Ecological Indicators 118 (2020) 106735

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Integrating multi-source data for wildlife habitat mapping: A case study of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) in Yunnan, China

Guiming Zhang[a,*], A-Xing Zhu[b,c,d,e,f], Yu-Chao He[g], Zhi-Pang Huang[h,i], Guo-Peng Ren[h,i], Wen Xiao[h,i]

[a] *Department of Geography and the Environment, University of Denver, Denver, USA*
[b] *Department of Geography, University of Wisconsin-Madison, Madison, USA*
[c] *Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing, China*
[d] *State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, China*
[e] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China*
[f] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China*
[g] *Administration of Yunling Provincial Nature Reserve, Lanping County, China*
[h] *Institute of Eastern-Himalaya Biodiversity Research, Dali University, Dali, China*
[i] *Collaborative Innovation Center for Biodiversity and Conservation in the Three Parallel Rivers Region of China, Dali, China*

ARTICLE INFO

Keywords:
Habitat suitability mapping
Wildlife sightings
Local villagers
Patrol records
Data integration
Black-and-white snub-nosed monkey
(*Rhinopithecus bieti*)

ABSTRACT

Wildlife habitat mapping is a widely used tool for supporting decision making in conservation. It requires data indicating wildlife habitat use to model and map habitat suitability. Collecting wildlife data, however, requires much effort, especially for species in remote mountainous regions of limited accessibility. Such circumstances often necessitate the integration of limited amounts of data available from multiple sources for habitat mapping. To that end, this study presents a framework for integrating multi-source wildlife data for habitat mapping. For evaluating the integration framework, a case study of mapping habitat suitability of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) by integrating sightings elicited from local volunteer villagers and obtained from official patrol records was conducted in Yunnan, China. The integration was explored at three levels: data-, knowledge- and model-level following disparate principles. The predicted habitat suitability maps were validated against monkey occurrence data independently collected though field-tracking. Results show the suitability maps predicted based on data integration were more accurate compared to maps predicted based on individual data sources. Data- and model-level integration achieved higher accuracy compared to knowledge-level integration. Further, data- and model-level integration following a conservative principle, i.e., the 'minimum' operator, led to higher mapping accuracy. The integration framework is generally applicable for integrating data from multiple sources for habitat mapping. It is also easy to implement and thus can be conveniently adopted by practitioners. Habitat suitability maps generated based on integrated data from multiple sources could better supporting decision making in biodiversity monitoring and conservation.

1. Introduction

Wildlife habitat mapping, also referred to as species distribution modeling (SDM), is widely used to support decision making in conservation (e.g., biological invasions management, habitats protection, reserve selection, re-introduction, etc.) (Guisan et al., 2013). From a theoretical perspective, SDM is also a useful tool for understanding the multiscale biological dynamics as it could reveal the effects of environmental heterogeneities over space, time and biological scales of organization on species specific patterns and macroecological patterns (Franklin and Miller, 2009; Levin, 1992). Both environmental data characterizing the environmental conditions and wildlife data indicating habitat use are required for habitat mapping. Relatively abundant environmental data are increasingly available due to the rapid development of geospatial technologies such as remote sensing (Kerr and Ostrovsky, 2003; van Zyl, 2001; Viña et al., 2008). The

[*] Corresponding author.
E-mail address: guiming.zhang@du.edu (G. Zhang).

https://doi.org/10.1016/j.ecolind.2020.106735
Received 5 March 2020; Received in revised form 10 June 2020; Accepted 15 July 2020

collection of wildlife data, nevertheless, requires much effort. Techniques such as radio telemetry, infrared trapping cameras and GPS collars could be deployed to collect high-quality wildlife data (Burton et al., 2012; Campbell and Sussman, 1994; Hemson et al., 2005). However, these techniques may not work well in regions with highly variable terrains and are excessively expensive for conservation programs with limited budgets. The high cost often renders them unsuitable for conservation programs in poor and remote mountainous regions (Danielsen et al., 2003).

Researchers and practitioners have been exploring cost-effective alternatives for wildlife data collection. Examples are local ecological knowledge (Anadón et al., 2009) and ranger-based monitoring programs (Critchlow et al., 2016). On one hand, local villagers in remote rural areas whose livelihoods are closely linked to ecosystem services are valuable information sources for obtaining wildlife data. Substance farmers, shepherds and hunters have spent a great deal of time in the field and encountered wildlife in their natural habitats. They have accumulated rich local ecological knowledge about wildlife occurrences in their local areas (Anadón et al., 2009). Wildlife sightings, therefore, can be elicited from local volunteer villagers for habitat mapping (Zhang et al., 2018a; Zhu et al., 2015). On the other hand, many protected area administrations (e.g., nature reserves, parks) have set up routine ranger-based patrols to monitor wildlife populations, illegal poaching and deforestation and wildlife encounters are recorded during patrols (Burton, 2010). The official patrol records can also provide wildlife data for habitat mapping (Zhang et al., 2018b).

Nonetheless, wildlife data from the above-mentioned sources may be of limited amount and subject to certain data quality issues, which have adverse effects on the accuracy of habitat mapping. For example, sightings elicited from villagers are likely to be spatially biased due to the non-random and non-systematic observation effort (Zhu et al., 2015). Villagers do not intentionally track the wildlife; Instead, they opportunistically encounter the wildlife en route to other activities such as farming and pasturing. Sightings elicited from the villagers can also suffer positional uncertainty as villagers may not be able to locate the exact location of wildlife occurrence because of their blurry memories or incompetence of pinpointing the occurrence location due to a lack of map-reading skills (Zhu et al., 2015).

Patrol data are less susceptible to spatial bias as the patrol routes often cover the whole area with approximately even patrolling effort. Patrol records are also accurately georeferenced, for example, using a GPS (global positioning system) receiver. However, sightings in patrol records are still prone to positional errors (Zhang et al., 2018b). A location recorded by the patrol is where the patrol stands when sighting the wildlife; It is not the actual occurrence location of the wildlife. Without information regarding the distance and direction between the recorded location and the sighted wildlife, it is difficult to recover the actual location of the wildlife.

Even though data from multiple sources each has their own limitations and may be of limited amount, they may be the only available data that can be used for wildlife habitat mapping to support conservation decision making in real-world scenarios. Moreover, combining data from different sources may increase the amount of data and overcome their respective limitations and thus could improve habitat mapping accuracy. This study aims to develop a general framework for integrating multi-source wildlife data for habitat mapping and to evaluate its effectiveness.

Fletcher et al. (2019) reviewed and identified five typical ways for combining multiple sources of data (e.g., data from citizen science projects, atlas, museums, planned surveys, etc.) for modeling species distribution. First, simply pooling species data (e.g., occurrence locations) is the most commonly used method as it can increase sample size for modeling. Second, data from individual sources are used to develop independent models and the models are then combined in some way. Third, secondary data (e.g., range maps) are used to inform modeling the species of interest (e.g., to guide background point selection). Fourth, one source of data is used to provide a prior distribution for model parameters when modeling the second data. Lastly, different types of distribution data (e.g., coarse-grain atlas data, fine-grain data, ad hoc presence-only data, planned survey data, etc.) are formally combined under the inhomogeneous point process framework.

