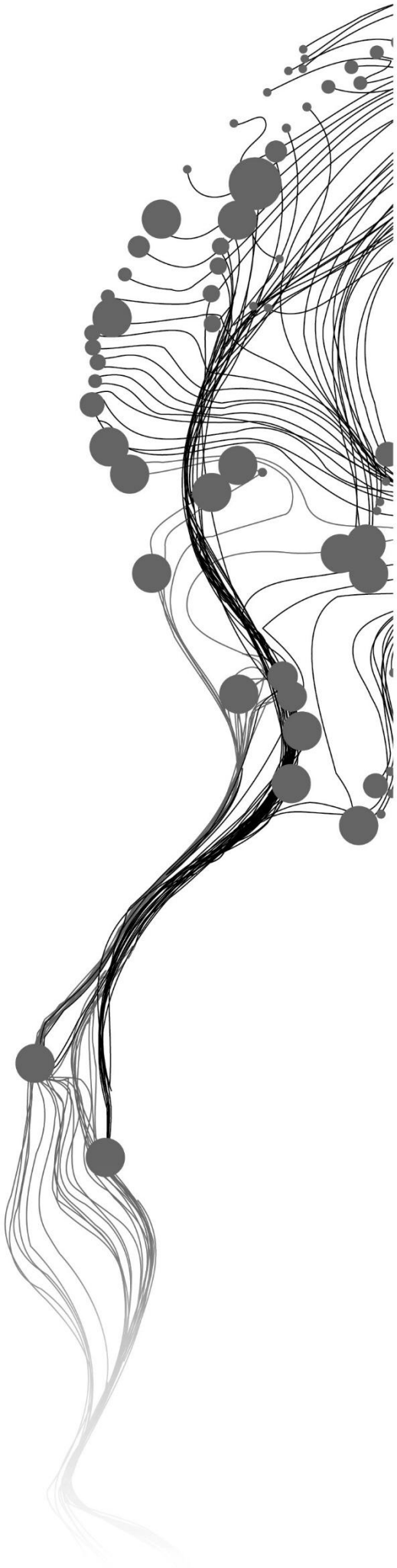
Modelling the giant panda habitat in China using MaxEnt: effects of sample size and extent of the study region

XUAN JIANG

March, 2015

SUPERVISORS:
Dr.Tiejun Wang (ITC, University of Twente)
Drs.E.H.Kloosterman (ITC, University of Twente)
ADVISOR:
Yiwen Sun (PhD candidate, ITC, University of Twente)



Modelling the giant panda habitat in China using MaxEnt: effects of sample size and extent of the study region

XUAN JIANG
Enschede, The Netherlands, March, 2015

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Natural Resource Management

SUPERVISORS:
Dr.Tiejun Wang (ITC, University of Twente)
Drs.E.H.Kloosterman (ITC, University of Twente)
ADVISOR:
Yiwen Sun (PhD candidate, ITC, University of Twente)

THESIS ASSESSMENT BOARD:
Dr.Y.A.Hussin (Chair, ITC, UT)
Dr.Ignas Heitkonig (External examiner, WUR)

ABSTRACT

Assessing the spatial distribution of giant panda is essential for efficient conservation management. GIS, remote sensing and statistics techniques have a great contribution to species distribution modelling. It has been proved that MaxEnt model is one of the most popular methods to predict species distribution and its potential suitable habitat by using presence-only data together with environmental variables. The overall objective of this study is to evaluate the effects of sample size and extent of study region on the prediction accuracy of the giant panda habitat in China using MaxEnt model.

In this research four extents of the study area for model training were selected: county level (i.e., extent of 54 administration counties with the presence of giant pandas), provincial level (i.e., extent of three provinces with the presence of giant pandas), regional level (i.e., historical regional areas with the presence of giant pandas) and national level (i.e., entire Mainland China). Ten partitions (i.e. 10%, 20%...100%) out of full giant panda occurrence records (i.e., 3032 points) were used after processing. Depending on proper environmental variables of giant panda's living condition, topographic data, climatic data, SPOT NDVI data and human disturbance data were selected. In order to evaluate model fitting for different scenarios, three accuracy measures: Area Under the receiver operating characteristic Curve (AUC), Kappa and True Skill Statistic (TSS) were used. Before systematically testing of the sample size and extent effects, a test for selecting 5,000 pseudo-absences for modelling has been carried out.

The results show that the prediction accuracy of the giant panda habitat rises with increasing sample size based on Kappa evaluation which turned out to be the best evaluation method for this study among AUC, Kappa and TSS. The value of Kappa levels off when at least 70% of the presence data were used to calibrate the model. On the other hand, the county level for predicting giant panda habitat proved to be the best extent among the four extents of the study region by areas comparison and overlay with the habitat estimated from the Third National Survey. Besides, the areas predicted by MaxEnt from the best scenario is 28,269 km^2 which is bigger than habitat estimated by the third national survey with 23,049 km^2. The most probable reason for that is both continuous suitable areas and potential living areas for giant panda has been predicted by MaxEnt modelling while the ground survey estimated practical discontinuous habitat. In general, MaxEnt is an efficient method for species distribution modelling, but sample size and extent of specific study area should be considered properly.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all the people who have helped me along the way in my MSc study especially in my thesis. I offer my sincere appreciation for the most helpful people I meet in the faculty of Geo-information Science and Earth Observation (ITC) of University of Twente.

First of all, I cannot express enough thanks to Dr. Tiejun Wang who was my primary supervisor. Words cannot begin to describe his unwavering and continuing support throughout my MSc thesis. Every time when I met with difficulties, both in research and life, he was always standing behind me. The most helpful advice and support I got from him.

The completion of this research could not have been accomplished without the support of my second supervisor Drs. E. H. Kloosterman. I will never forget the encouragement that he gave to me. He kept discovering my talents in this study through every discussion we had. He made me become more confident with myself.

Also I have to thank Yiwen Sun, my advisor studying PhD in ITC. She gave me much support on solving technical problems and increased my knowledge of giant panda background. To the many friends who helped me to manage the software and improve my thesis writing, especially Bhawana, Hossein and Nyasha, thank you very much.

Finally, to my caring and loving parents. Your encouragement and financial support for my studies at ITC was one of the sweetest things in the world. Words cannot express how thankful and grateful I am to you. I offer my heartfelt thanks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background

1.1.1. Species distribution model

Nowadays, it has been proved that species distribution models are able to determine how species are distributed in space and quantify relation between species and environmental variables. A main reason behind popularity of species distribution models is that they produce expected continuous habitat suitability maps as outputs (Andelman & Willig, 2002; Austin, 2007; Wilson et al., 2005). Numerous species distribution modelling methods exist, for instance, distance metrics (Carpenter et al., 1993) bounding boxes (Busby, 1991), logistic regression (Buckland et al., 1996), Bayesian approaches (Hepinstall & Sader, 1997), artificial neural networks (Manel et al., 1999), genetic algorithms (Stockwell, 1999) and factor analysis (Hirzel etc al.,2002). Each unique with regard to their data requirements, statistical methods and overall ease of use (Elith & Burgman, 2003; Elith et al., 2006; Guisan & Zimmermann, 2000). The predictive performances of each method is different from each other as well (Elith et al., 2006; Ladle et al., 2004; Pearson et al., 2006). However, most of the traditional models such as logistic regression and generalized linear models should have presence-absence data to estimate the relationships between species and habitat. But, the presence-absence data are costly and are also difficult to obtain for most species. In most of the cases, only presence data is available to estimate the occurrence of the species (e.g., atlases, ground survey, herbarium records and museum databases). So, nowadays, a number of new approaches such as BIOCLIM, DOMAIN, GARP and Maximum Entropy software package (MaxEnt) have been developed that utilize only presence data for species distribution modelling (Baldwin, 2009).

MaxEnt is one of the most popular species distribution models which uses presence-only data with environmental predictors to predict the species distribution. It uses incomplete information to estimate a target probability distribution by finding a probability distribution of maximum entropy (Phillips et al., 2006). The MaxEnt is frequently used because it has competitive high accuracy prediction on model performance compare to other methods and is also easy to handle (Merow et al., 2013). Because of this, government and other organizations are widely adopting MaxEnt in large-scale mapping of real-world biodiversity (Jane Elith et al., 2011). In addition, the use of statistical techniques and GIS has led to a renaissance of species distribution modelling (Wiens & Graham, 2005).

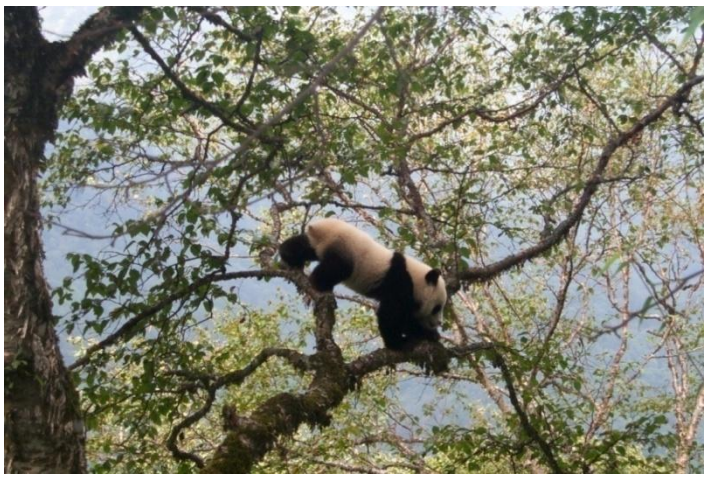


Figure 1. Giant Panda Photograph: Dr. Tiejun Wang

Figure 2. Giant Panda habitat Photograph: Dr. Tiejun Wang

1.1.2. The giant panda habitat

The giant panda, *Ailuropoda melanoleuca* (David, 1869) (Figure1), is one of the most endangered mammals in the world. In the past, fossil evidence suggests that the giant panda were widely distributed from northern Vietnam to Beijing and eastward as far as Fujian in China (Schaller, 1994). However, giant pandas have become endangered in the past few hundred years due to habitat loss, degradation and fragmentation (Wang & Xie, 2004). According to the Third Chinese National Survey conducted between 2000 and 2002, about only 1,590 pandas are living in the wild (State Forestry Administration of China, 2006). The remaining population are restricted to the Qinling area of Shaanxi Province and the high mountain ranges of Gansu and Sichuan Provinces (Hu & Wei, 2001). The Third National Survey (2000 to 2002) found 23,049 km² of panda habitat in total while it was 29,500 km² during the First National Survey (1974 to 1977). But, the Second National Survey (1985 to 1988) showed that the habitat was limited to 13,000 km² (State Forestry Administration of China, 2006). The survey showed loss of panda habitat between 1977 to 1988 while it increased between 1988 and 2002. One of the reasons of increasing of giant panda habitat was banned commercial logging across the giant pandas' habitat by Chinese government in 1998. As the methodology used during survey were different from each other, it is not possible to compare the results of the First and the Second Survey with the Third one. During First and Second Survey, sightings, spoor observation and the line-transect sampling technique were used. While, the remote sensing data and geo-spatial tools such as Global Positioning System (GPS) and GIS were used in the third survey (State Forestry Administration of China, 2006).

Assessing the spatial distribution of rare and endangered species is a key issue for efficient conservation and management (Margoluis & Salafsky, 1998; Stem et al., 2005). Accurate predictive species distribution maps are necessary to find suitable conditions and potential habitat for species. However, the prediction of giant pandas distribution is challenging because 1) giant pandas are widely dispersed in Sichuan, Shaanxi and Gansu provinces, 2) the estimated population of giant panda is low, 3) giant pandas live in solitary and 4) 99% of their diet are bamboos which are common and even dominant plants in the understory forests (Reid & Jien, 1999) (Figure2). Because of the difficulties, the previous survey extrapolated the giant panda distribution based on a sample area which cannot represent the entire range (State Forestry Administration of China, 2006). Therefore, it is important to accurately assess the distribution of remaining panda population and its habitat in China for its conservation and management.

1.1.3. Problem statement

Users of species distribution models are faced with a variety of otions, and it is not always clear how selecting one option over another (Syfert et al., 2013). In this study, we assessed the effects of numbers of pseudo-absences, sample size and extent of study region, while working with MaxEnt and giant panda presence data. That aspect of analyze the selection of pseudo-absence points, because that influences all model accuracy measures based on previous research (Lobo & Tognelli, 2011). Specifically, the quality and number of pseudo-absences can directly affect the accuracy (Barbet-Massin et al., 2012; Senay et al., 2013). While running MaxEnt, the pseudo-absence data are drawn at random from the entire region. The difference between occurrence collection and background sampling may lead to inaccurate models if the spatially biased presence data used (Park et al., 2009). Nevertheless, for this study, panda occurrence-free location data were used for generating pseudo-absence points. These panda occurrence-free location data can be considered as a true absence because the presence data were collected by an exhaustive survey throughout the study area during national survey (State Forestry Administration of China, 2006). However, it is still not clear on how many pseudo-absences should be used during modelling. Some research argue that pseudo-absences should be equally weighted to the presences while others recommend the use of a large number (e.g.10,000) of pseudo absences (Barbet-Massin et al., 2012).

Use of various numbers of presence points and extents of study area in models may also give different predictive performances (Vale et al., 2014). According to Hernandez et al., (2006), the accuracy of models is greater for species having small geographic ranges compare to wider range. The accuracy increases with increase in sample size until it approaches maximum accuracy (Hernandez, Graham, Master, Albert, & The, 2006). In contrast, some research have shown that MaxEnt is less sensitive to sample size than other algorithms (Baldwin, 2009; Wisz et al., 2008). Additionally, there is lack of general guidelines for threshold selection amongst different models (Liu et al., 2005; Nenzén & Araújo, 2011). On the other hand, the extent of study region also affects the model output. Anderson and Raza (2010) have concluded that use of small study region lead to more realistic predictions and higher estimates compare with larger study area. In addition, the study conducted by Barnes et al. (2014) reported lower accuracy of model performance when using all native range instead of incomplete one. However, there is no clear guide about selecting an appropriate extent of study region. Besides, most of study use presence points data for evaluating the model performance. However, the lack of accurate occurrence data at national and regional level is common for many countries, which makes less powerful to examine the effect of sample size and extent at a large spatial level (Kumar et al., 2014). For this study, we assumed that presence data and habitat estimated from the Third National Giant Panda Survey are accurate. Therefore, it is necessary to use the precise presence data and habitat for evaluating the model performance together with AUC, Kappa and TSS evaluations. That helps to test the effects of sample size and extent of study region in MaxEnt.

1.2. Research objectives

1.2.1. General objective

The aim of this study is to evaluate the effects of sample size and extent of study region on the prediction accuracy of the giant panda habitats in China using MaxEnt model.

