets, E., Alkemade, J. and Pérez, A. 2009. Species distribution modeling in the tropics: problems, potentialities, and the role of biological data for effective species conservation. *Tropical Conservation Science* 2: 319-352.

[5] Pearson, R. G., Raxworthy, C. J., Nakamura, M. and Townsend Peterson, A. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* 34: 102-117.

[6] Hernandez, P. A., Franke, I., Herzog, S. K., Pacheco, V., Paniagua, L., Quintana, H. L., Soto, A., Swenson, J. J., Tovar, C., Valqui, T. H., Vargas, J. and Young, B. E. 2008. Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation* 17: 1353-1366.

[7] Bean, W. T., Stafford, R. and Brashares, J. S. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography* 35: 250-258.

[8] Clements, G. R., Mark Rayan, D., Aziz, S. A., Kawanishi, K., Traeholt, C., Magintan, D., Yazi, M. F. A. and Tingley, R. 2012. Predicting the distribution of the Asian Tapir (*Tapirus indicus*) in Peninsular Malaysia using maximum entropy modelling. *Integrative Zoology* 7: 400-406.

[9] Norris, D., Peres, C. A., Michalski, F. and Hinchsliffe, K. 2008. Terrestrial mammal responses to edges in Amazonian forest patches: a study based on track stations. *Mammalia* 72: 15-23.
[10] Salas, L. A. 1996. Habitat of the lowland tapir (*Tapirus terrestris* L) in the Tabaro River valley, southern Venezuela. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 74: 1452-1458.
[11] Sánchez-Núñez, E., Nájera, H. E. O. and Nicolás, E. A. 2011. Abundancia y uso de hábitat del tapir (*Tapirus bairdii*) en Frontera Corozal, Selva Lacandona, Chiapas, México. *Tapir Conservation* 20: 25-29.
[12] Bodmer, R. 1991. Influence of digestive morphology on resource partitioning in Amazonian ungulates. *Oecologia* 85: 361-365.
[13] Hill, K., Padwe, J., Bejyvagi, C., Bepurangi, A., Jakugi, F., Tykuarangi, R. and Tykuarangi, T. 1997. Impact of hunting on large vertebrates in the Mbaracayu reserve, Paraguay. *Conservation Biology* 11: 1339-1353.
[14] Salas, L. A. and Fuller, T. K. 1996. Diet of the lowland tapir (*Tapirus terrestris* L) in the Tabaro River valley, southern Venezuela. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 74: 1444-1451.
[15] Instituto Florestal 2008. *Parque Estadual da Serra do Mar Plano de Manejo*. Instituto Florestal do Estado de Sao Paulo, Sao Paulo.
[16]. Mantovani, W. 1993. *Estrutura e dinâmica da floresta Atlântica na Juréia, Iguape-SP*. Tese de Livre Docência, Universidade de São Paulo, São Paulo, Brazil.
[17] Veloso, H. P., Rangel-Filho, A. L. R. and Lima, J. C. A. 1991. *Classificação da vegetação brasileira adaptada a um sistema universal*. IBGE, Rio de Janeiro, Brazil.
[18] Norris, D., Moreira Ramírez, J., Zacchi, C. and Galetti, M. 2012. A survey of mid and large bodied mammals in Núcleo Caraguatatuba, Serra do Mar State Park, Brazil. *Biota Neotropica* 12: 127-133.
[19] Licona, M., McCleery, R., Collier, B., Brightsmith, D. J. and Lopez, R. 2011. Using ungulate occurrence to evaluate community-based conservation within a biosphere reserve model. *Animal Conservation* 14: 206-214.
[20] Naveda, A., de Thoisy, B., Richard-Hansen, C., Torres, D. A., Salas, L., Wallance, R., Chalukian, S. and de Bustos, S. 2008. *Tapirus terrestris*. In: *IUCN 2013. IUCN Red List of Threatened Species*. Version 2013.2. www.iucnredlist.org
[21] Garcia, M. J., Medici, E. P., Naranjo, E. J., Novarino, W. and Leonardo, R. S. 2012. Distribution, habitat and adaptability of the genus Tapirus. *Integrative Zoology* 7: 346-355.
[22] Wallace, R., Ayala, G. and Viscarra, M. 2012. Lowland tapir (*Tapirus terrestris*) distribution, activity patterns and relative abundance in the Greater Madidi-Tambopata Landscape. *Integrative Zoology* 7: 407-419.
[23] Fox, J. 2010. *Polycor: Polychoric and Polyserial Correlations. R package version 0.7-8*. http://CRAN.R-project.org/package=polycor
[24] Hengl, T. 2009. *A practical guide to geostatistical mapping.* 2nd edn. Office for Official Publications of the European Communities, Luxembourg.
[25] Fragoso, J. M. V., Silvius, K. M. and Correa, J. A. 2003. Long-distance seed dispersal by tapirs increases seed survival and aggregates tropical trees. *Ecology* 84: 1998-2006.
[26] Phillips, S. J., Anderson, R. P. and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231-259.
[27] R Core Team. 2013. *R: A language and environment for statistical computing 3.0.2*. http://www.R-project.org/