This study presents a framework integrating multi-source data for habitat mapping at three different levels: data-, knowledge- and model-level. For data-level integration, sightings from different sources were pooled for habitat suitability modeling. For knowledge-level integration, knowledge regarding the relationships between species habitat suitability and environmental gradients was discovered from individual data sources and then synthesized to build a model. For model-level integration, independent models were built using individual data sources and the models were then combined. Knowledge-level integration is a novel means of combining multi-source data in addition to those identified by Fletcher et al. (2019). Further, knowledge- and model-level data integration in this study were tested under disparate principles ranging from conservative to liberal. Results and findings in this regarding could make new contributions to the existing literature. The proposed integration framework is simple, easy to implement and generally applicable compared to other alternatives as identified in Fletcher et al. (2019) (e.g., secondary data, prior distribution, inhomogeneous point process) and thus it is more likely to be adopted by conservation practitioners (more discussion in Section 4.4).

To evaluate the data integration framework, a case study of mapping habitat suitability of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) by integrating sightings elicited from voluntary local villagers and obtained from official patrol records was conducted at Mt. Lasha in northwestern Yunnan, China. *R. bieti* is an endangered species of historical and cultural significance to communities in the mountainous regions of southwest China (Long et al., 1994). Data availability for *R. bieti* distribution is very limited and therefore integrating multi-source data for habitat mapping is urgently needed for its conservation. Separately, the two data sources have been used for *R. bieti* habitat mapping but the focus was on devising geospatial analysis methods to correct for spatial bias in sightings data elicited from villagers (Zhu et al., 2015) and to reduce positional errors in patrol data (Zhang et al., 2018b). This study makes new contributions by exploring whether integrating the two data sources for habitat mapping would overcome their respective data quality limitations and thus improve mapping accuracy.

2. Materials and methods

2.1. Study area

The study area (Fig. 1) is at Mt. Lasha (99°15′E, 26°20′N) in northwest Yunnan, China, near the southern-most part of the geographic range of *R. bieti*, an *Endangered* species on the International Union for Conservation of Nature Red List (IUCN 2016) endemic to the eastern Himalayas between the upper Mekong and Yangtze Rivers (Long et al., 1994; Xiao et al., 2003). The study area (~20 km^2) is home for a group of approximately 100 *R. bieti* individuals (Huang et al., 2012) and it became part of the Yunling Provincial Nature Reserve since 2006. The elevation ranges from about 2500–4000 m. Ridgelines surrounding the area at the high elevations are largely deforested and mostly used as grazing land. Farmland and villages are at the low elevations in the east. The vegetation transitions from deciduous broad-leaved forest from lower elevations to dark conifer forest to higher elevations with mixed deciduous-conifer forest in between (Huang et al., 2017, 2012; Huang, 2009).

2.2. R. bieti occurrence data

2.2.1. Local ecological knowledge

Sightings of *R. bieti* were elicited from local villagers through

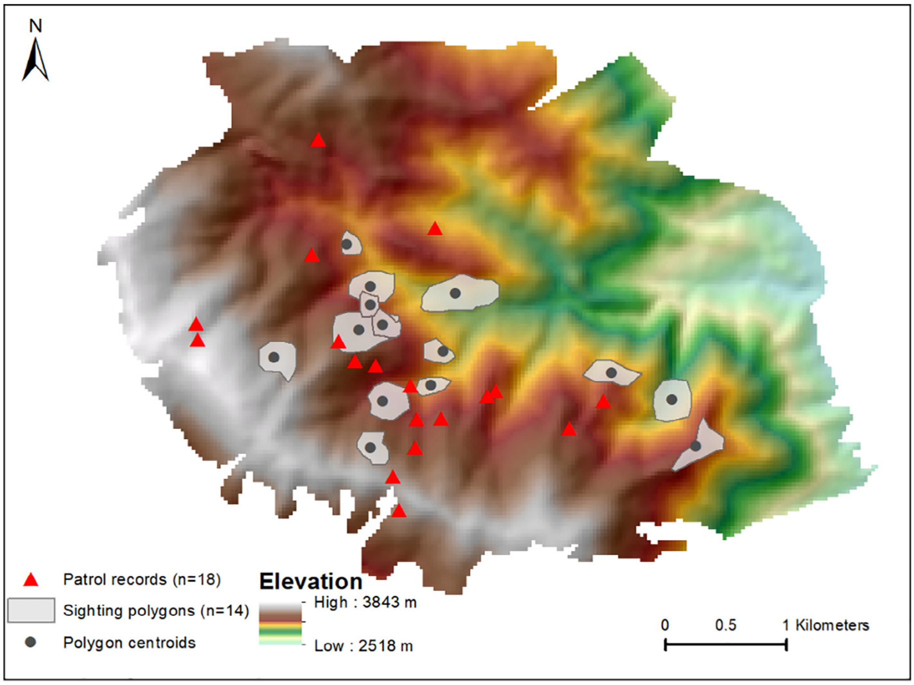

Fig. 1. *R. bieti* sighting polygons elicited from local villagers and sightings obtained from patrol records (summer months of 2008 and 2009).

structured interview sessions conducted in Jul. and Aug. 2010. Structured interview is a standard protocol and best practice for eliciting wildlife sighting data from residents (Anadón et al., 2009; Zhu et al., 2015). The interviews were carried out by a field biologist using geovisualization software that integrates high-resolution digital elevation model (DEM) and satellite imagery to produce an intuitive 3-dimensional view of the study area for the villagers to locate wildlife sightings. Data were recorded as polygons indicating the approximate areas where monkeys were sighted. 70 local villagers in total from all the 5 nearby villages who had extensive experience in the field were interviewed. More details regarding the data elicitation processes can be found in Zhang et al. (2018a) and Zhu et al. (2015).

Sightings in the summer months (Jun., Jul., Aug.) of 2008 and 2009 were extracted and used in this study (14 sightings in total) given that *R. bieti* exhibits seasonality in habitat use (Huang, 2009) and local villagers were more active in the field in summer months and therefore provided more data. Moreover, sightings in the two recent years were assumed to be more accurate as the memories of the villagers were still fresh. Also, the two-year period matches timeframe of the validation data (Section 2.6). Centroids of the sighting polygons were used as approximate monkey occurrence locations for habitat mapping in this study (Fig. 1). Centroids are the simplest reasonable approximation of species occurrence locations as villagers delineated polygons around sites where *R. bieti* occurred.

2.2.2. *Patrol records*

In the study area, one forest ranger was employed and trained by the Administration of Yunling Provincial Nature Reserve to conduct regular patrols 5 days per month on a series of main routes and secondary routes covering the whole area. Wildlife encounters during the patrols were recorded in patrol forms. A hand-held GPS receiver was used to read the latitude and longitude of the location at which he sighted wildlife. The patrol date, geographic coordinates, species name, number of wildlife encountered, behaviors of the wildlife, and habitat type were recorded in the form. More details regarding the patrol program can be found in Zhang et al. (2018b).

The patrol records were obtained from the Administration of Yunling Provincial Nature Reserve and recorded sightings of *R. bieti* during summer months of 2008 and 2009 were extracted and used in this study (18 records in total) (Fig. 1).

2.3. *Environmental data*

Based on existing knowledge of the ecology of the species (Huang et al., 2017, 2012; Huang, 2009), elevation, tangent of slope, aspect category (0–360° discretized into 8 equal 45° categories), least-cost distance to rivers (tangent of slope as cost), least-cost distance to villages (tangent of slope as cost), and plant type (10 types) were used as environmental covariates for modeling and mapping habitat suitability for *R. bieti* in this study. These covariates represent the environmental factors influencing the habitat use of *R. bieti* in the study area (e.g., terrain condition, water source, shelter or food, and human-posed disturbance) and have been used in previous studies (Zhang et al., 2018a,b; Zhu et al., 2015). The covariates were at 30-m spatial resolution.