1.2.2. Specific objectives

- To determine the optimal number of pseudo-absence points in MaxEnt model for predicting the suitable panda habitat
- To examine the effects of the sample size on the prediction accuracy of the panda habitat

- To examine the effects of the extent of the study region on the prediction accuracy of the panda habitat
- To assess the difference between the panda habitat predicted by MaxEnt model and the one estimated from the ground survey

1.3. Research questions

- What are the differences between 5,000 pseudo-absence points and 10,000 pseudo-absence points on the prediction accuracy of the panda habitat?
- What are the effects of the sample size on the prediction accuracy of the panda habitat?
- What are the effects of the extent of the study region on the prediction accuracy of the panda habitat?
- What are the differences between the panda habitat predicted by MaxEnt model and the one estimated from the ground survey?

1.4. Research hypotheses

- H_0: There are no statistically significant differences on the prediction accuracy of giant panda habitat in different sample sizes.
  H_1:The sample size has statistically significant effect on the prediction accuracy of the giant panda habitat.

- H_0: There are no statistically significant differences on the prediction accuracy of giant panda habitat in different extents of the study region.
  H_1:The extent of the study region has statistically significant effect on the prediction accuracy of the giant panda habitat.

- H_0: There is no statistically significant difference between giant panda habitat predicted by the MaxEnt model and the one estimated from the ground survey.
  H_1: The giant panda habitat predicted by the MaxEnt model is statistically significantly larger than the panda habitat estimated from the ground survey.

1.5. Organization of the thesis and research approach

Chapter 1 introduces a general background of this study, research problem, objectives, research questions and hypotheses. Chapter 2 provides outline of research including study area, datasets and methods. Chapter 3 lists the results relevant to research questions proposed. Chapter 4 discusses methods taken in the study and gap between predictive distribution and actual habitat. Last but not the least, chapter 5 gives conclusion of the research and recommends further studies.

Figures 3 and Figure 4 present the framework of research approaches. The Figure 3 shows how to determine numbers of pseudo-absence points in MaxEnt by comparing model performances between using 5,000 pseudo-absences and using 10,000 pseudo-absences. Took the selected numbers of pseudo-absence from this step to examine effect of sample sizes and extent. Three accuracy measures (i.e. AUC, Kappa and TSS) were used to evaluate model fitting for different scenarios. Finally, high suitability maps were found after evaluation and comparison between predicted habitats and habitat from ground survey.

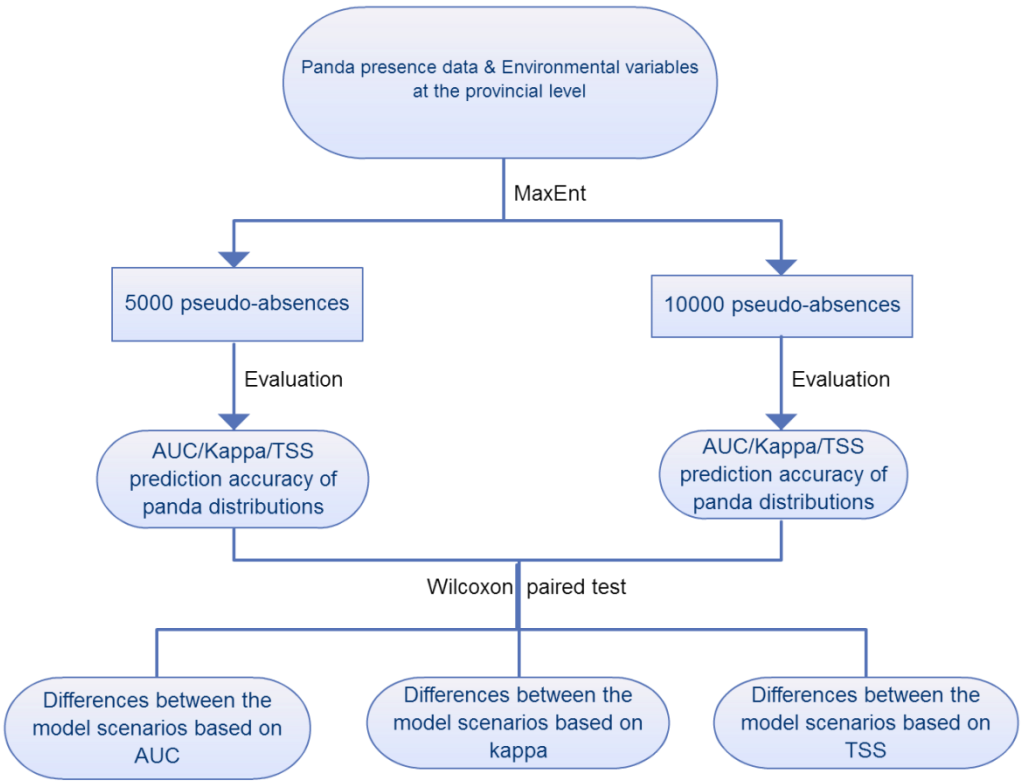


Figure 3. Approach to determine number of pseudo-absences in MaxEnt on modelling the distribution of giant panda

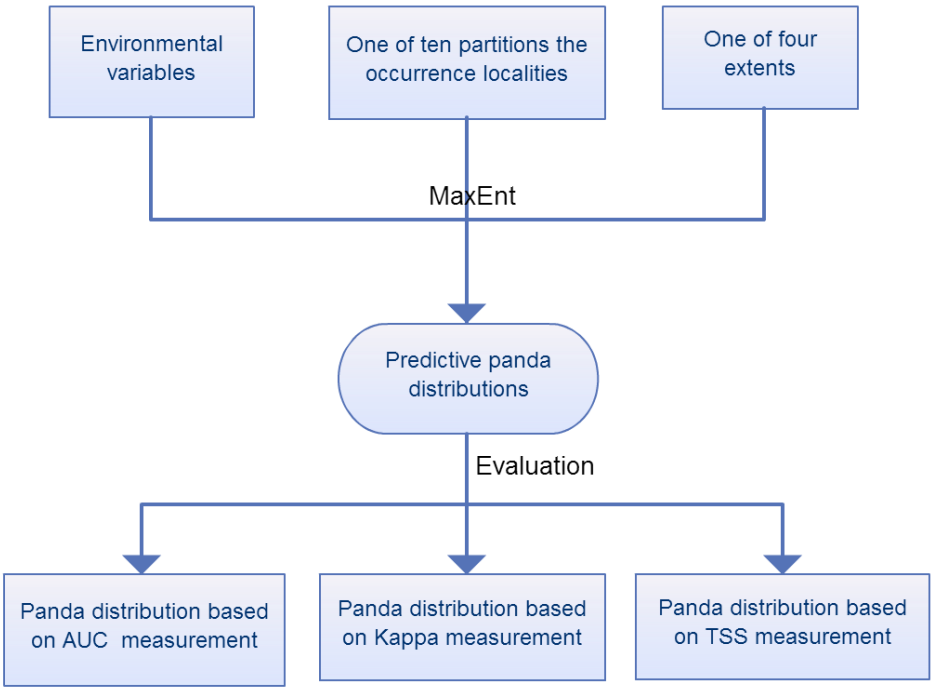


Figure 4. Approach to evaluate the effects of sample size and extent in MaxEnt on modelling the distribution of giant panda

2. MATERIALS AND METHODS

2.1. Extent of the study region

Figure 5 shows the four extents of the study region namely county level, provincial level, regional level and national level. According to the Third National Panda Survey from 2000 to 2002， the giant panda was observed in 54 administration counties with area about 160,000 km² in China. So this study defined the boundary of these 54 counties as the first study area extent. The second extent of the study region is at the provincial level where wild panda existed in the past decades. The provincial level includes Shaanxi, Gansu and Sichuan provinces of China having approximately 1,000,000 km² area (Reid & Jien, 1999). The historical and regional distribution range of the giant pandas inside China is used as the third extend of the study region which is about 3,000,000 km². The boundary of Mainland China with an area of about 9,600,000 km² was selected as last extent for the study.

The red part in Figure 5 and the green patches in Figure 6 show the current giant panda habitat which is about 23,049 km² according to Third National Panda Survey (State Forestry Administration of China, 2006). The giant panda habitat range is located on 102⁰00'-108⁰11'E longitude to 27⁰53'-35⁰35'N latitude (Hu & Wei, 2001). The habitat ranges between 1,000-3,500 m elevation which include five mountain ranges: Qinling, Minshan, Qionglai, Xiangling (includes both Greater and Lesser Xiangling) and Liangshan (Hu, 2001; Schaller, 1994). These mountains have bamboo as the dominant understory species which is a prominent source of food for giant panda.

.

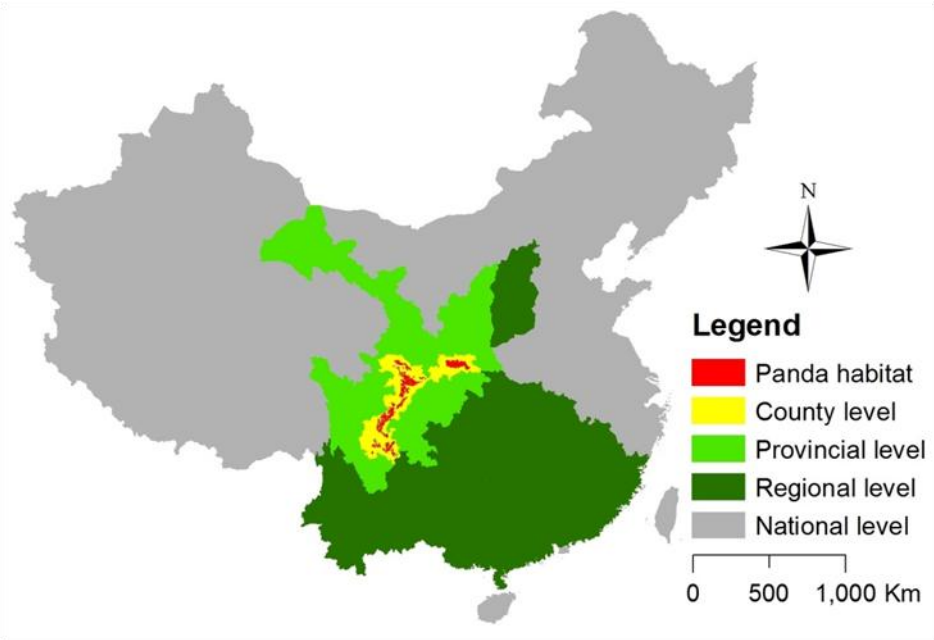


Figure 5. The extend of the four study regions for giant panda habitat modelling

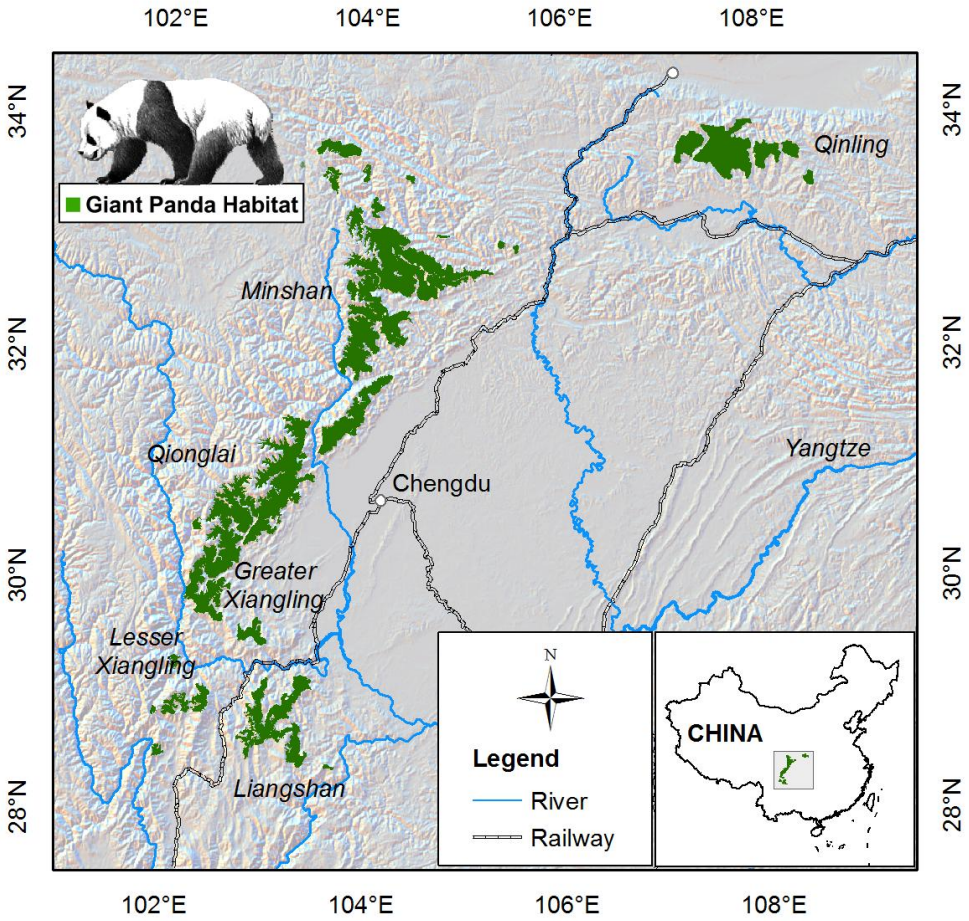


Figure 6. The remaining panda habitats (shown by green patches) in the west part of China estimated from the Third National Giant Panda Survey (2000 to 2002)

2.2. Data preparation and pre-processing

2.2.1. Giant panda occurrence data and re-sampling

A shapefile including 4,964 giant panda occurrence points (i.e., the direct sighting of pandas and its signs) were derived from the Third National Giant Panda Survey conducted by the State Forestry Administration of China during 2000 to 2002. This survey covered the whole area known to have a panda population as well as the areas thought to potentially have populations via a dragnet investigation approach. The whole investigation area was plotted out 11,174 plots in total with an average plot size of 2 km^2 (State Forestry Administration of China, 2006). These points represent locations where pandas and their traces were observed. The location of plots were recorded by GPS in GCS_WGS_1984 system. 3,032 points were left after removing duplicate points in each 1 km*1 km resolution square. Then, remaining 3,032 points were sub-sampled into ten partitions randomly (i.e. 10%, 20%...100%). After that, the partitions were extracted and converted to csv format for processing in MaxEnt. Figure 7 shows the ten partitions of giant panda presence points at county level.