[28] Baddeley, A. and Turner, R. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software* 12: 1-42.
[29] Warren, D. L., Glor, R. E. and Turelli, M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62:2868-2883.
[30] Warren, D. L., Glor, R. E. and Turelli, M. 2010. ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography* 33: 607-611.
[31] VanDerWal, J., Falconi, L., Januchowski, S., Shoo, L. and Storlie, C. 2012. *SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1-13.* http://CRAN.R-project.org/package=SDMTools
[32] Hijmans, R. J., Phillips, S., Leathwick, J. and Elith, J. 2013. *dismo: Species distribution modeling. R package version 0.9-3*. http://CRAN.R-project.org/package=dismo
[33] Hutchinson, G. E. 1959. Homage to Santa Rosalia or why are there so many kinds of animals? *American Naturalist* 93: 145-159.
[34] Booth, T. H., Nix, H. A., Hutchinson, M. F. and Jovanic, T. 1988. Niche analysis and tree species introduction. *Forest Ecology and Management* 23: 47-59.
[35] Norris, D., Rocha-Mendes, F., de Barros Ferraz, S. F., Villani, J. P. and Galetti, M. 2011. How to not inflate population estimates? Spatial density distribution of white-lipped peccaries in a continuous Atlantic forest. *Animal Conservation* 14:492-501.
[36] Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33: 1677-1688.
[37] Michalski, F., Valdez, F. P., Norris, D., Zieminski, C., Kashivakura, C. K., Trinca, C. S., Smith, H. B., Vynne, C., Wasser, S. K., Metzger, J. P. and Eizirik, E. 2011. Successful carnivore identification with faecal DNA across a fragmented Amazonian landscape. *Molecular Ecology Resources* 11: 862-871.
[38] de Oliveira, M. L., Norris, D., Ramirez, J. F. M., Peres, P. H. D., Galetti, M. and Duarte, J. M. B. 2012. Dogs can detect scat samples more efficiently than humans: an experiment in a continuous Atlantic Forest remnant. *Zoologia* 29: 183-186.
[39] Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. and Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17: 43-57.
[40] Syfert, M. M., Smith, M. J. and Coomes, D. A. 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *Plos One* 8: e55158.
[41] Nenzén, H. K. and Araújo, M. 2011. Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling* 222: 3346-3354.
[42] Liu, C., White, M. and Newell, G. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography* 40: 778-789.
[43] Noss, A. J., Gardner, B., Maffei, L., Cuellar, E., Montano, R., Romero-Munoz, A., Sollman, R. and O'Connell, A. F. 2012. Comparison of density estimation methods for mammal populations with camera traps in the Kaa-Iya del Gran Chaco landscape. *Animal Conservation* 15: 527-535.
[44] Norris, D., Rocha-Mendes, F., Marques, R., Nobre, R. d. A. and Galetti, M. 2011. Density and Spatial Distribution of Buffy-tufted-ear Marmosets (*Callithrix aurita*) in a Continuous Atlantic Forest. *International Journal of Primatology* 32: 811-829.
[45] Buckley, L. B., Urban, M. C., Angilletta, M. J., Crozier, L. G., Rissler, L. J. and Sears, M. W. 2010. Can mechanism inform species' distribution models? *Ecology Letters* 13: 1041-1054.
[46] Kearney, M. and Porter, W. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters* 12: 334-350.