2.4. *Habitat suitability mapping*

Two habitat mapping methods based on species presence-only data were adopted for habitat suitability mapping in this study: the rule-based mapping method (Zhang et al., 2018c) which allows knowledge-level data integration and the most widely used Maxent (Phillips et al., 2006). Both methods model species distribution or habitat suitability as a function of environmental variables considering probability distributions.

2.4.1. *Rule-based mapping*

In essence, the rule-based mapping method (Zhang et al., 2018c) models species habitat suitability – environment relationship based on the probability distribution of species occurrences over environmental gradients. This method consists of two major steps, as described below. Interested readers are referred to Zhang et al. (2018c) for full details.

In the first step, probability distribution of species occurrences over a covariate is estimated based on the covariate values at the occurrence locations using kernel density estimation (KDE), a non-parametric method for estimating continuous probability distribution from discreate sample data values (Silverman, 1986):

$$f_{occurrence}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_x}K\left(\frac{x-x_i}{h_x}\right) \tag{1}$$

in which $f_{occurrence}(x)$ is the estimated probability distribution of species occurrence with respect to environmental factor x, x_i is the value of x at occurrence location i, n is the total number of occurrence locations. K is a kernel function for which the Gaussian kernel was adopted (Silverman, 1986). h_x is the bandwidth for x, which was determined following the maximum likelihood principle based on cross-validation (Brunsdon, 1995). This presence distribution $f_{presence}(x)$ reflects species habitat use regarding covariate x.

Similarly, the probability distribution of the covariate is also estimated based on the covariate values in the whole mapping area using KDE. This background distribution $f_{background}(x)$ reflects resource availability with respect to x. The ratio of the presence distribution and the background distribution thus indicates species habitat preferences and therefore habitat suitability regarding covariate x is modeled using Eqs. (2) and (3) below:

$$S(x) = \frac{1}{1+e^{1-f_{ratio}(x)}} \tag{2}$$

where:

$$f_{ratio}(x) = \frac{f_{occurrence}(x)}{f_{background}(x)} \tag{3}$$

Species habitat suitability regarding each of the covariates is modeled in this way (For continuous covariates KDE is used to estimate probability distributions, whereas for categorical covariates relative frequency distributions are used). For any location (e.g., raster cell) in the mapping area, a vector of suitability values regarding individual covariates can be computed based on values of the covariates at that location.

At the second step, suitability values regarding individual covariates are synthesized to compute the overall suitability considering all covariates. Here a simple arithmetic mean is adopted (Eq. (4)) as previous studies has shown it results in better model performance compared to other alternatives (Zhang et al., 2018c):

$$S(x^1, x^2, x^j, \cdots, x^m) = \frac{1}{m}\sum_{j=1}^{m} S(x^j) \tag{4}$$

where $S(x^1, x^2, x^j, \ldots, x^m)$ is the overall suitability, x^j is the j^{th} environmental factor, m is the total number of covariates involved, and $S(x^j)$ is the habitat suitability with respect to x^j.

2.4.2. Maxent

Maxent is the most widely used SDM method that requires species presence-only data (Elith et al., 2011; Phillips et al., 2006). It estimates species distribution over geographic space as the probability distribution that has the maximum entropy (i.e., closest to a uniform distribution) while respecting constraints implied in the environmental conditions at species occurrence localities (Phillips et al., 2006). The most recent Maxent software (version 3.4.0) (Phillips et al., 2020) was downloaded and used in this study. Its default model parameter settings, which were fine-tuned based on a large dataset and thus are supposed to be generally applicable (Phillips and Dudík 2008), were used in this study (e.g., auto features, cloglog output format, add samples to background, remove duplicate presence records, etc.). The Cloglog output from Maxent can be interpreted as a species habitat suitability map.

2.5. Integrating multi-source data for habitat mapping

The integration of *R. bieti* sightings from local ecological knowledge and patrol records for habitat mapping was explored at three levels: data-, knowledge- and model-level (Fig. 2). These three levels of integration correspond to the three stages of habitat suitability mapping, namely, input, model, and output, respectively. Data-level integration creates a new set of integrated species occurrence data as input to habitat suitability modeling and mapping methods. For data-level integration, occurrences from the two sources were simply pooled together for modeling and mapping habitat suitability using the rule-based method and Maxent.

Knowledge-level integration creates a new model encoding the integrated knowledge regarding species habitat suitability – environment relationships synthesized from the knowledge embedded in individual data sources. Knowledge-level integration was examined using only the rule-based method, as the Maxent software provides no such flexibility. Occurrences from individual sources were used to estimate presence distribution and derive the relationship between habitat suitability and environmental covariates (knowledge) (Eq. (2)). The suitability – environment relationships derived from the two data sources regarding the same covariate were then synthesized using the ‘minimum’, ‘mean’ or ‘maximum’ operator (Fig. 3). The ‘minimum’ operator implies a conservative view in data integration as the synthesized relationship is very stringent. The suitability value under a given environmental condition is determined by the lowest among the suitability values under that environmental condition derived from individual data sources. On the other end, the ‘maximum’ operator represents an optimistic view as the synthesized relationship tends to be overly tolerant. The suitability value under the environmental condition is determined by the highest among the individual suitability values. The ‘mean’ operator stands for a middle ground by assigning the suitability value as the average of the individual suitability values.

Model-level integration creates a new output habitat suitability map by integrating suitability maps predicted from models built based on individual data sources. Model-level integration was tested using the rule-based mapping method and Maxent. Occurrences from each source were used to build a model and predict a suitability map. The two suitability maps were then integrated through a pixel-wise ‘minimum’, ‘mean’ or ‘maximum’ operator to produce a final suitability map. The implication of the three operators are the similar to those discussed above in knowledge-level integration.

2.6. Accuracy assessment

R. bieti occurrence locations recorded during field tracking were used as independent validation data to evaluate the accuracy of the predicted habitat suitability maps. Tracking was conducted by one field biologist and two assistants primarily for behavioral study purposes in 2008 and 2009 (Huang et al., 2012). Location of the monkeys was recorded on a topographic map every 30 min from 7 am to 8 pm. These field tracking locations were the most accurate data available reflecting the distribution of *R. bieti* in the study area during the study period (2008 to 2009). Recorded monkey occurrence locations in summer months of 2008 and 2009 (Fig. 4) along with 1000 background locations (i.e., pseudo-absences) randomly chosen from the study area were used as validation data.

The area under the curve (AUC) was adopted as an accuracy measure of the predicted suitability maps. AUC can be computed based on the occurrence locations and background locations (Phillips and Dudík, 2008). AUC ranges from 0.5 to 1.0, with 0.5 indicating that the predictions are no better than random predictions and 1.0 indicating perfect predictions. Models with AUC values greater than 0.75 are considered potentially useful (Elith et al., 2002; Phillips and Dudík, 2008). AUC provides a single accuracy measure that is independent of any choice of suitability threshold. It has been widely used for evaluating performance of species distribution models and habitat suitability models (Elith et al., 2011; Phillips et al., 2006; Phillips and Dudík, 2008; Zhang, 2019; Zhang et al., 2018c, 2018b; Zhang and Zhu, 2019).