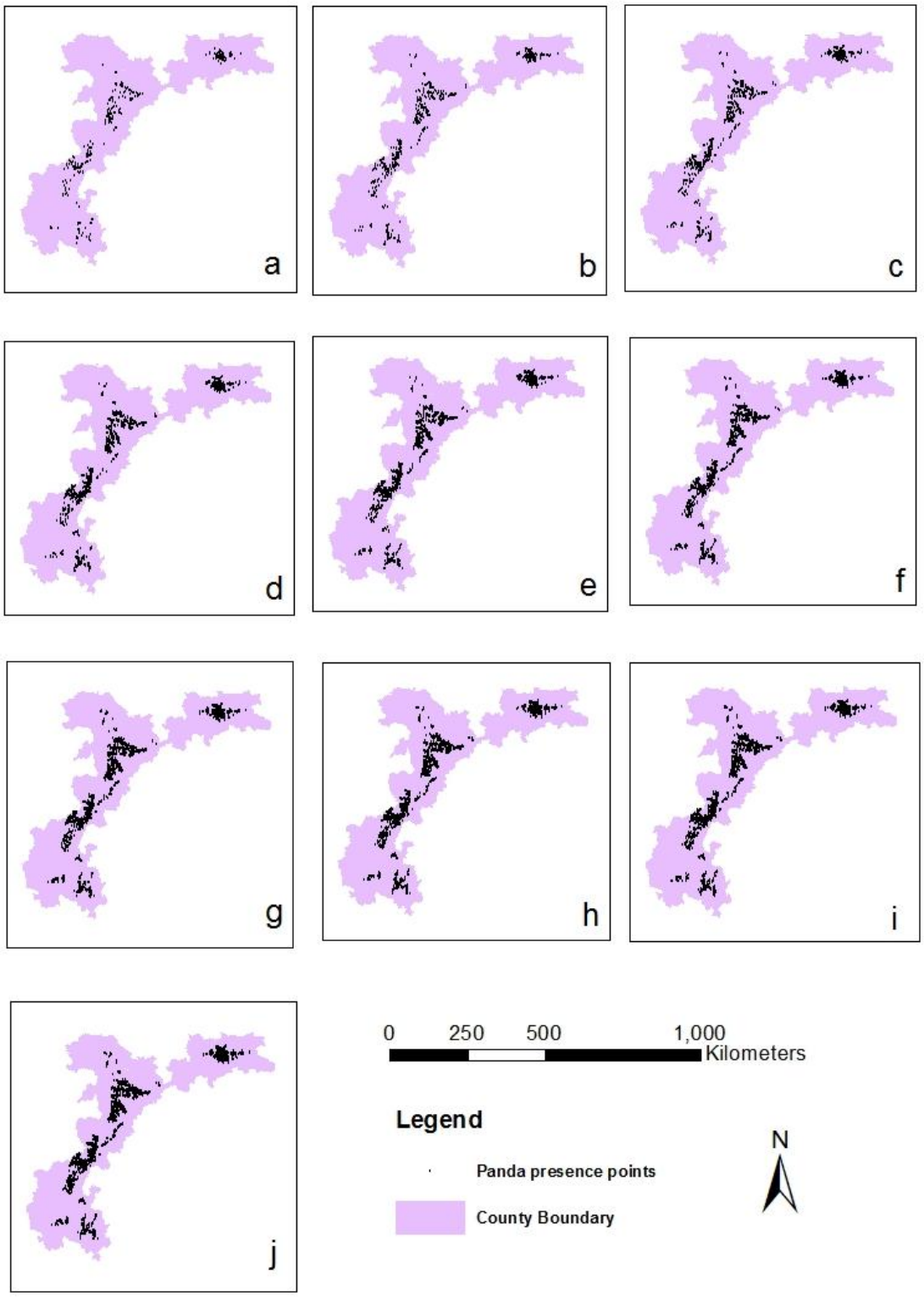


Figure 7. Maps showing ten partitions of giant panda presence points at county level:(a)using 10% of presences; (b)using 20% of presences; (c)using 30% of presences; (d)using 40% of presences; (e)using 50% of presences; (f)using 60% of presences; (g)using 70% of presences; (h)using 80% of presences; (i)using 90% of presences and (j)using full presences.

2.2.2. Environmental variables

- Topographic data

  Topographic variable is a key driver of biodiversity. For this study, these variables were derived from the WorldClim-Global Climate Database (http://www.worldclim.org/). These variable are continuous layers with one square kilometer spatial resolution in GCS_WGS_1984 projection (Rosenzweig, 1995). Also, the DEM were derived from same database. The ancillary data such as elevation, slope and aspect maps were extracted from DEM in ENVI. Finally, the ancillary data were clipped into four subsequent extent of study area (Figure 3).

- Climate data

  Climatic data were also obtained from the WorldClim-Global Climate Database (http://www.worldclim.org/). It is a set of continuous global climate layers (climate grids) with a spatial resolution of one square kilometer recording from the 1950-2000 period (Hijmans et al., 2005). The climate data include monthly precipitation, mean, minimum, and maximum temperature (Hijmans et al., 2005). Eighteen climatic layers were used in this study except "Precipitation of driest quarter" because of its bad quality (Table1).

- SPOT NDVI data

  The Normalized Difference Vegetation Index (NDVI) is often used as a simple graphical indicator to observe the vigor of green vegetation. It is calculated from individual measurements of NIR and VIS, as shown below:

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{VIS})}{(\mathrm{NIR} + \mathrm{VIS})}$$

  where, VIS and NIR stand for the visible (red) and near-infrared regions respectively.

  In this study, ten-day synthesis of SPOT-VEGETATION images were obtained from VITO website (http://www.vito-eodata.be/PDF/portal/Application.html#Home) from year 2000 to 2002. The images area projected in plate carree with 1 km resolution. After stacking 12-month multi-temporal NDVI data into one image, these time series images were smoothed in ENVI. Additionally, the maximum NDVI, mean NDVI, minimum NDVI, amplitude NDVI and NDVI standard deviation were extracted and calculated in ENVI

- Human population density

  The raster layer of human population density was obtained from the Land Administration Bureau of China. The pixel size of raster layer is 1 km by 1 km and the population density is in number of people per square kilometer. It was collected by the National Bureau of Statistics in China during the Fifth Population Census 2000.

- Roads

  The raster layer of distance to roads was also obtained from Land Administration Bureau of China. The pixel size is 1 km*1 km and the distance is measure in kilometer.

All the environmental variable layers were rasterized into the same bounds, cell size and same coordinate system as the layer of occurrence localities in ArcGIS. Then environmental variable layers were re-projected in GCS_WGS_1984 with one square kilometre spatial resolution. Finally, all these layers were converted to the ASCII format for further calculation at MaxEnt.

Table 1. Environmental variables used for modelling the habitat of giant panda

Data source	Category	Variables	Abbreviation	Units
WorldClim	Bio-climatic	Annual mean temperature	Bio1	^{0}C
		Mean diurnal range	Bio2	^{0}C
		Isothermality	Bio3	Dimensionless
		Temperature seasonality	Bio4	Dimensionless
		Max temperature of warmest month	Bio5	^{0}C
		Min temperature of coldest quarter	Bio6	^{0}C
		Temperature annual range	Bio7	^{0}C
		Mean temperature of wettest quarter	Bio8	^{0}C
		Mean temperature of driest quarter	Bio9	^{0}C
		Mean temperature of warmest quarter	Bio10	^{0}C
		Mean temperature of coldest quarter	Bio11	^{0}C
		Annual precipitation	Bio12	mm
		Precipitation of wettest month	Bio13	mm
		Precipitation of driest quarter	Bio14	mm
		Precipitation seasonality	Bio15	Dimensionless
		Precipitation of wettest quarter	Bio16	mm
		Precipitation of driest quarter	Bio17	mm
		Precipitation of warmest quarter	Bio18	mm
		Precipitation of coldest quarter	Bio19	mm
WorldClim	Topographic	Altitude	Altitude	m
		Slope	Slope	Degree
		Aspect	Aspect	Degree
SPOT-VGT	Vegetation	Annual minimum NDVI	NDVI_min	Dimensionless
		Annual mean NDVI	NDVI_mean	Dimensionless
		Annual maximum NDVI	NDVI_max	Dimensionless
		Standard deviation NDVI	NDVI_std	Dimensionless
Administration in China	Human population	Population density	Pop_den	Number of people /km^2
	Roads	Distance to road	Road_dis	kilometer

2.3. Selection of number of pseudo- absence points

It is important to decide what number of pseudo-absences should be used before running the model. Before testing of sample size and extent effects, two different pseudo-absence points (i.e. 5,000 and 10,000) were selected to compare which number of pseudo-absence points give a higher accuracy for model performance. Out of four types of extent, provincial extent was used as wild panda existing today only in these three provinces of China (Reid & Jien, 1999). According to Barbet-Massin et al. (2012), a larger spatial extent is needed to optimise model performance at a given spatial resolution for ensuring the selection of enough informative pseudo-absences. However, the sensitivity of pseudo-absence point become lower with increasing extent such as national and regional extent of study area. Provincial level is neither too large nor too small compared with the other three extents. So, provincial level was chosen to determine the number of pseudo-absence points. The other input indicators, for instance, the number of presences and the environmental layers, were same for running MaxEnt. After MaxEnt running, AUC and predicted probability for both presence points and pseudo-absence points were obtained. After that, Kappa and TSS were calculated in R program by the probability prediction of presences and pseudo-absences. The probabilities were used to test the difference between 5,000 pseudo-absences scenario and 10,000 pseudo-absences scenario by Wilcoxon signed-rank test. Finally the optimal one was selected based on higher accuracy for further analysis.

2.4. Modelling approach - MaxEnt

MaxEnt, also called ecological niche modelling, is based on a machine learning method with precise mathematical formulation to make predictions for species distribution modelling (Phillips et al., 2006). The MaxEnt approach was chosen for this study because it does not requires true absence points reducing workload for collecting data and has very good predictive performance even using sparse or noisy input information (Elith et al., 2006). Besides, MaxEnt provides output data in three formats i.e. raw, cumulative and logistic formats in comparison to other modelling methods. The logistic format is easy to conceptualize as it gives an estimate between 0 and 1 of probability of presence. Also, the MaxEnt has ability to run the Jackknife test which estimates the significance of environmental variables in computing the species distribution (Phillips & Dudík, 2008). The important environmental variables for giant panda showed in Appendix.

In order to examine how sample size affects the model accuracy, this study sub-sampled ten partitions (i.e. 10%, 20%...100%) from presence points on four different extents respectively. The four extents are 54 counties with the presence of giant pandas, three provinces with the presence of giant pandas, historical areas with the presence of giant pandas and the Mainland China. In other words, to know the effect of sample size, each out of four extents was taken and compared accuracy difference within ten partitions While, for testing the effect of extent, each of ten partitions was taken and compared accuracy amongst four extents.

2.5. Measures of model performance

In this study, three methods were used to evaluate the accuracy of model performance. They are Area Under the receiver operating characteristic Curve (AUC), Kappa and True skill statistic (TSS).

Area Under the receiver operating characteristic Curve (AUC)

Receiver operating characteristic (ROC) evaluates the performance of the model when there was no absence data. Based on Allouche et al. (2006), ROC curve is created by plotting the true positive against the false positive (equal to 1-specificity) rate (Table 2). The AUC of the ROC plot is considered an effective indicator for model performance, which provides a single measure of overall accuracy that is not

dependent upon a particular threshold (Fielding & Bell, 1997). The AUCs ranges from 0 to 1, where 1 indicates perfect model, ≥0.750 indicates best model category, 0.5 is random model while ≤ 0.5 is a worse model than random (Phillips & Dudík, 2008).

Kappa

Kappa is one of the most widely used measures of model performance in ecology (Allouche etc al., 2006). The Kappa index gives a less biased measure of predictability as it considers both omission and commission errors (Table 2). However, several studies have criticized it for being inherently dependent on prevalence (Allouche et al., 2006). The Kappa value ranges from -1 to 1, where +1 indicates perfect fit and 0 or less indicate a performance no better than random (Cohen, 1960).

True Skill Statistic (TSS)

TSS corrects for the dependence of prevalence while keeping all Kappa advantages. It takes both omission and commission errors into account, and successes as a result of random guessing (Table 2). The values range is from -1 to 1, where 1 indicates perfect agreement and 0 or less indicates a performance no better than random (Allouche et al., 2006).

Kappa and TSS are threshold-dependent methods. An threshold value is needed to transform the results of species distribution modelling from probabilities to a binary map (Liu et al., 2005). However, this is no clear value of threshold identified. In some ecological researches, the probability threshold classifies all the areas of probability greater than 0.5 as suitable areas for species while all the areas below 0.5 as absent. In this case, the subjective dichotomy value of 0.5 seems arbitrary, lacking ecological basis (Osborne et al., 2001). Nowadays, more advanced techniques for selecting a probability threshold have been developed. The sensitivity and specificity of model makes the result more powerful are required during analysis, while the sensitivity-specificity sum maximization approach turns out to be one of good approaches for threshold determination, which can be processed by PresenceAbsence package in R program (Liu et al., 2005). Hence, threshold of maximum TSS was used to differentiate the suitable and non-suitable habitat for giant panda prediction were used in this study.

Table 2. Measures of predictive accuracy

Measure	Formula
Overall accuracy	$\frac{a+d}{n}$
Sensitivity	$\frac{a}{a+c}$
Specificity	$\frac{d}{b+d}$
Kappa statistic	$\frac{\left(\frac{a+d}{n}\right)-\frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1-\frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}$
TSS	Sensitivity + specificity - 1

In all formulae: n=a+b+c+d, (a)True positive, (b)False positive, (c) False negative and (d) True negative

2.6. Statistical Analysis

The Wilcoxon Signed-rank test, a non-parametric equivalent of a paired t-test, was used to compare the differences in accuracy assessed by three measurements (i.e. AUC, Kappa and TSS) between the model scenarios. The null hypothesis of Wilcoxon Signed-rank test is that two populations are the same against

an alternative hypothesis. Significant difference at $p<0.05$ between model scenarios was considered as non-identical populations. These tests were conducted in R.

After calculating the accuracies of model performances and comparing the differences between model scenarios, the most accurate predictive models based on AUC/Kappa/TSS were obtained. Differences between habitat predicted by the most accurate predictive models and habitat estimated from the ground survey were assessed by overlaying analysis and then area was calculated in ArcGIS.

3. RESULTS

3.1. Effects of the numbers of pseudo-absence points on model prediction accuracy

Table 3 shows the p-values are less than 0.05 for all partitions of sample sizes based on AUC and Kappa, indicating using 5,000 pseudo-absence points and 10,000 pseudo-absence points are statistically significant different. Therefore, it was accepted that using 5,000 background points are different from using 10,000 background points. However, the difference was not statistically significant for almost all the scenarios of sample size based on TSS evaluation. That means TSS was not sensitive to the numbers of pseudo-absence points. In order to select the optimal number of pseudo-absences, we compared the accuracy of each scenario. The average accuracy graphs for two scenarios of pseudo-absences are shown in Figure 8, Figure 9 and Figure 10.