[47] Keith, D. A., Akçakaya, H. R., Thuiller, W., Midgley, G. F., Pearson, R. G., Phillips, S. J., Regan, H. M., Araújo, M. B. and Rebelo, T. G. 2008. Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters* 4: 560-563.

[48] Renner, I. W. and Warton, D. I. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*.

[49] Elith, J., Kearney, M. and Phillips, S. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* 1: 330-342.

[50] Lyra-Jorge, M., Ribeiro, M., Ciocheti, G., Tambosi, L. and Pivello, V. 2010. Influence of multi-scale landscape structure on the occurrence of carnivorous mammals in a human-modified savanna, Brazil. *European Journal of Wildlife Research* 56: 359-368.

[51] Sappington, J., Longshore, K. M. and Thompson, D. B. 2007. Quantifying landscape ruggedness for animal habitat analysis: a case study using bighorn sheep in the Mojave Desert. *The Journal of wildlife management* 71: 1419-1426.

[52] Baddeley, A. 2008. *Analysing spatial point patterns in R*. www.csiro.au/resources/pf16h.html

Appendix 1 Environmental variables used to predict the distribution of *Tapirus terrestris* in the Serra do Mar State Park, Brazil

Variable	Description/Ecological characteristics	Grid	Transect mean values (range)	[b] Survey area mean values (range)	[c] Pred. area mean values (range)	[d] Modeling area mean values (range)	Source/Instrument
Altitude	Digital elevation model (DEM): topography	30m	353.2 (21.0-803)	608.5 (19-1286)	612.3 (27-1615)	477.4 (0-1615)	[e] ASTER
Slope	Topographic morphometry/suitability for quadraped locomotion	30m	10.0 (0.1-32.3)	15.6 (0-57.6)	16.6 (0.08-86)	10.3 (0-86.0)	From DEM
Topographic wetness	Topographic wetness index	30m	9.4 (4.8-16.5)	6.9 (3.2-19.7)	6.8 (1.1-19.7)	9.2 (1.1-21.4)	From DEM
[K] Terrain ruggedness	Topographic morphometry/suitability for quadraped locomotion	30m	0.0 (0-0.08)	0.0 (0-0.18)	0.0 (0.0-0.67)	0.0 (0-0.67)	From DEM
Distance to river (km)	Straight line distance to nearest river channel	30m	0.3 (0-1.0)	0.4 (0-1.4)	0.4 (0-2.0)	0.4 (0-3.3)	From stream-channel network derived from DEM
Land cover (2 classes)	Forest (primary, secondary and plantation) and non-forest (pasture, urban areas etc).	100m	Categorical				Derived from land cover shapefile [L]
Land cover (10 classes)	Land cover type	100m	Categorical				Derived from land cover shapefile [L]
% forest in 5km radius [j]	Coarse scale influence of forest cover	200m	82.1 (11.0-100.0)	91.9 (0-100)	93.4 (0-100.0)	47.7 (0-100)	Derived from land cover shapefile
Distance to border of PA (km)	Distance from park border/human accessibility and disturbance (negative values outside/ positive inside)	30m	1.1 (-0.9-4.2)	2.0 (-0.1-6.9)	1.9 (0.1-6.9)	-1.5 (-10.1-6.9)	[f] Derived from park boundary vector
Distance to asphalt road (km)	Distance from asphalt road/human accessibility	30m	4.5 (0.4-8.5)	6.3 (0-12.9)	5.6 (0.0-12.9)	3.5 (0-14.1)	[g] Derived from road network vector
Distance to urban area (km)	Distance from urban area/ proxy for human impact and accessibility	30m	4.6 (0.6-10.6)	6.7 (0-16.1)	5.9 (0.0-16.0)	6.8 (0-20.6)	[h] Derived from urban area vector

[a] Native cell resolution (meters) of raster grid.
[b] Survey area: the area encompassed by the survey trails (i.e. Caragua: 49,953 ha)
[c] Prediction area: includes any of the modeling area that was within the Serra-do-Mar protected area (totaling 82,561 ha)
[d] Modeling area: the survey area plus a 10km buffer used in MaxEnt modeling to avoid analytic edge effects (totaling 218,675 ha)