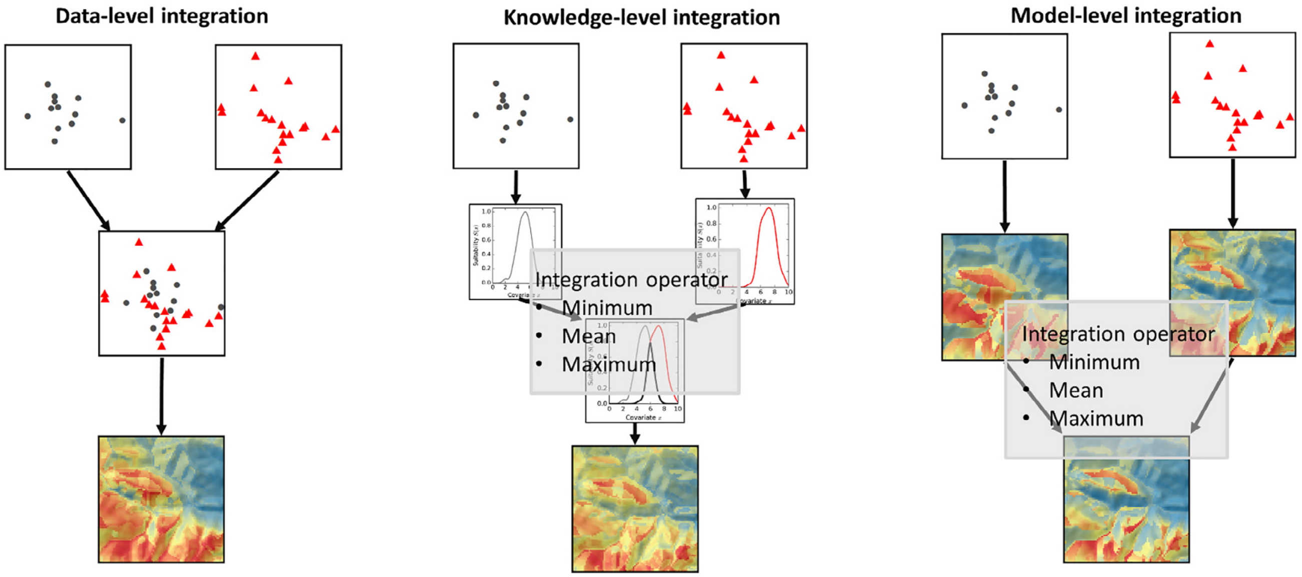

Fig. 2. Integration of multi-source species data for habitat mapping at three different levels.

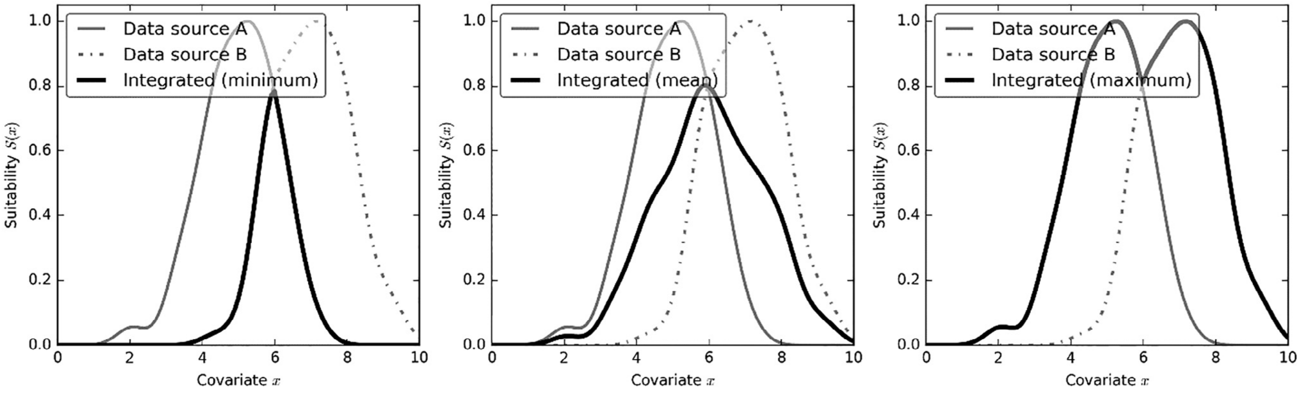

Fig. 3. A schematic example of knowledge-level data integration for habitat suitability mapping.

2.7. *Experiment design*

Accuracies of suitability maps predicted based on data integration were compared to accuracies of suitability maps predicted based on individual data sources. This allows examining whether multi-source data integration can improve mapping (prediction) accuracy. Moreover, accuracies of suitability maps based on data integration at different levels (data-, knowledge or model-level) were compared to identify the integration level achieving the highest mapping accuracy. Finally, accuracies of suitability maps based on knowledge- or model-level integration using different operators ('minimum', 'mean' or 'maximum') were compared to examine the effects of the operators.

3. Results

3.1. *Mapping based on individual data sources*

The accuracies of habitat suitability maps predicted based on individual data sources (Fig. 5) were shown in Table 1. The spatial patterns of the suitability maps predicted from different methods based on the same data source were similar. Overall, high suitability areas predicted from local ecological knowledge (LEK) were more geographically constrained than those predicted from patrol records (PAT). Suitability maps predicted from PAT were generally of higher accuracy than those predicted from LEK, which indicates data from PAT were of better quality than data from LEK. Nevertheless, all AUC values were below 0.75. The unsatisfactory mapping accuracies may be attributed to the limitations of individual data sources, i.e., spatial bias in LEK sightings (Zhu et al., 2015) and positional errors in PAT records (Zhang et al., 2018b).

3.2. *Data-level integration*

Data-level integration (i.e., pooling occurrences from LEK and PAT) resulted in habitat suitability maps (Fig. 6) that were more accurate than those predicted based on individual data sources (Table 2). Suitability maps predicted using the two mapping methods had similar spatial patterns, although Maxent (AUC = 0.778) achieved a slightly higher mapping accuracy than the rule-based mapping method (AUC = 0.761). The AUC values were all above 0.75, suggesting that simply pooling occurrences from the two data sources can overcome the limitations of individual data sources and thus result in potentially useful predictions.

3.3. *Knowledge-level integration*

Knowledge-level integration of LEK and PAT using the 'minimum' or 'mean' operator improved mapping accuracy (Fig. 7; Table 3) compared to mapping based on individual data sources (Table 1). Integrating knowledge from the two sources with the 'minimum' operator achieved the highest accuracy (AUC = 0.751), followed by the 'mean' operator (AUC = 0.747). Using the 'maximum' operator at knowledge-level