Table 3. p-values based on AUC/Kappa/TSS and ten partitions of sample sizes

p-value / Sample size (%)	AUC	Kappa	TSS
10	0.000	0.000	0.064
20	0.000	0.000	0.898
30	0.000	0.000	0.097
40	0.000	0.000	0.076
50	0.000	0.000	0.898
60	0.000	0.000	0.202
70	0.000	0.000	0.870
80	0.000	0.000	0.246
90	0.000	0.000	0.000
100	0.000	0.000	0.729

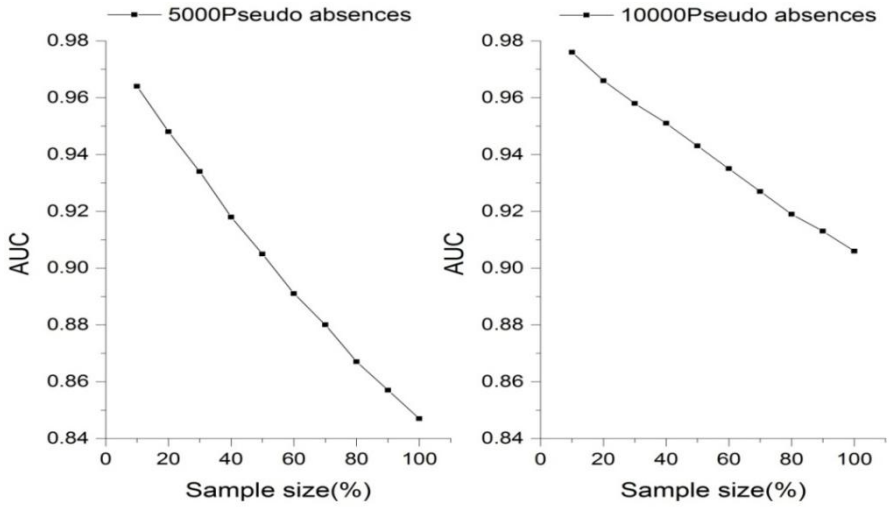


Figure 8. Prediction accuracy of different pseudo-absences based on AUC

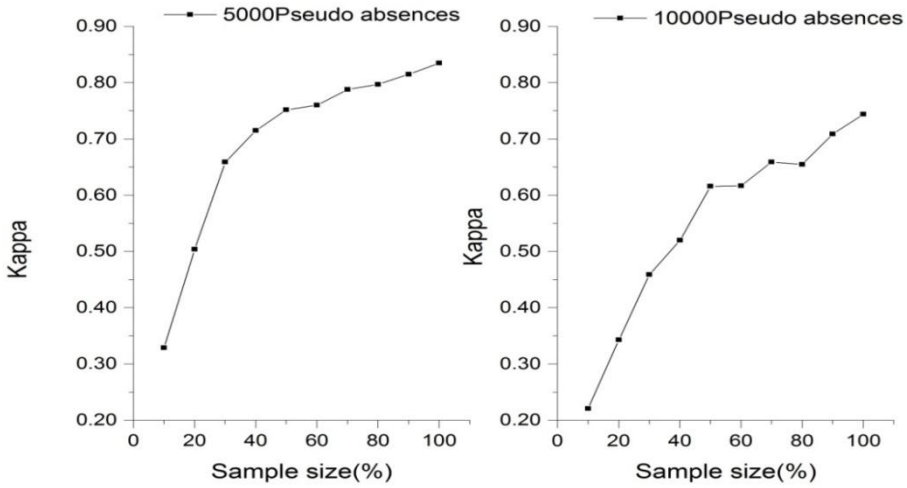


Figure 9. Prediction accuracy of different pseudo-absences based on Kappa

In general, the accuracy from 10,000 pseudo-absences scenario was higher than 5,000 pseudo-absences scenario based on AUC evaluation (Figure 8). On the other hand, Kappa evaluation method provided the opposite trend, where the accuracy from 5,000 pseudo-absences was higher than 10,000 pseudo-absences in every sample size scenario (Figure 9). Even though the result shows that TSS was not sensitive to number of pseudo-absences, the graphs show interesting results (Figure 10). The trend of TSS graphs were similar with Kappa graphs, which means the accuracy was increasing with increased number of presence data.

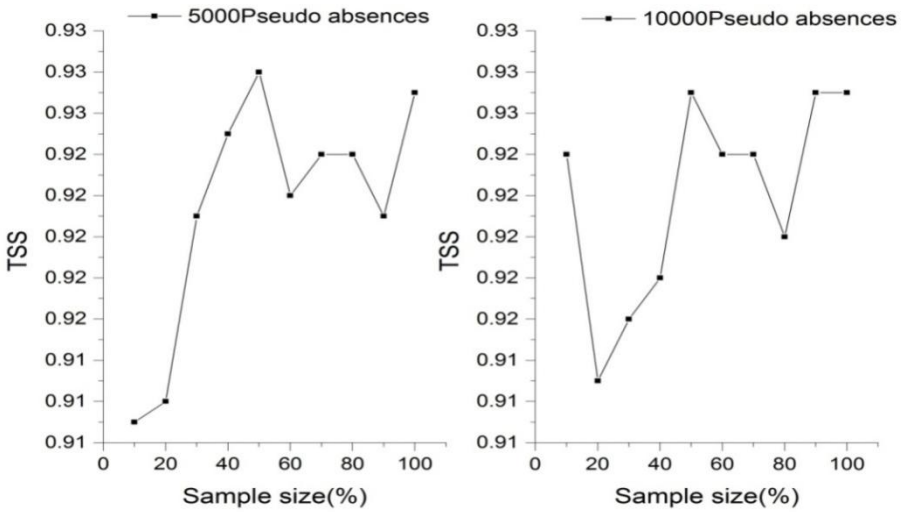


Figure 10. Prediction accuracy of different pseudo-absences based on TSS

3.2. Effect of the sample size on model prediction accuracy

3.2.1. Prediction accuracy based on AUC

Five thousands pseudo-absence points were used to further test according to the analysis on Chapter 3.1. The p-values were ascertained for each pair of sample sizes groups as shown in Table 4 to Table 7. These tables revealed that there were differences among ten sample sizes in AUC. The county level, provincial level, historical level and national level follow the same trend.

Table 4. Wilcoxon paired test (p-value) for AUC to test effect of sample size at county level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.086	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.000	0.000	0.000	0.000	0.000
60						0.000	0.000	0.000	0.000
70							0.000	0.000	0.000
80								0.000	0.000
90									0.003

Table 5. Wilcoxon paired test (p-value) for AUC to test effect of sample size at provincial level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.000	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.000	0.000	0.000	0.000	0.000
60						0.000	0.000	0.000	0.000
70							0.000	0.000	0.000
80								0.000	0.000
90									0.000

Table 6. Wilcoxon paired test (p-value) for AUC to test effect of sample size at regional level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.000	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.000	0.000	0.000	0.000	0.000
60						0.000	0.000	0.000	0.000
70							0.000	0.000	0.000
80								0.000	0.000
90									0.000

Table 7. Wilcoxon paired test (p-value) for AUC to test effect of sample size at national level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.000	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.000	0.000	0.000	0.000	0.000
60						0.000	0.000	0.000	0.000
70							0.000	0.000	0.000
80								0.000	0.000
90									0.000

The study also analyzed how sample size affects AUC at each extent level. Figure 11 demonstrates AUC varying in different situations. The graphs show AUC were gradually decreasing from 10% of panda presences to 100% of panda presences at all four levels. In specific, AUC decreased from 0.906 to 0.809 at county level while accuracy fell from 0.964 to 0.847 at provincial level. Also, at regional level and national level, AUC were decreasing from 0.975 to 0.852 and from 0.979 to 0.855, accordingly.

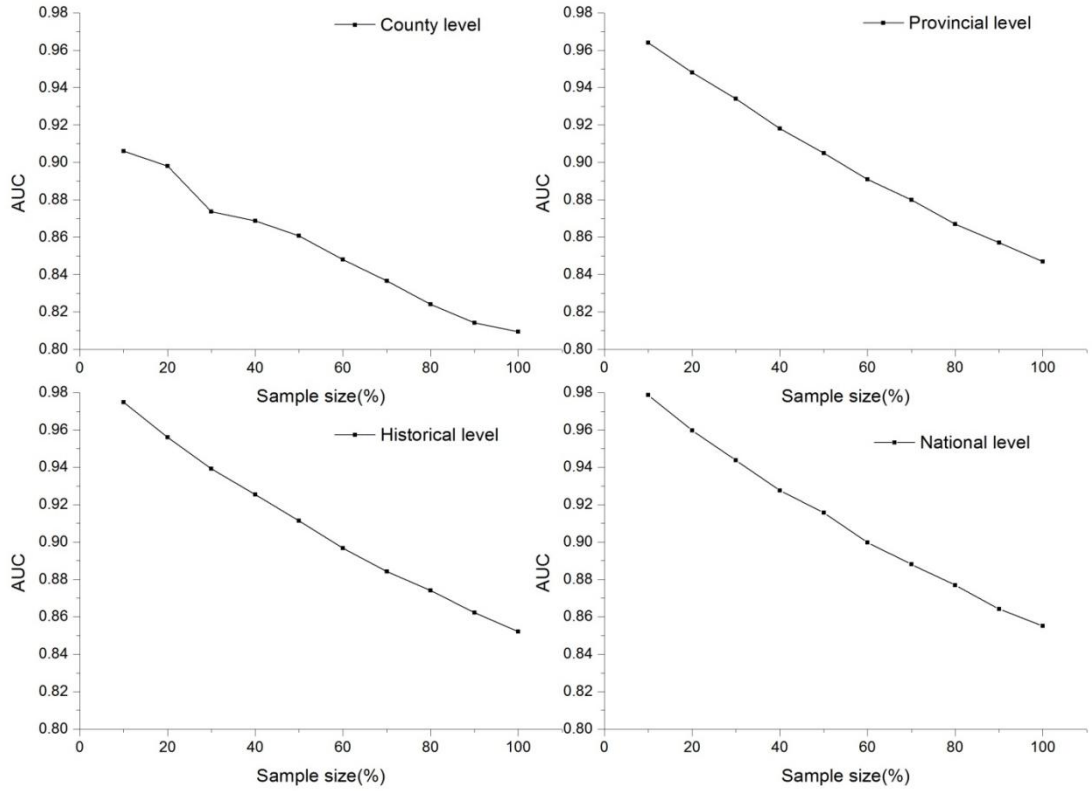


Figure 11. AUC vary in ten partitions of sample sizes based on four extents of the study region

3.2.2. Prediction accuracy based on Kappa

At four extents of the study region, the statistic differences among ten partitions of panda occurrences were tested respectively. The p-values from Wilcoxon Signed-rank paired test were obtained for each pair of sample sizes demonstrating in Table 8 to Table 11. In general, the statistics show that the sample size does affect Kappa accuracy of modelling.

Table 8. Wilcoxon paired test (p-value) for Kappa to test effect of sample size at county level

Sample Size (%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.000	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.004	0.000	0.000	0.000	0.000
60						0.000	0.000	0.000	0.000
70							0.277	0.000	0.000
80								0.000	0.000
90									0.000

Table 9. Wilcoxon paired test (p-value) for Kappa to test effect of sample size at provincial level

Sample Size (%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.001	0.000	0.000	0.000	0.000	0.000	0.000
40				0.001	0.001	0.000	0.000	0.000	0.000
50					0.330	0.000	0.000	0.000	0.000
60						0.001	0.000	0.000	0.000
70							0.044	0.000	0.000
80								0.000	0.000
90									0.000

Table 10. Wilcoxon paired test (p-value) for Kappa to test effect of sample size at regional level

Sample Size (%)	20	30	40	50	60	70	80	90	100
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.000	0.000	0.000	0.000	0.000	0.000	0.000
40				0.005	0.001	0.000	0.000	0.000	0.000
50					0.756	0.058	0.000	0.000	0.000
60						0.058	0.000	0.000	0.000
70							0.000	0.000	0.000
80								0.898	0.154
90									0.090

Table 11. Wilcoxon paired test (p-value) for Kappa to test effect of sample size at national level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30			0.898	0.000	0.000	0.000	0.000	0.000	0.000
40				0.000	0.000	0.000	0.000	0.000	0.000
50					0.956	0.985	0.133	0.000	0.000
60						0.648	0.064	0.000	0.000
70							0.154	0.000	0.000
80								0.000	0.000
90									0.064

Figure 12 describes how Kappa varying from 10% of panda presence points to entire panda presences. Kappa showed increasing trend from the 10% of presences to the full presences which was opposite to AUC trend. In addition, Kappa rose from 0.131 to 0.554 and from 0.329 to 0.835 as sample sizes increased at county level and provincial level respectively. While, at historical level and national level, Kappa gradually increased from 0.520 to 0.890, and 0.706 to 0.956 respectively.

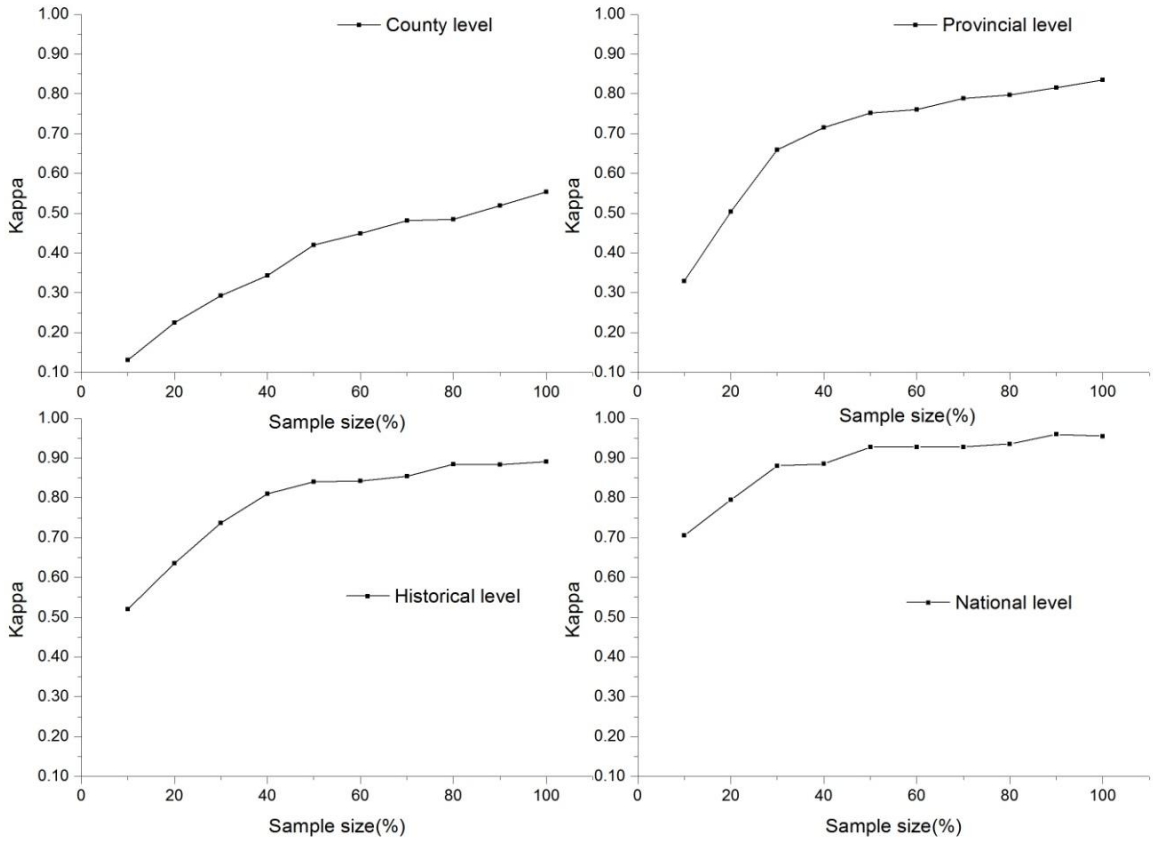


Figure 12. Kappa vary in ten partitions of sample sizes based on four extents of the study region

3.2.3. Prediction accuracy based on TSS

The p-values from Wilcoxon Signed-rank paired test were calculated for each pair of sample sizes groups (Table 12 to Table 15). TSS showed that there were no statistically significant differences among ten partitions of sample sizes which was different from AUC and Kappa.