[e] ASTER global digital elevation model (product of METI and NASA 2009), distributed by the Land Processes Distributed Active Archive Center (LPDAAC), located at the US Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center (lpdaac.usgs.gov). Downloaded via https://wist.echo.nasa.gov

[f] Compiled and maintained by the Brazilian environment agency (IBAMA). Downloaded via http://www.ibama.gov.br/zoneamento-ambiental

[g] From the remote-sensing center (CSR – Centro de Sensoriamento Remoto) of the Brazilian environmental agency (IBAMA). Downloaded via http://siscom.ibama.gov.br/shapes/

[h] Produced by the SOS Mata Atlântica/INPE 2008. Atlas dos remanescentes florestais da Mata Atlântica, período de 2000 a 2005. http://www.sosmatatlantica.org.br

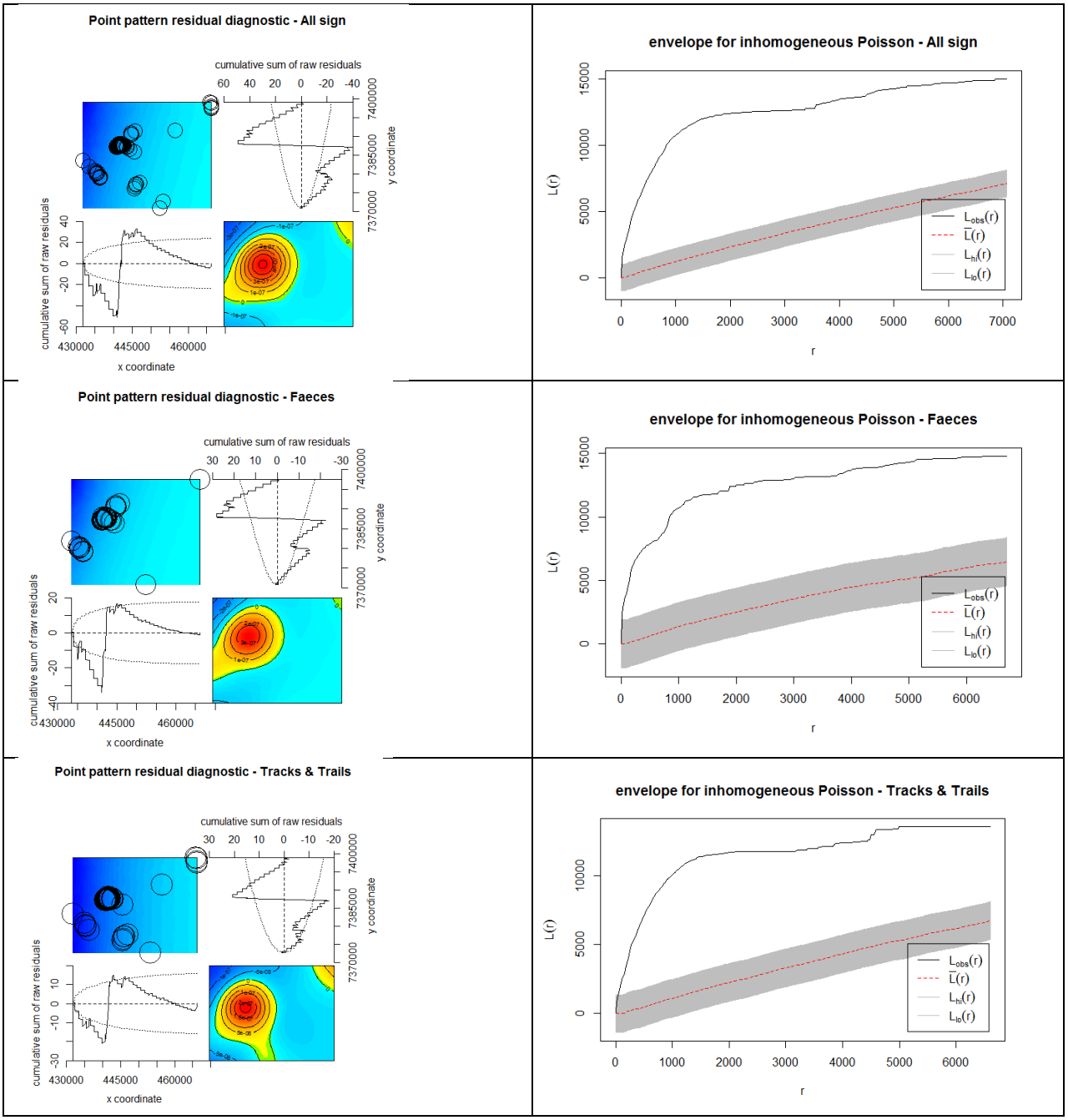
[j] See [50] for GIS procedures used. Forest cover derived from Atlantic Forest vegetation shapefile, reference year 2005 (www.sosma.org.br and www.inpe.br) produced by the SOS Mata Atlântica/INPE 2008. Atlas dos remanescentes florestais da Mata Atlântica, período de 2000 a 2005. http://www.sosmatatlantica.org.br.