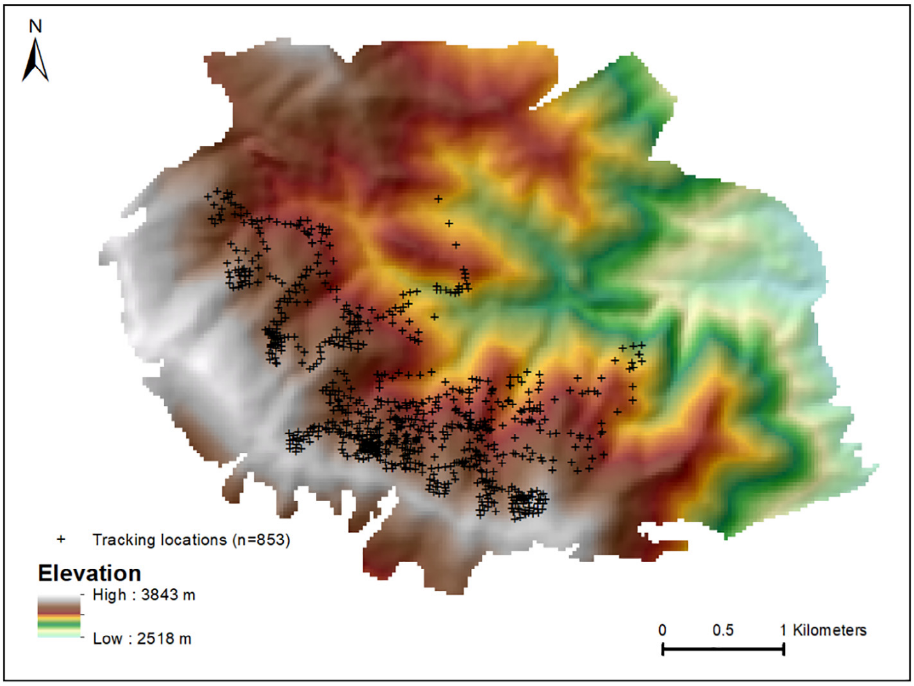

Fig. 4. *R. bieti* occurrence locations recorded during field-tracking of the monkeys (summer months of 2008 and 2009).

integration (AUC = 0.711) did not improve mapping accuracy compared to mapping using knowledge only from PAT.

3.4. Model-level integration

Model-level integration of LEK and PAT using the 'minimum' or 'mean' operator increased mapping accuracy (Fig. 8; Table 4) compared to mapping based on individual data sources (Table 1). Integrating models (i.e., suitability maps) built from the two sources with the 'minimum' operator achieved higher accuracy (AUC = 0.760 and AUC = 0.771 for rule-based mapping and Maxent, respectively) than using the 'mean' operator (AUC = 0.748 and AUC = 0.755 for the two mapping methods, respectively). Using the 'maximum' operator at model-level integration (AUC = 0.699 and AUC = 0.713 for the two

Table 1
Accuracy (AUC) of the habitat suitability maps predicted based on individual data sources.

	LEK	PAT
Rule-based mapping	0.702	0.711
Maxent	0.677	0.730

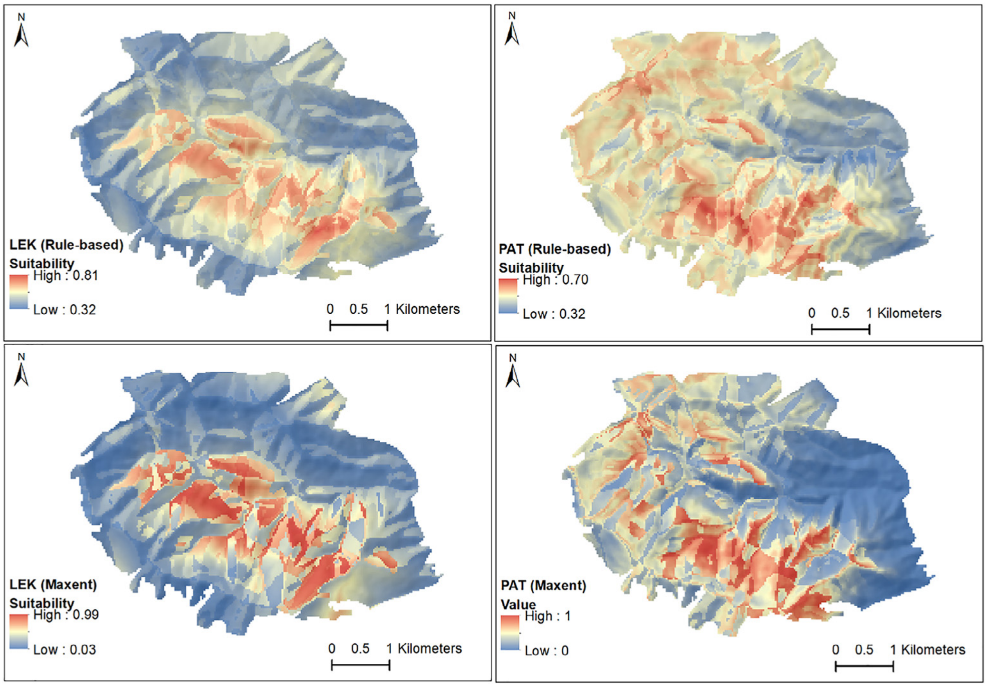

Fig. 5. Suitability maps predicted based on individual data sources (LEK: local ecological knowledge; PAT: patrol records).

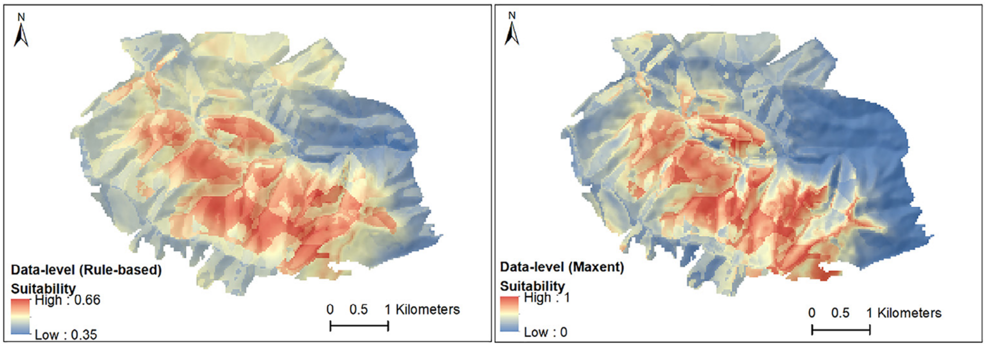

Fig. 6. Suitability maps predicted based on data-level integration.

Table 2
Accuracy (AUC) of the habitat suitability maps predicted based on data-level integration.

	Rule-based mapping	Maxent
AUC	0.761	0.778

Table 3
Accuracy (AUC) of the habitat suitability maps predicted based on knowledge-level integration with the three operators (rule-based mapping method).

	'Minimum' operator	'Mean' operator	'Maximum' operator
AUC	0.751	0.747	0.711

mapping methods, respectively) did not achieve clear accuracy improvement over mapping based on individual data sources.

4. Discussion

4.1. Effectiveness of data integration

Compared to habitat suitability mapping based on individual data sources, integrating data from LEK and PAT at all three levels could effectively improve mapping accuracy. Notably, the highest AUC values of suitability maps predicted based on data integration (using the 'minimum' operator) were all above the 0.75 threshold. That is, mapping by integrating multi-source data can result in potentially useful habitat suitability maps (AUC $>$ 0.75) (Elith et al., 2002; Phillips and Dudík, 2008) even when mapping based on individual data falls short. The improved mapping accuracy indicates that integrating multi-source data can overcome the limitations of individual data sources.

4.2. Impact of integration level

In this study, data-level and model-level integration generally produced more accurate suitability maps compared to knowledge-level integration. *R. bieti* sightings from LEK overall were in slightly steeper areas compared to sightings from PAT (Fig. 9). Data-level integration (i.e., pooling occurrences from LEK and PAT to estimate occurrence probability distribution) can effectively expand the spatial coverage of the occurrence data, increases sample size for modeling and mapping and therefore improves mapping accuracy. Model-level integration produced a more accurate suitability map by combining the two suitability maps predicted from individual data sources. It is a simple form of model averaging (Dormann et al., 2018). Knowledge-level integration synthesizes knowledge on the relationship between *R. bieti* habitat suitability and environment covariates derived from individual data sources to form an integrated model (Fig. 10).