Table 12. Wilcoxon paired test (p-value) for TSS to test effect of sample size at county level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.648	0.701	0.701	0.027	0.114	0.452	0.784	0.571	0.097
20		0.985	0.956	0.000	0.002	0.083	0.097	0.076	0.004
30			0.261	0.000	0.000	0.012	0.017	0.004	0.000
40				0.001	0.012	0.076	0.409	0.294	0.005
50					0.177	0.006	0.001	0.004	0.312
60						0.044	0.015	0.006	0.756
70							0.312	0.674	0.044
80								0.674	0.001
90									0.003

Table 13. Wilcoxon paired test (p-value) for TSS to test effect of sample size at provincial level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.756	0.076	0.014	0.000	0.036	0.003	0.015	0.033	0.001
20		0.003	0.000	0.000	0.006	0.000	0.001	0.001	0.000
30			0.036	0.002	0.430	0.097	0.277	0.812	0.011
40				0.216	0.312	0.898	0.522	0.058	0.261
50					0.007	0.017	0.036	0.002	0.956
60						0.097	0.475	0.898	0.044
70							0.870	0.076	0.123
80								0.388	0.021
90									0.001

Table 14. Wilcoxon paired test (p-value) for TSS to test effect of sample size at regional level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.956	0.701	1.000	0.812	0.648	0.216	0.648	0.498	0.571
20		0.756	0.898	0.648	0.841	0.165	0.368	0.596	0.522
30			0.596	0.498	0.596	0.294	0.330	1.000	0.898
40				0.729	0.956	0.064	0.246	0.729	0.648
50					0.216	0.004	0.898	0.133	0.123
60						0.048	0.522	0.870	0.870
70							0.001	0.021	0.036
80								0.231	0.053
90									0.898

Table 15. Wilcoxon paired test (p-value) for TSS to test effect of sample size at national level

Sample Size(%)	20	30	40	50	60	70	80	90	100
10	0.430	0.294	0.294	0.870	0.173	0.287	0.065	0.452	0.956
20		0.648	0.952	0.105	0.784	0.522	0.498	0.003	0.012
30			0.701	0.143	0.812	0.870	0.388	0.001	0.024
40				0.246	0.956	0.756	0.349	0.004	0.012
50					0.083	0.277	0.008	0.033	0.430
60						0.360	0.202	0.000	0.004
70							0.058	0.000	0.014
80								0.000	0.000
90									0.294

Figure 13 shows TSS accuracies based on four extent levels. There was more or less no change in TSS from 10% of panda occurrences to 100% of panda presences. TSS value ranges from 0.733 to 0.756 at county level which was lower than the other three extents of study region. Whereas, the TSS values at provincial level, regional level and national level are [0.911, 0.958], [0.953, 0.958] and [0.977, 0.983], respectively.

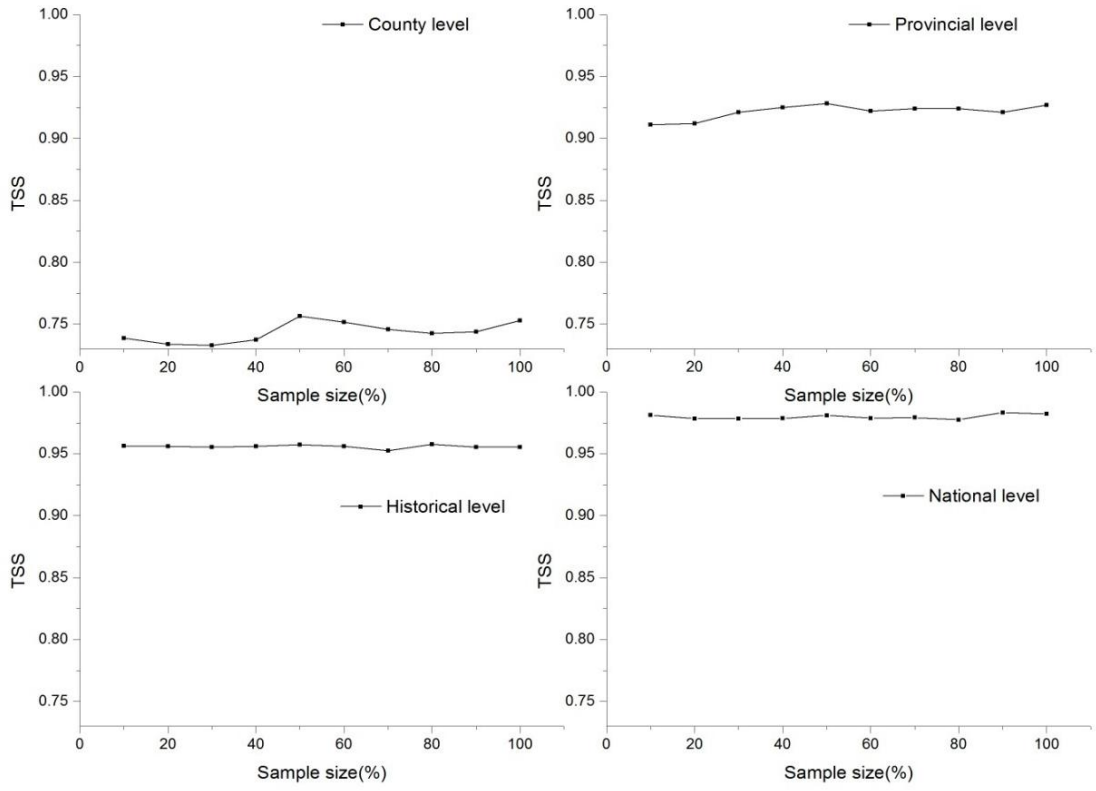


Figure 13. TSS vary in ten partitions of sample sizes based on four extents of the study region

3.3. Effect of extent of the study region on model prediction accuracy

Among four extents of study region, giant panda distribution prediction at national level was the best for all the sample sizes based on AUC, following by regional level, provincial level and county level (Figure 14). The differences among four extents of study region were analyzed in vertical direction. From each sample size, the ranking of AUC of four extents of study region were same. For instance, the prediction at national level gave the highest AUC following regional, provincial and county levels when we used 10% of presences for modelling while the same phenomenon as using 20% of presences.

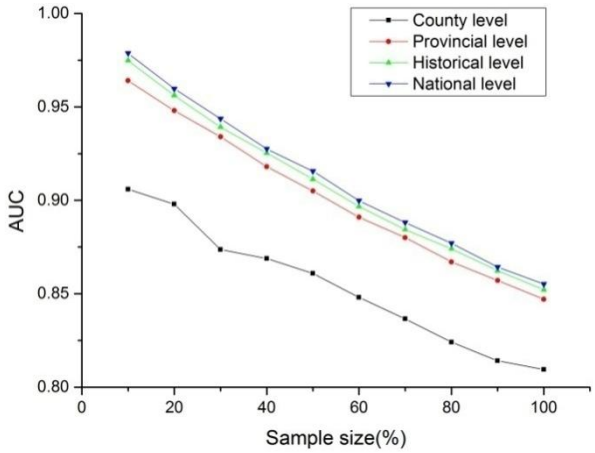


Figure 14. AUC variation in four extents of the study region

The prediction at national level also had the highest Kappa/TSS among four extent levels on each sample size (shown in Figure 15 and Figure 16). The graphs clearly show that the national level had the highest accuracy followed by regional level, provincial level and county level, respectively.

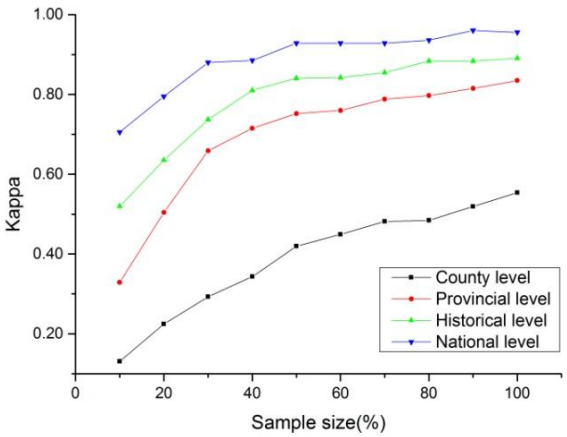


Figure 15. Kappa variation in four extents of the study region

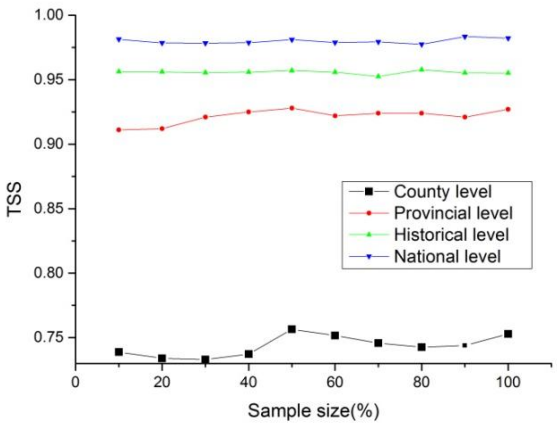


Figure 16. TSS variation in four extents of the study region

The Wilcoxon Signed-rank paired test was done between each pair of two extent levels to test the model differences (Table 16 to Table 21). The results revealed that the four extents of study region had statistically significant differences from each other based on AUC, Kappa and TSS.

Table 16. Wilcoxon paired test for AUC, Kappa and TSS to test difference between county and provincial levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.000	0.000	0.000
30	0.000	0.000	0.000
40	0.000	0.000	0.000
50	0.000	0.000	0.000
60	0.000	0.000	0.000
70	0.000	0.000	0.000
80	0.000	0.000	0.000
90	0.000	0.000	0.000
100	0.000	0.000	0.000

Table 17. Wilcoxon paired test for AUC, Kappa and TSS to test difference between provincial and regional levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.000	0.000	0.000
30	0.001	0.000	0.000
40	0.000	0.000	0.000
50	0.000	0.000	0.000
60	0.000	0.000	0.000
70	0.000	0.000	0.000
80	0.000	0.000	0.000
90	0.000	0.000	0.000
100	0.000	0.000	0.000

Table 18. Wilcoxon paired test for AUC, Kappa and TSS to test difference between provincial and national levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.000	0.000	0.000
30	0.000	0.000	0.000
40	0.000	0.000	0.000
50	0.000	0.000	0.000
60	0.000	0.000	0.000
70	0.000	0.000	0.000
80	0.000	0.000	0.000
90	0.000	0.000	0.000
100	0.000	0.000	0.000

Table 19. Wilcoxon paired test for AUC, Kappa and TSS to test difference between county and regional levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.000	0.000	0.000
30	0.000	0.000	0.000
40	0.000	0.000	0.000
50	0.000	0.000	0.000
60	0.000	0.000	0.000
70	0.000	0.000	0.000
80	0.000	0.000	0.000
90	0.000	0.000	0.000
100	0.000	0.000	0.000

Table 20. Wilcoxon paired test for AUC, Kappa and TSS to test difference between county and national levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.000	0.000	0.000
30	0.000	0.000	0.000
40	0.000	0.000	0.000
50	0.000	0.000	0.000
60	0.000	0.000	0.000
70	0.000	0.000	0.000
80	0.000	0.000	0.000
90	0.000	0.000	0.000
100	0.000	0.000	0.000

Table 21. Wilcoxon paired test for AUC, Kappa and TSS to test difference between regional and national levels

p-value / Sample size(%)	AUC	Kappa	TSS
10	0.000	0.000	0.000
20	0.002	0.000	0.000
30	0.000	0.000	0.000
40	0.018	0.000	0.000
50	0.001	0.000	0.000
60	0.004	0.000	0.000
70	0.002	0.000	0.000
80	0.009	0.000	0.000
90	0.113	0.000	0.000
100	0.007	0.000	0.000

3.4. Probability of suitable giant panda habitats

Probability maps of suitable habitat for giant panda for forty scenarios were derived from MaxEnt modelling (Figure 17 to Figure 20). The four groups of maps represent probability maps at county level, provincial level, regional level and national level, respectively. The larger extent of the study region, the larger area of predicted habitat. However, there were no big area differences among ten sample sizes within the same extent of the study region.