[K] Algorithm of [51], implemented in SAGA GIS (module "Vector ruggedness measure").

[L] Available from the Brazilian research program BIOTA-FAPESP (http://www.fapesp.br/6259).

Appendix 2 Spatial diagnostics

Spatial diagnostics were calculated using functions ("diagnose.ppm" and "envelope") in the R package spatstat version 1.35-0 [28] as described in [52]. See details presented in [28,52] for theoretical background and interpretation. From these diagnostics it is clear that the different sign have similar spatial characteristics. Any difference between MaxEnt distributions is therefore unlikely to be attributable to differences in spatial patterns.



Appendix 3 MaxEnt modeling

MaxEnt is a machine-learning method that estimates the probability distribution of a particular species following the principle of maximum entropy, and assumes that all environmental constraints that regulate the presence of a species are included in the estimation procedure [26]. It is now 10 years since MaxEnt became available and during this time there have been substantial refinements/additions to the algorithms and options available to users [2,39-40]. As my objective was to compare the distribution models obtained from the different types of sign, I followed recent studies that provide general guides [2,39,49], using program default settings unless otherwise stated.

It is possible to model species occurrence with MaxEnt [26], but this can only be done when the following assumption is valid [39]: "as long as we can interpret logistic output as corresponding to a temporal and spatial scale of sampling that results in a 50% chance of the species being present in suitable areas". In other words for MaxEnt to represent species occurrence the logistic output must reflect the probability of presence in relation to sites with ''typical'' conditions for the species [39]. This assumption cannot be accepted for *Tapirus terrestris* in Caragua due to a combination of species natural history and sampling design. Therefore, I refer to MaxEnt output as representing "habitat suitability" not species occurrence.

I defined three areas in the analysis: survey area refers to the area encompassed by survey trails (i.e. Caragua: 49,953ha), modeling area was the survey area plus a 10km buffer used in MaxEnt modeling to avoid analytic edge effects, and finally the prediction area includes any of the Serra-do-Mar protected area [3,24,35,49]within the modeling area, a subset (totaling 82,561ha) of the modeling area that contains abiotic and biotic conditions similar to those of the survey area. Unless otherwise stated, all results correspond to the prediction area.

Presence records (excluding duplicate locations within the same $1km^2$ pixel) and 11 environmental variables were used for MaxEnt modeling. The location datasets were sampled by the subset method with a random seed to obtain 50 different random test/training partitions with 70% of presence records used for training (calibration) and 30% for testing/evaluating models [2,26]. All runs (MaxEnt model replicates) were set with a convergence threshold of 1.0 E-5 with 500 iterations and 10,000 background points (MaxEnt software default). I changed the default regularization parameter from 1 to 2.5 [49]. Regularization refers to model smoothing, and increasing the value avoids fitting locally complex models, reducing the effects of spatial autocorrelation and model over-fitting [39,49]. I used only linear and quadratic features for model parametization [2,39,49], as this option tends to reduce the effects of model over-fitting compared with the default ("Auto") setting.

The final logistic output ("habitat suitability") was obtained from the mean value for each pixel based on the 50 MaxEnt replicates. I used the area under the curve (AUC) to evaluate the appropriateness of the models [26]. Although the use of AUC has been challenged for presence-only data it remains the *de facto* standard for MaxEnt model evaluation [2]. This index provides a rank, where a random map has, on average, an AUC of 0.5, and a map with perfect predictions achieves the best possible AUC of 1.0, although when presence-only locations and random points are used to calculate AUC, its maximum value is <1.0 [26].