4.3. Impact of integration operator

The integration operators imply three different philosophies in data integration (Fig. 10). The 'minimum' operator is very conservative and stringent. For example, a high suitability is assigned for a given environmental condition only if this assignment is supported by both data sources. The 'maximum' operator, however, tends to be liberal and tolerant. A high suitability is assigned for the given environmental condition even if this assignment is supported by only one data source. The 'mean' operator is moderate and assigns the given environmental condition a suitability value that is the average of the suitability assignments from the two sources. In this study, integrating data using the 'minimum' operator (at the knowledge- or model-level) produced more accurate suitability maps than using the 'mean' or 'maximum' operator, suggesting the conservative standpoint is preferred for multi-source data integration for habitat mapping.

4.4. Applicability of the integration framework

Integrating multi-source data for wildlife habitat suitability mapping at the data- and model-level level is easy to implement. Occurrence data pooling and model averaging can be done without tweaking the modeling method in use (i.e., the modeling method can be treated as a

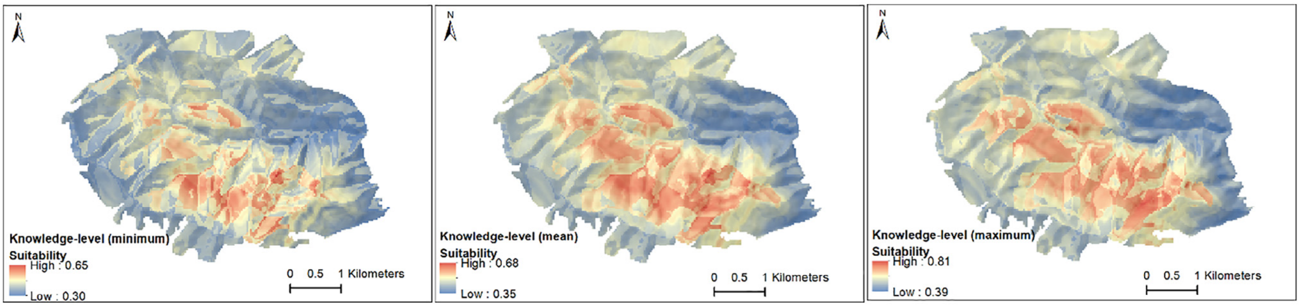

Fig. 7. Suitability maps predicted based on knowledge-level integration (rule-based mapping method).

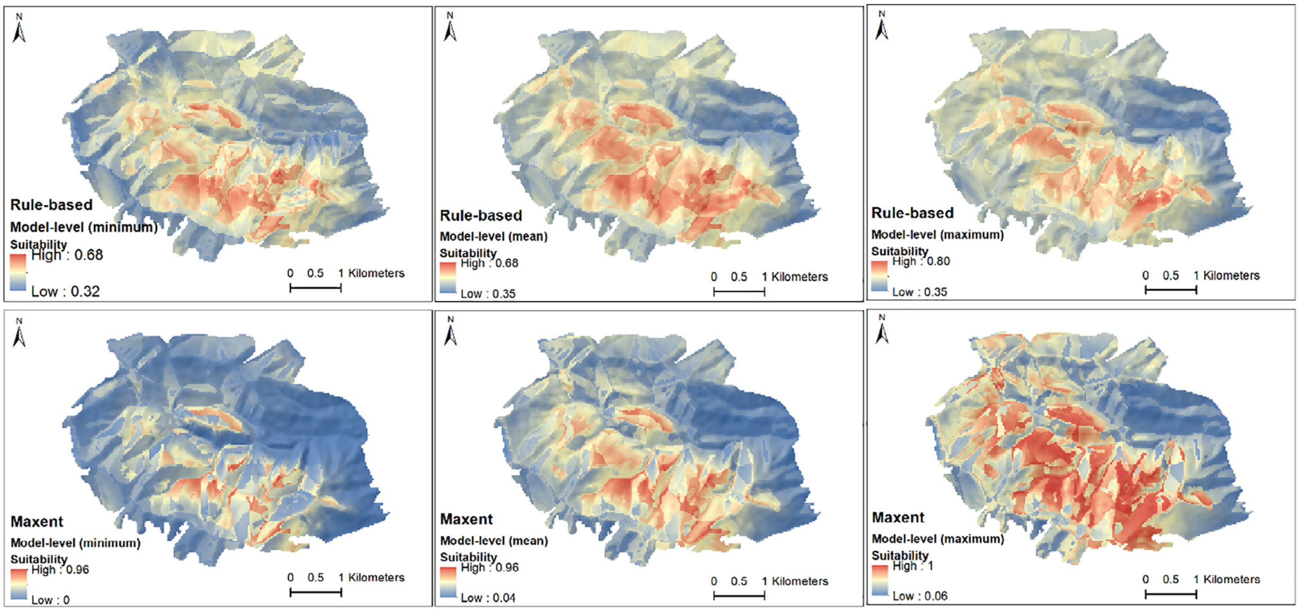

Fig. 8. Suitability maps predicted based on model-level integration.

Table 4
Accuracy (AUC) of the habitat suitability maps predicted based on model-level integration with the three operators.

	'Minimum' operator	'Mean' operator	'Maximum' operator
Rule-based	0.760	0.748	0.699
Maxent	0.771	0.755	0.713

'black box'). For data-level integration, multi-source occurrence data can be simply pooled and provided as input to the modeling method to produce an output suitability map. For model-level integration, intermediate suitability maps produced based on individual data sources using the modeling method can be synthesized following certain integration principles (e.g., pixel-wise minimum, mean, or maximum) to produce a final suitability map. All these procedures are straightforward and easy to implement.

Knowledge-level integration, on the other end, requires understanding of the mechanism of the modeling method such that proper procedures can be designed to implement knowledge-level integration. In this study, with the rule-based mapping method, knowledge regarding species habitat suitability – environment relationships derived from different sources were integrated by synthesizing the suitability – covariate response curves (Fig. 10). The Maxent software as is does not provide such flexibility. Fletcher et al. (2019) pointed out that one source of data can be used to provide a prior distribution for model parameters when modeling the second data. This can be taken as a form of knowledge-level integration (knowledge regarding model parameters). Implementing this idea is model-specific and requires detailed knowledge of the models.

Results of the *R. bieti* case study suggest that data integration produced more accuracy suitability maps than mapping based on individual data sources. It is supporting evidence that the proposed data integration framework is generally applicable for integrating multi-source data for improving habitat mapping accuracy and its applications for mapping habitats of other species are encouraged. More broadly, multi-source data integration for habitat mapping and species distribution modeling offers more opportunities for unveiling the dynamics of biodiversity, ecosystem states and multiple future trajectories under environmental variability, human impact and positive controls (Convertino et al., 2011). Species distribution modeling and predictions can also unveil the complexity and simplicity of the structure and function of biodiversity and thus guide optimal monitoring and multi-scale ecological engineering interventions that aim to protect multiple co-dependent species at the same time (Convertino et al., 2015).