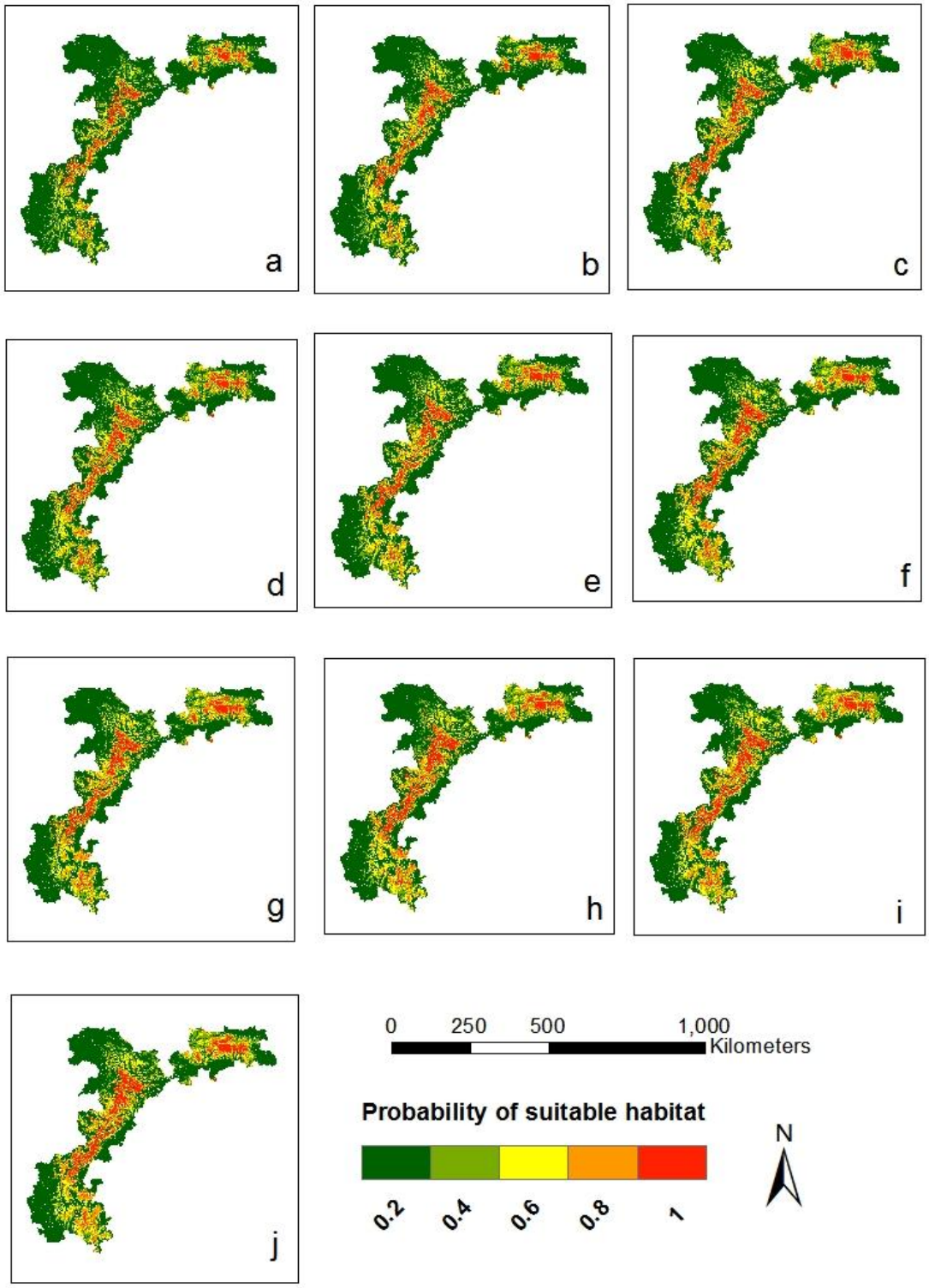


Figure 17. Maps showing the probability of suitable habitat of giant panda at county level for ten models:(a)using 10% of presences; (b)using 20% of presences; (c)using 30% of presences; (d)using 40% of presences; (e)using 50% of presences; (f)using 60% of presences; (g)using 70% of presences; (h)using 80% of presences; (i)using 90% of presences and (j)using full presences.

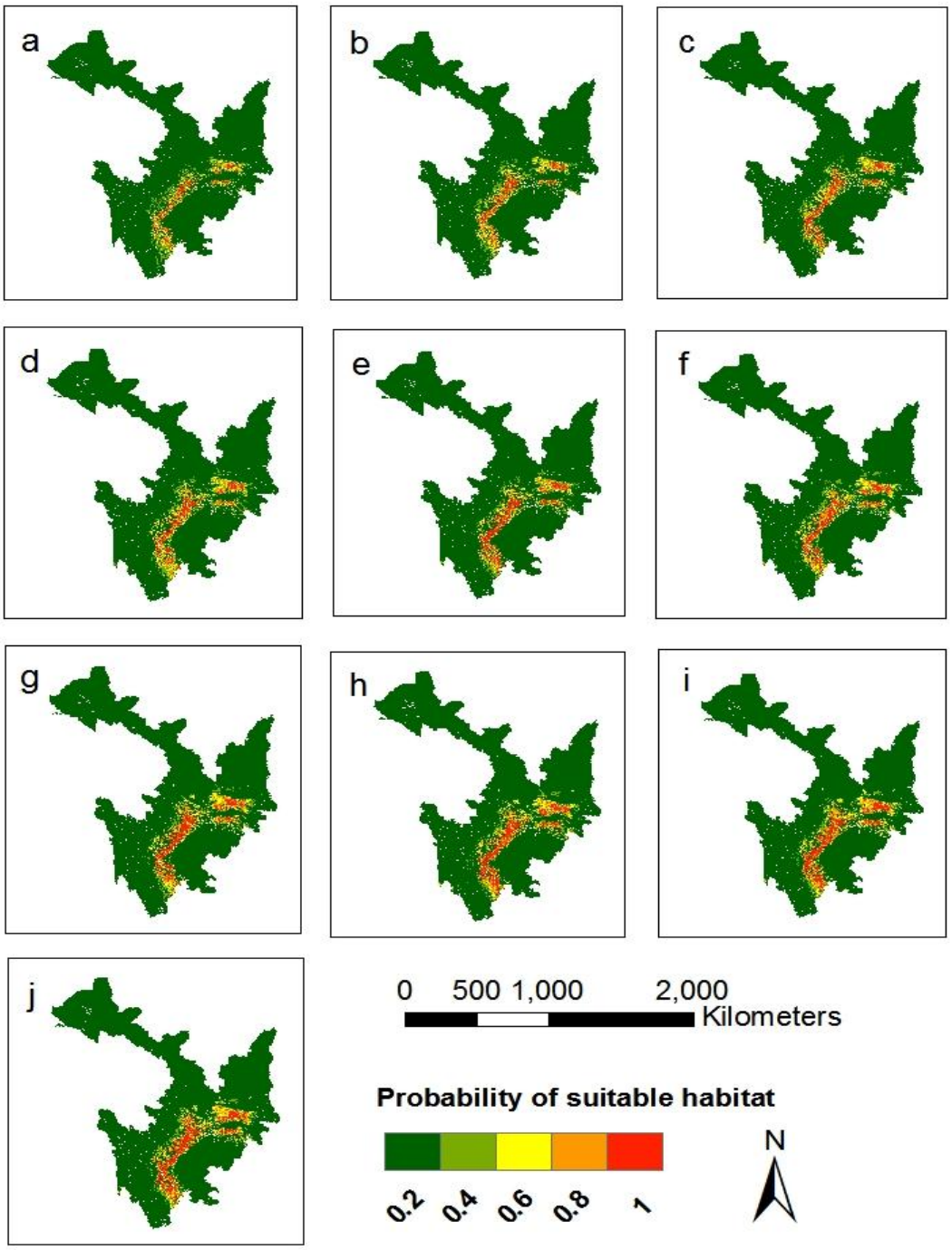


Figure 18. Maps showing the probability of suitable habitat of giant panda at provincial level for ten models:(a)using 10% of presences; (b)using 20% of presences; (c)using 30% of presences; (d)using 40% of presences; (e)using 50% of presences; (f)using 60% of presences; (g)using 70% of presences; (h)using 80% of presences; (i)using 90% of presences and (j)using full presences.

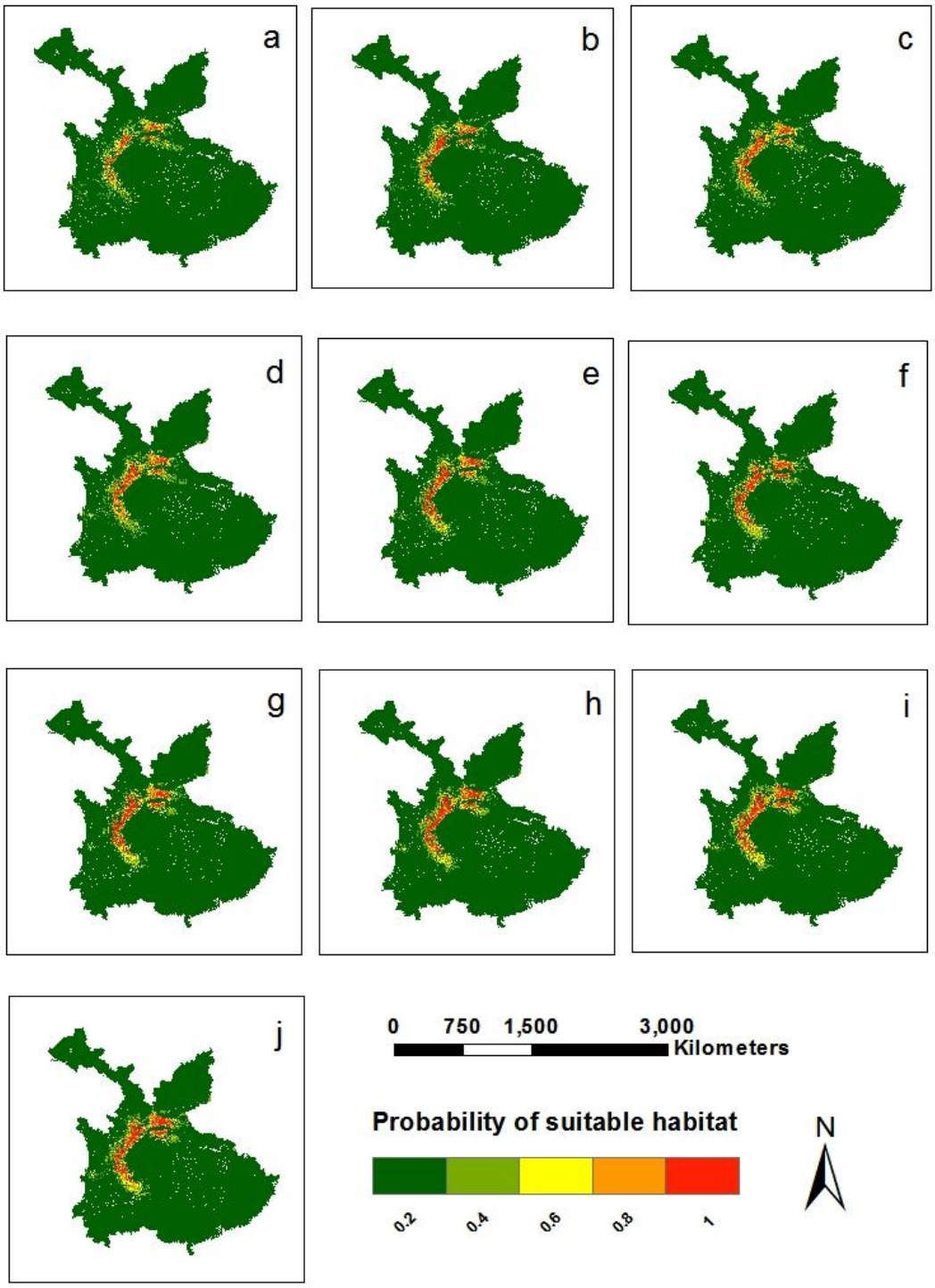


Figure 19. Maps showing the probability of suitable habitat of giant panda at historical regional level for ten models:(a)using 10% of presences; (b)using 20% of presences; (c)using 30% of presences; (d)using 40% of presences; (e)using 50% of presences; (f)using 60% of presences; (g)using 70% of presences; (h)using 80% of presences; (i)using 90% of presences and (j)using full presences.

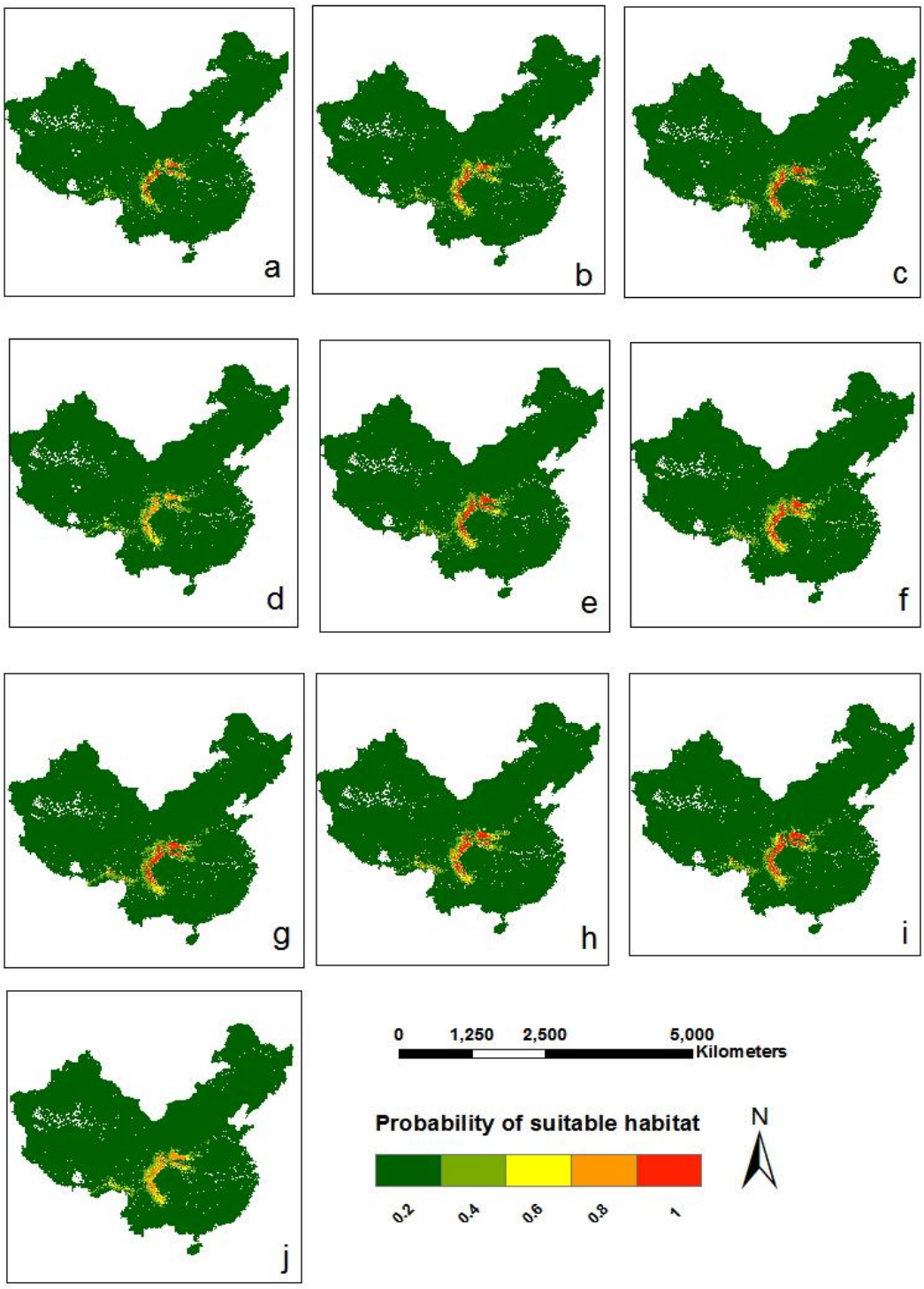


Figure 20. Maps showing the probability of suitable habitat of giant panda at national level for ten models:(a)using 10% of presences; (b)using 20% of presences; (c)using 30% of presences; (d)using 40% of presences; (e)using 50% of presences; (f)using 60% of presences; (g)using 70% of presences; (h)using 80% of presences; (i)using 90% of presences and (j)using full presences.