5. Analysis of covariate contributions

The environmental covariates representing diverse drivers of species distribution used for mapping *R. bieti* habitat suitability may be of varied importance in suitability modeling. The Maxent software provides an analysis of covariate contributions (Phillips et al., 2006). According to estimates of relative contributions of the environmental variables to the Maxent model (Table 5), distance to village or road and elevation seem to be the most important covariates to the model trained using occurrence locations pooled from LEK and PAT whilst slope and distance to river seem to be the least important. Based on the results of

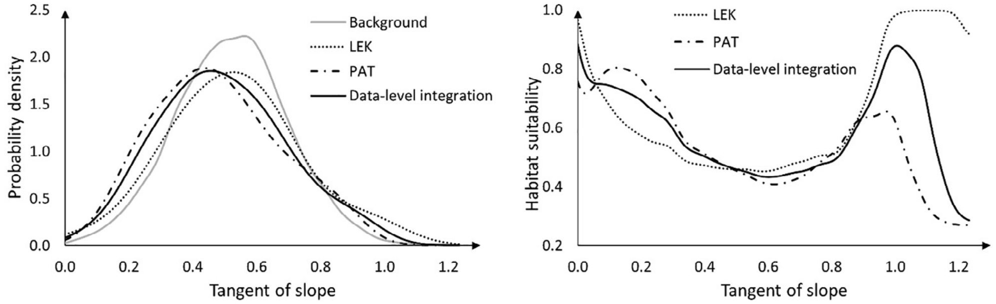

Fig. 9. Data-level integration in the rule-based mapping method illustrated with the slope covariate. Left figure shows the background distribution and occurrence probability distributions estimated from LEK occurrences, PAT occurrences, and the pooled LEK and PAT occurrences (data-level integration). Right figure shows the relationship between habitat suitability and slope derived from occurrences from individual data sources and the pooled occurrences.

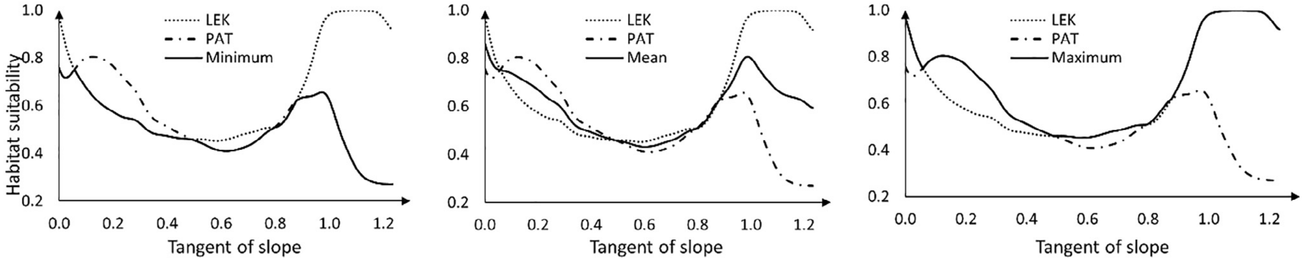

Fig. 10. Knowledge-level integration in the rule-based mapping method illustrated with the slope covariate. Figures show the relationship between habitat suitability and slope synthesized using the 'minimum', 'mean' or 'maximum' operator.

Table 5
Estimates of relative contributions of the environmental variables to the Maxent model trained using occurrence locations pooled from LEK and PAT.

Covariate	Percent contribution	Permutation importance
dist2vilroad	34.3	30.5
elevation	30.6	32.8
planttype	19.3	15.9
aspect	10.3	8.6
slope	3.8	6. 9
dist2river	1.7	5.2

the jackknife test of variable importance (Fig. 11), the covariate with highest regularized training gain when used in isolation is plant type, which therefore appears to have the most useful information by itself. The covariate that decreases the gain the most when it is omitted is distance to village or road, which therefore appears to have the most information that is not present in the other covariates.

The covariates were chosen primarily based on existing knowledge of the ecology of the species (Huang et al., 2017, 2012; Huang, 2009); Data-driven covariate selection was not appropriate for this study given the relatively small amount of species occurrence data. It was not pursued neither in this study how different combinations of the covariates would affect the results of data integration for habitat mapping. Instead, the same set of covariates were used in habitat mapping experiments throughout in this study; This control allows comparing the effects of the input species occurrence data. Admittedly, the analysis of covariate contributions and importance is only one aspect of the more general global sensitivity and uncertainty analysis (Convertino et al., 2014), which itself is an interesting topic but is out of the scope of this study.

6. Conclusion

This study presents a framework for integrating multi-source wildlife data for habitat mapping at three levels: data-, knowledge- and model-level. To evaluate the effectiveness of the framework, a case study of mapping habitat suitability of the black-and-white snub-nosed monkey (*Rhinopithecus bieti*) by integrating sightings elicited from local villagers and obtained from official patrol records was conducted in Yunnan, China. Results suggest that data integration can effectively improve mapping accuracy over mapping based on individual data sources. Data- and model-level integration generally achieved higher accuracy. Knowledge- and model-level integration using the 'minimum' operator (a conservative principle) resulted in higher mapping accuracy, compared to using the 'maximum' or 'mean' operator. This data integration framework is generally applicable for integrating data from two or more sources for habitat mapping. It is also easy to implement and thus can be conveniently adopted by practitioners. Habitat suitability maps generated based on integrated species data from multiple sources could better support decision making in biodiversity monitoring and conservation.

One limitation of this study is the integration framework was tested for mapping habitat suitability of one species in a smaller study area as a proof of concept. The authors invite interested parties to apply the integration framework for habitat mapping of other species in other geographic areas and to more comprehensively evaluate its general applicability and effectiveness for integrating multi-source data for wildlife habitat mapping or species distribution modeling.

CRediT authorship contribution statement

Guiming Zhang: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **A-Xing Zhu:** Conceptualization, Methodology, Writing - review & editing. **Yu-Chao He:** Conceptualization, Writing - review & editing. **Zhi-Pang Huang:** Conceptualization, Writing - review & editing. **Guo-Peng Ren:** Conceptualization, Writing - review & editing. **Wen Xiao:** Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

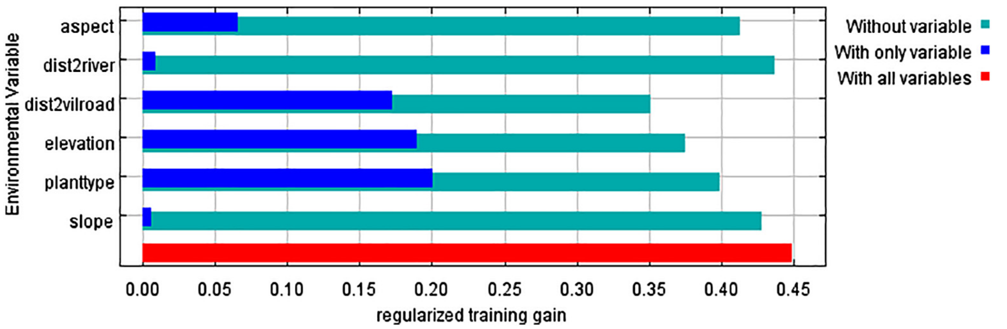

Fig. 11. Contributions of environmental covariates to regularized training gain in Maxent modeling training.

Acknowledgements

This study was supported by the Internationalization Grant and Faculty Research Fund at the University of Denver and the National Natural Science Foundation of China (Grant No. 41871300, 31860164, 31860168). Supports to A-Xing Zhu through the Vilas Associate Award, the Hammel Faculty Fellow Award, and the Manasse Chair Professorship from the University of Wisconsin-Madison are greatly appreciated. The Provincial Innovation Team of Biodiversity Conservation and Utility of the Three Parallel Rivers Region at Dali University was supported by the Second Tibetan Plateau Scientific Expedition and Research Program (Grant No. 2019QZKK0402).

References

Anadón, J.D., Giménez, A., Ballestar, R., Pérez, I., 2009. Evaluation of local ecological knowledge as a method for collecting extensive data on animal abundance. Conserv. Biol. 23, 617–625.

Brunsdon, C., 1995. Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. Comput. Geosci. 21, 877–894. https://doi.org/10.1016/0098-3004(95)00020-9.

Burton, A.C., 2010. Wildlife Monitoring and Conservation in a West African Protected Area. University of California, Berkeley, Berkeley, Environ. Sci. Policy Manag.

Burton, A.C., Sam, M.K., Balangtaa, C., Brashares, J.S., 2012. Hierarchical multi-species modeling of carnivore responses to hunting, habitat and prey in a West African protected area. PLoS One 7, e38007.

Campbell, A.F., Sussman, R.W., 1994. The value of radio tracking in the study of neotropical rain forest monkeys. Am. J. Primatol. 32, 291–301.

Convertino, M., Kiker, G.A., Muñoz-Carpena, R., Chu-Agor, M.L., Fischer, R.A., Linkov, I., 2011. Scale- and resolution-invariance of suitable geographic range for shorebird metapopulations. Ecol. Complex. 8, 364–376. https://doi.org/10.1016/j.ecocom.2011.07.007.

Convertino, M., Muñoz-Carpena, R., Chu-Agor, M.L., Kiker, G.A., Linkov, I., 2014. Untangling drivers of species distributions: global sensitivity and uncertainty analyses of MaxEnt. Environ. Model. Softw. 51, 296–309. https://doi.org/10.1016/j.envsoft.2013.10.001.

Convertino, M., Muñoz-Carpena, R., Kiker, G.A., Perz, S.G., 2015. Design of optimal ecosystem monitoring networks: hotspot detection and biodiversity patterns. Stoch. Environ. Res. Risk Assess. 29, 1085–1101. https://doi.org/10.1007/s00477-014-0999-8.

Critchlow, R., Plumptre, A.J., Alidria, B., Nsubuga, M., Driciru, M., Rwetsiba, A., Wanyama, F., Beale, C.M., 2016. Improving law-enforcement effectiveness and efficiency in protected areas using ranger-collected monitoring data. Conserv. Lett. 1–9.

Danielsen, F., Mendoza, M.M., Alviola, P., Balete, D.S., Enghoff, M., Poulsen, M.K., Jensen, A.E., 2003. Biodiversity monitoring in developing countries: what are we trying to achieve? Oryx 37, 407–409.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wüest, R.O., Hartig, F., 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol. Monogr. 88, 485–504. https://doi.org/10.1002/ecm.1309.

Elith, J., Burgman, M.A., Regan, H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecol. Modell. 157, 313–329. https://doi.org/10.1016/s0304-3800(02)00202-8.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x.

Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100, 1–15. https://doi.org/10.1002/ecy.2710.

Franklin, J., Miller, J.A., 2009. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., Mcdonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16, 1424–1435. https://doi.org/10.1111/ele.12189.

Hemson, G., Johnson, P., South, A., Kenward, R., Ripley, R., Macdonald, D., Mcdonald, D., 2005. Are kernels the mustard? Data from global positioning system (GPS) collars suggests problems for kernel home-range analyses with least-squares cross-validation. J. Anim. Ecol. 74, 455–463. https://doi.org/10.1111/j.1365-2656.2005.00944.x.

Huang, Z.-P., Scott, M.B., Li, Y.-P., Ren, G.-P., Xiang, Z.-F., Cui, L.-W., Xiao, W., 2017. Black-and-white snub-nosed monkey (Rhinopithecus bieti) feeding behavior in a degraded forest fragment: clues to a stressed population. Primates. https://doi.org/10.1007/s10329-017-0618-7.

Huang, Z.P., 2009. Foraging, reproduction and sleeping site selection of black-and-white snub-nosed monkey (Rhinopithecus bieti) at the southern range. Fac. Conserv. Biol. Southwest Forestry University, Kunming.

Huang, Z.P., Cui, L.W., Scott, M., Wang, S.J., Xiao, W., 2012. Seasonality of reproduction of wild black-and-white snub-nosed monkeys (Rhinopithecus bieti) at Mt. Lasha, Yunnan, China. Primates 53, 237–245. https://doi.org/10.1007/s10329-012-0305-7.

Kerr, J.T., Ostrovsky, M., 2003. From space to species: ecological applications for remote sensing. Trends Ecol. Evol. 18, 299–305. https://doi.org/10.1016/S0169-5347(03)00071-5.

Levin, S.A., 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology 73, 1943–1967.

Long, Y.C., Kirkpatrick, C.R., Zhong, T., Xiao, L., 1994. Report on the distribution, population, and ecology of the yunnan snub-nosed monkey (Rhinopithecus bieti). Primates 35, 241–250. https://doi.org/10.1007/bf02382060.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Modell. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026.

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography (Cop.) 31, 161–175.

Phillips, S.J., Dudík, M., Schapire, R.E., 2020. Maxent software for modeling species niches and distributions (Version 3.4.0) [WWW Document]. URL http://biodiversityinformatics.amnh.org/open_source/maxent/ (accessed 2.8.20).

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, UK.

van Zyl, J.J., 2001. The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. Acta Astronaut. 48, 559–565. https://doi.org/10.1016/S0094-5765(01)00020-0.

Viña, A., Bearer, S., Zhang, H., Ouyang, Z., Liu, J., 2008. Evaluating MODIS data for mapping wildlife habitat distribution. Remote Sens. Environ. 112, 2160–2169. https://doi.org/10.1016/j.rse.2007.09.012.

Xiao, W., Ding, W., Cui, L.W., Zhou, R.L., Zhao, Q.K., 2003. Habitat degradation of Rhinopithecus bieti in Yunnan, China. Int. J. Primatol. 24, 389–398. https://doi.org/10.1023/a:1023009518806.

Zhang, G., 2019. Enhancing VGI application semantics by accounting for spatial bias. Big Earth Data 3, 255–268. https://doi.org/10.1080/20964471.2019.1645995.

Zhang, G., Zhu, A.-X., 2019. A representativeness directed approach to spatial bias mitigation in VGI for predictive mapping. Int. J. Geogr. Inf. Sci. 33, 1873–1893. https://doi.org/10.1080/19475683.2018.1501607.

Zhang, G., Zhu, A.-X., Huang, Z.-P., Ren, G., Qin, C.-Z., Xiao, W., 2018a. Validity of historical volunteered geographic information: evaluating citizen data for mapping historical geographic phenomena. Trans. GIS 22, 149–164. https://doi.org/10.1111/tgis.12300.

Zhang, G., Zhu, A.-X., Huang, Z.-P., Xiao, W., 2018b. A heuristic-based approach to mitigating positional errors in patrol data for species distribution modeling. Trans. GIS 22, 202–216. https://doi.org/10.1111/tgis.12303.

Zhang, G., Zhu, A.-X., Windels, S.K., Qin, C.-Z., 2018c. Modelling species habitat suitability from presence-only data using kernel density estimation. Ecol. Indic. 93, 387–396. https://doi.org/https://doi.org/10.1016/j.ecolind.2018.04.002.

Zhu, A.-X., Zhang, G., Wang, W., Xiao, W., Huang, Z.-P., Dunzhu, G.-S., Ren, G., Qin, C.-Z., Yang, L., Pei, T., Yang, S., 2015. A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. Int. J. Geogr. Inf. Sci. 29, 1864–1886. https://doi.org/10.1080/13658816.2015.1058387.