3.5. Comparison between predicted habitat and ground survey habitat

In order to compare differences between predicted habitat and habitat obtained from the Third National Survey, the area of habitats was calculated in ArcGIS (Table 22). According to the accuracy comparison, AUC was highest at 10% of presence points at national level whereas, Kappa and TSS were highest while using full presences at national level. Taking the effect of extent into consideration, this study selected predicted suitable habitats from county level, provincial level and regional level using 10% of presences and full presences as well. Therefore, predicted suitable habitats were selected from these eight scenarios. Additionally, area comparison by overlaying between one out of eight scenarios and habitat from ground survey was done as shown in Figure 21 to Figure 24.

Table 22. Habitat area predicted by MaxEnt

Level of extents Sample size(%)	Habitat area			
	County Level (km^2)	Provincial Level (km^2)	Regional Level (km^2)	National Level (km^2)
10	28,073	66,592	80,404	120,318
20	28,589	65,044	99,877	148,615
30	27,499	58,530	100,310	128,030
40	29,857	57,993	84,610	152,762
50	27,916	63,638	84,660	120,214
60	28,382	58,153	92,764	129,570
70	26,594	60,830	97,936	148,165
80	29,519	63,120	90,140	132,947
90	27,924	62,372	95,915	141,560
100	28,269	65,009	96,748	145,609

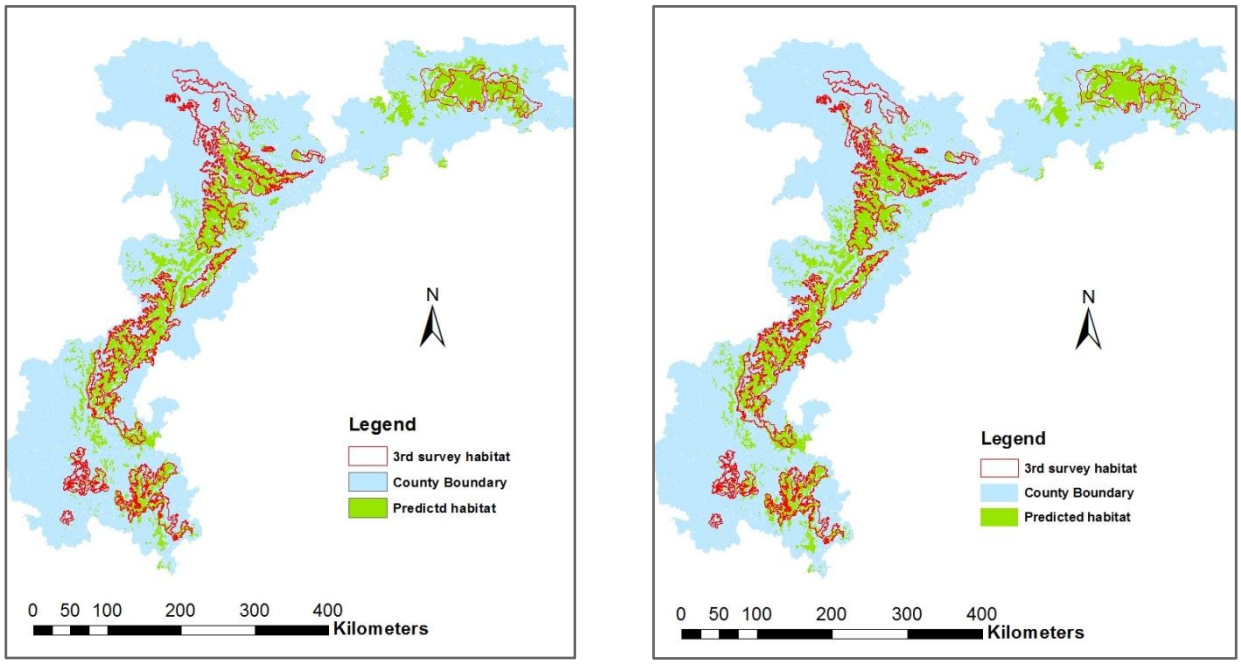


Figure 21. Overlay between the Third National Survey habitat and predicted habitat at county level: with 10% of presences on the left and with full presences on the right

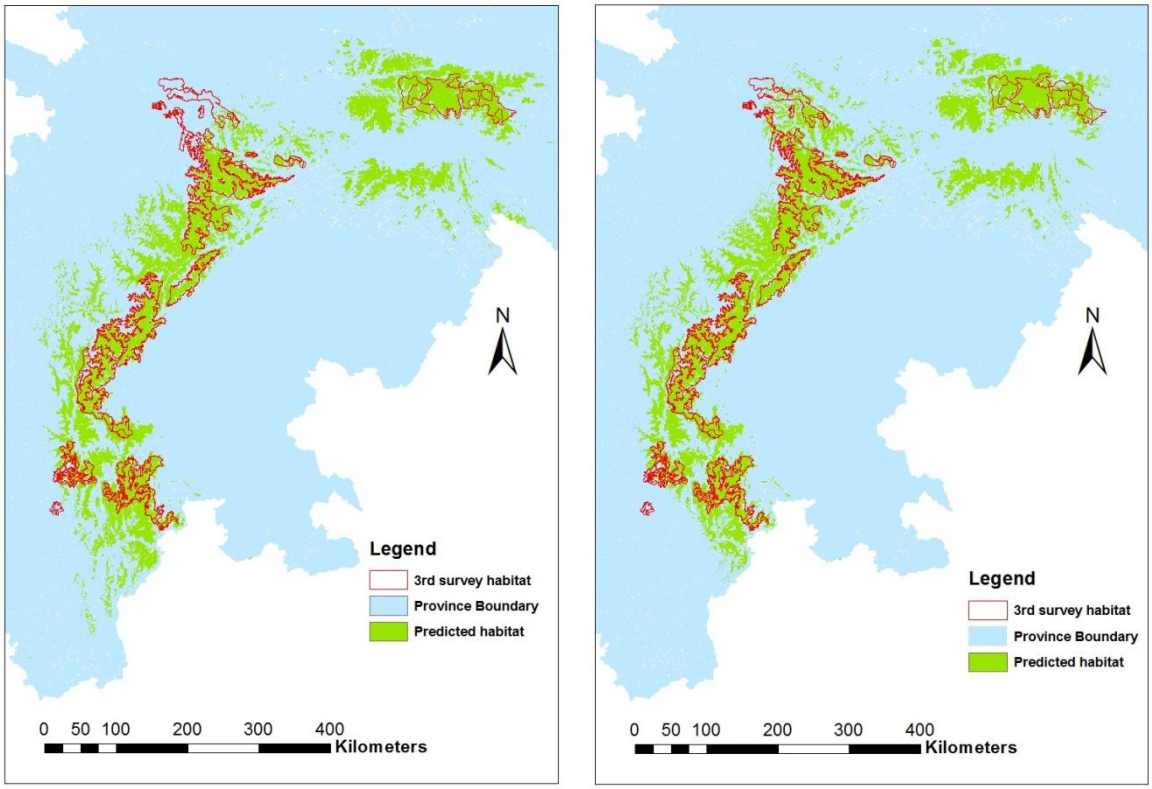


Figure 22. Overlay between the Third National Survey habitat and predicted habitat at provincial level: with 10% of presences on the left and with full presences on the right

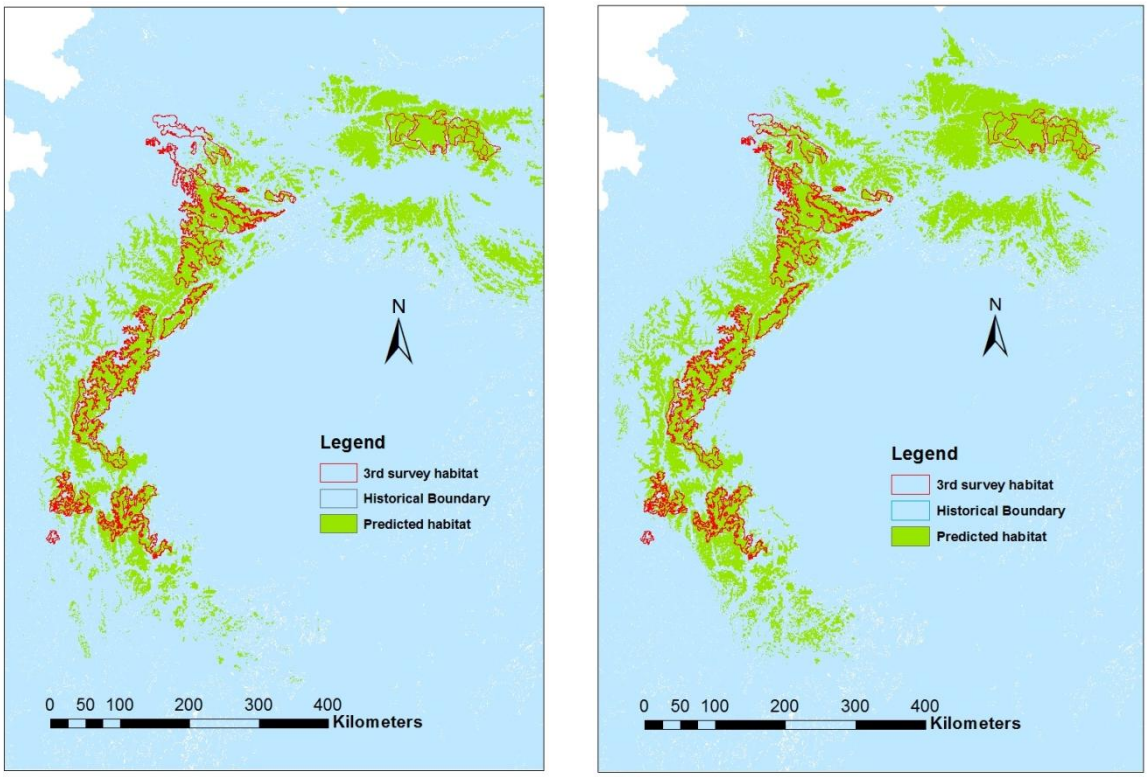


Figure 23. Overlay between the Third National Survey habitat and predicted habitat at regional level: with 10% of presences on the left and with full presences on the right

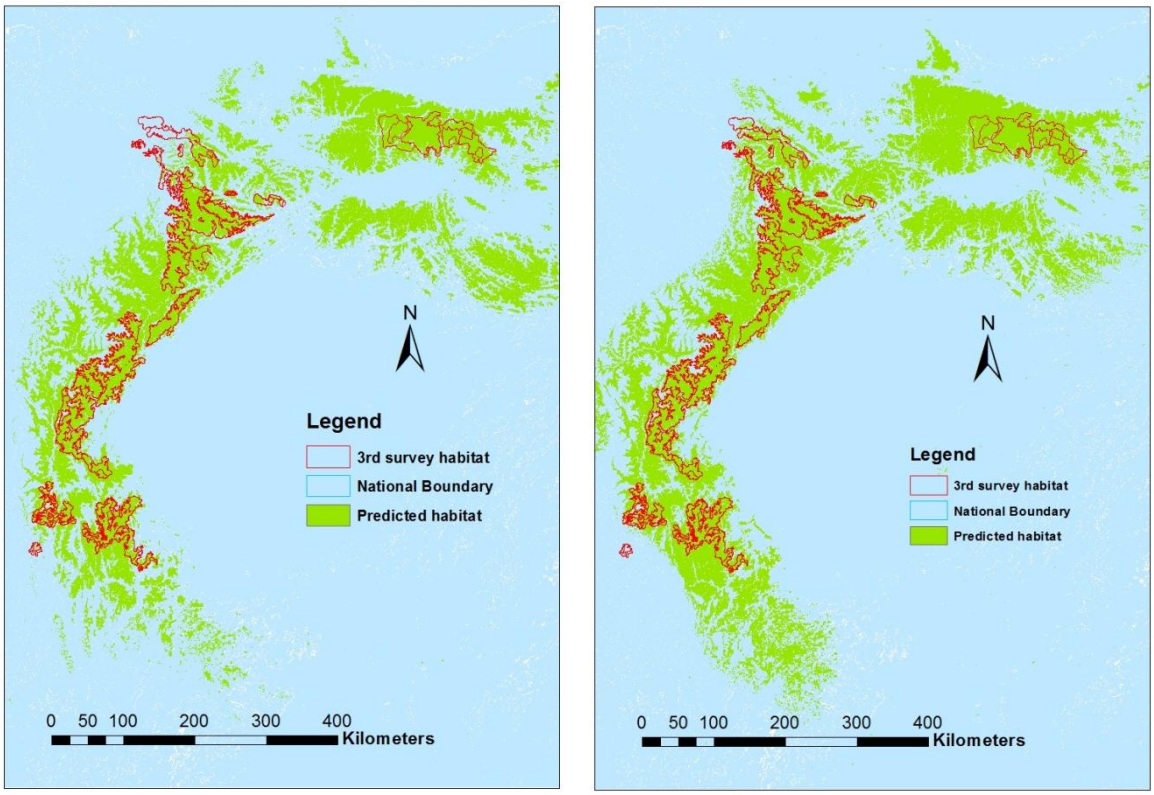


Figure 24. Overlay between the Third National Survey habitat and predicted habitat at national level: with 10% of presences on the left and with full presences on the right

4. DISCUSSION

4.1. Effect of the number of pseudo-absence points on the model prediction accuracy

The different evaluation methods (i.e. AUC, Kappa and TSS) investigated in this study behave differently on accuracy of model performances. The Wilcoxon Signed-rank test revealed that the AUC evaluation using 5,000 pseudo-absences are statistically different from using 10,000 pseudo-absences (Table 3). According to Figure 8, the higher AUC value was obtained from 10,000 pseudo-absences scenario than 5,000 pseudo-absences scenario. However, the trend of AUC value from this study is not logical compared to previous work. Hernandez et al. (2006) reported that accuracy should increase with increase in sample size. But the opposite accuracy trend occurs while using ten ascending sample sizes in both 5,000 pseudo-absences and 10,000 pseudo-absences scenarios (the AUC trend shown in Figure 8). While for Kappa evaluation, the accuracy trend is similar with the results of Hernandez et al. (2006). The accuracy of 5,000 pseudo-absences was statistically significant higher than 10,000 pseudo-absences. The TSS evaluation showed no statistically significant difference between using 5,000 pseudo-absences and 10,000 pseudo-absences. Besides, using 5,000 pseudo-absence points for further analysis is less workload than using 10,000 pseudo-absence points. Therefore, 5,000 pseudo-absences were used to test continue effects of sample size and extent tests.

4.2. Effect of the sample size on the model prediction accuracy

For AUC evaluation, the accuracy between each pair of sample sizes were statistically significant different. Also the accuracy of models gradually decreases from 10% of panda presences to full panda presences. The trend of accuracy is not similar compared to previous work. However, for Kappa evaluation, the value level off after 70% of presences which about 2,100 occurrence records (the accuracy shown in Figure 12). That means the Kappa increases until achieving its potentially maximum accuracy as sample size increases as mentioned by Hernandez et al. (2006). In case of TSS, the Wilcoxon Signed-rank test showed there is no statistically significant differences among ten sample sizes while the accuracy stays constant from 10% of presences to full presences. This indicates that TSS is not sensitive to sample size.

On the other hand, the TSS graphs showed more fluctuate than AUC and Kappa graphs. One of the reasons behind is that TSS is very sensitive to the location of presence points. While taking different 10% out of full presences, the result shows in Figure 25 that the accuracies are different with different random 10% of presences. TSS of 0.924 in original scenario and TSS of 0.908 from test are the average of 20 times model running. So, it can be concluded that TSS are sensitive to location of presences but not sensitive with pseudo-absences location. In this case, TSS is not the best evaluation method for testing the effects of sample size and extent of study region when the value is uncertain. Therefore, the Kappa evaluation is the best method to test the effect of sample size in this study. Buckland et al. (1996) have mentioned that the accuracy increased when more restrictive thresholds were used. In case of giant panda data, models with more restrictive thresholds such as 2,100 presence data tend to be more accurate.

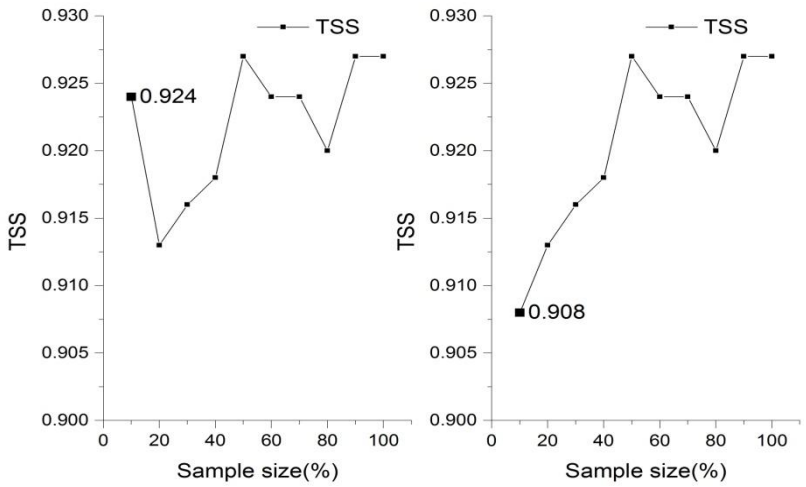


Figure 25. TSS sensitivity on location of presences test: former TSS on the left and test TSS on the right

4.3. Effect of the extend of study regions on the model prediction accuracy

The three evaluation methods showed that the differences exist between each pair of extents of study region. Generally, it is supposed that the national level, the highest level, provided highest accuracy among four extents of study region following regional level, provincial level and county level. However, other research concluded that smaller study region led to more realistic predictions and higher estimated of niche conservatism (Anderson & Raza, 2010). For instance, as Vale et al. (2014) presented, continental models tended to overestimate species distribution while regional models show better fitting to presence data. Also, according to the study conducted by Anderson and Raza (2010), the selection of reasonable study region than the extremely large ones which are in common use. One of the reasons for avoiding extremely large extent is that too large spatial extent is prone to model over-fitting (Anderson & Raza, 2010). However, we could not simply selected the best extent of study region only based on AUC, Kappa and TSS. The other evaluation method, area comparison and overlay between the predicted habitat by MaxEnt and estimated habitat from ground survey, should be added to evaluate the model performance. The result would be more accurate when using the actual habitat for assessing.

4.4. The difference between the panda habitat predicted by MaxEnt model and the one derived from the ground survey

For this study, it is assumed that the third national giant panda survey is accurate and reliable. The habitat area is 23,049 km² which represents the realized giant panda habitat. However, the results from predicted models provided different result compared to the third national giant panda survey. Among all, the area from predicted at county level i.e. 28,269 km² is closer with the areas estimated from the third national giant panda survey. However, the predicted habitat areas at provincial level, regional level and national level are much larger than the Third Ground Survey (Table 22). This results may be due to effects of extents. Provincial extent, regional extent and national extent are much larger than the real habitat which can be seen from Figure 5. Based on the study by Anderson and Raza (2010), it is clear that the large extent is prone to over-fitting the model. Hence, the result from county level is the most reasonable among different scenarios according to area comparison and overlay. However, that predicted habitat

areas at county level with 28,269 km² was larger than habitat from the Third National Giant Panda Survey with 23,049 km². The main reason is that the predicted habitat from MaxEnt includes all the continuous suitable areas and potential living areas for giant panda while most of habitats investigated from ground survey are discrete practical areas. On the other hand, the area differences among different sample sizes within the one extent of study region are not obvious according to Table 22, Figure 21 to Figure 24. In summary, MaxEnt is more sensitive to extent of study region than sample size.

5. CONCLUSIONS AND RECOMMENDATIONS

This study fulfils all the objectives and answers all four research questions. More specifically, the aim of this study is to evaluate the effects of sample size and extent of study region on the prediction accuracy of giant panda habitats using MaxEnt for species distribution modelling. In order to reach the overall objective, four aspects have been explored: 1) determining the number of pseudo-absence points, which is important for model running; 2) effects of sample size on the prediction accuracy of giant panda habitat; 3) effects of study region on the prediction accuracy of giant panda habitat; 4) comparison between predicted giant panda habitat and habitat estimated from the third national ground survey. The conclusions from the study are summarized as follows:

- Five thousand pseudo-absence points are chosen for modelling in this study based on accuracy assessed by AUC, Kappa and TSS measurements. Specifically, AUC and Kappa show statistically significant difference between using 5,000 pseudo absences and 10,000 pseudo absences. However, the accuracy gotten from 10,000 pseudo-absences scenario is higher based on AUC evaluation while the accuracy from 5,000 pseudo-absences scenario is higher based on Kappa. All the accuracies evaluated by AUC method are demonstrating high values ranging from 0.84 to 0.98, while accuracy range evaluated by Kappa is moderate.

- Prediction accuracy of giant panda habitat rises with increasing sample size based on Kappa evaluation. The value of Kappa level off after 70% of presences were used, which are about 2,100 occurrence records. That means, accuracy rises by increasing sample size in MaxEnt would reach a saturation point. So, neither too few presence points nor too many presences are good for selection.

- The county level for predicting giant panda habitat turns out to be the best extent of study region among county level, provincial level, regional level and national level by areas comparison and overlay with habitat estimated from the third national survey. The provincial level, regional level and national level are too big, which are prone to over model- fitting. Therefore, the proper extent of study region should be used for species distribution modelling, which shouldn't be too big.

- The areas predicted by MaxEnt from the best scenario is 28,269 km^2 which is larger than habitat estimated by the Third National Survey 23,049 km^2. This difference is coming from including all the continuous suitable and potential areas by MaxEnt. In general, MaxEnt is more sensitive to extent of study region than sample size.

All in all, results of this study demonstrate that sample size and extent of the study region do affect on prediction accuracy of giant panda habitat. Furthermore, the accuracy of modelling depends on many factors, for instance, the sample size of presence data, the extent of study area, the quality and spatial resolution of the environmental and species data and modelling method itself (Hernandez et al., 2006; Vale et al., 2014). It is recommended to take sample size, extent and resolution into consideration in future study. Besides, we suggest to use more presence points for modelling if possible to confirm the existence of the maximum accuracy asymptote. It is also worthy to use predicted habitat from this study to compare with habitat which will be estimated from the fourth national survey in a few years.

LIST OF REFERENCES

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. doi:10.1111/j.1365-2664.2006.01214.x

Andelman, S. J., & Willig, M. R. (2002). Alternative Configurations of Conservation Reserves for Paraguayan Bats: Considerations of Spatial Scale. *Conservation Biology*, *16*(5), 1352–1363. doi:10.1046/j.1523-1739.2002.01119.x

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, *37*(7), 1378–1393. doi:10.1111/j.1365-2699.2010.02290.x

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*(1-2), 1–19. doi:10.1016/j.ecolmodel.2006.07.005

Baldwin, R. a. (2009). Use of Maximum Entropy Modeling in Wildlife Research. *Entropy*, *11*(4), 854–866. doi:10.3390/e11040854

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. doi:10.1111/j.2041-210X.2011.00172.x

Buckland, S. T., Elston, D. A., & Beaney, S. J. (1996). Predicting distributional change,with application to bird distribution in northeast Scotland. *Global Ecology and Biogeography Letters*, *5*, 66–84.

Busby, J. R. (1991). BIOCLIM-a bioclimate analysis and prediction system. *Nature Conservation:cost Effective Biological Surveys and Data Analysis(ed.by C.R.Margules and M.P.Austin)*, (CSIRO,Canberra,ACT,Australia), 64–68.

Carpenter, G., Gillison, A. N., & Winter, J. (1993). DOMAIN:a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, *2*, 667–680.

Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Elith, J., & Burgman, M. A. (2003). Habitat Models for Population Viability Analysis. In C. A. Brigham & M. W. Schwartz (Eds.), *Population Viability in Plants* (Vol. 165, pp. 203–235). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-662-09389-4

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. doi:10.1111/j.2006.0906-7590.04596.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. doi:10.1111/j.1472-4642.2010.00725.x

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. doi:10.1017/S0376892997000088

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2-3), 147–186. doi:10.1016/S0304-3800(00)00354-9

Hepinstall, J. A., & Sader, S. A. (1997). Using Bayesian statistics,Thematic Mapper satellite imagery,and breeding bird survey data to model bird species probability of occurrence in Maine. *Photogrammetric Engineering and Remote Sensing*, *63*, 1231–1237.

Hernandez, P. A., Graham, C. H., Master, L. L., Albert, D. L., & The, A. D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *5*(June), 773–785.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. doi:10.1002/joc.1276

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecologial niche factors analysis: How to compute habitat suitability maps without absence data? *Ecology*, *83*(7), 2027–2036. doi:10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2

Hu, J. (2001). Research on the giant panda. *Shanghai Scientific and Technological Education Publishers*, *Shanghai*, 402pp.

Hu, J., & Wei, F. (2004). Comparative Ecology of Giant Pandas in the Five Mountain Ranges of Their Distribution in China. In L. Donald (Ed.), *Giant Pandas: Biology and Conservation* (pp. 137–148). doi:10.1525/california/9780520238671.003.0015

Kumar, S., Graham, J., West, A. M., & Evangelista, P. H. (2014). Using district-level occurrences in MaxEnt for predicting the invasion potential of an exotic insect pest in India. *Computers and Electronics in Agriculture*, *103*, 55–62. doi:10.1016/j.compag.2014.02.007

Ladle, R. J., Jepson, P., Araújo, M. B., & Whittaker, R. J. (2004). Dangers of crying wolf over risk of extinctions. *Nature*, *428*(6985), 799. doi:10.1038/428799b

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385–393. doi:10.1111/j.0906-7590.2005.03957.x

Lobo, J. M., & Tognelli, M. F. (2011). Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, *19*(1), 1–7. doi:10.1016/j.jnc.2010.03.002

Manel, S., Dias, J.-M., & Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, *120*(2-3), 337–347. doi:10.1016/S0304-3800(99)00113-1

Margoluis, R., & Salafsky, N. (1998). Measures of success: designing, managing and monitoring conservation and development projects. *Island Press*.

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069. doi:10.1111/j.1600-0587.2013.07872.x

Nenzén, H. K., & Araújo, M. B. (2011). Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling*, *222*(18), 3346–3354. doi:10.1016/j.ecolmodel.2011.07.011

Osborne, P. E., Alonso, J. C., & Bryant, R. G. (2001). Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, *38*(2), 458–471. doi:10.1046/j.1365-2664.2001.00604.x

Park, F., Impacts, C., Zealand, N., & Change, C. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.

Pearson, R., Thuiller, W., Araujo, M., Martinez-Meyer, E., Brotons, L., McClean, C., … Lees, D. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, *33*(10), 1704–1711. doi:10.1111/j.1365-2699.2006.01460.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3-4), 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008a). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. doi:10.1111/j.0906-7590.2008.5203.x

Phillips, S. J., & Dudík, M. (2008b). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. doi:10.1111/j.0906-7590.2008.5203.x

Prates-Clark, C. D. C., Saatchi, S. S., & Agosti, D. (2008). Predicting geographical distribution models of high-value timber trees in the Amazon Basin using remotely sensed data. *Ecological Modelling*, *211*(3-4), 309–323. doi:10.1016/j.ecolmodel.2007.09.024

Reid, D. G., & Jien, G. (1999). Giant panda conservation action plan. In S. Christopher, H. Stephen, & P. Bernard (Eds.), *Bears* (pp. 241–254). IUCN,Gland,Switzerland and Cambridge,UK.

Rosenzweig, M. L. (1995). Species diversity in space and time. In L. R. Michael (Ed.), . The press sydnicate of the University of Cambridge.

Schaller, G. B. (1994). *The last panda. University of Chicago Press, Chicago* (Vol. 299). University of Chicago.

Senay, S. D., Worner, S. P., & Ikeda, T. (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PloS One*, *8*(8), e71218. doi:10.1371/journal.pone.0071218

State Forestry Administration of China. (2006). The Third National Survey Report on Giant Panda in China (in Chinese). *Beijing: Science Press*.

Stem, C., Margoluis, R., Salafsky, N., & Brown, M. (2005). Monitoring and Evaluation in Conservation: a Review of Trends and Approaches. *Conservation Biology*, *19*(2), 295–309. doi:10.1111/j.1523-1739.2005.00594.x

Stockwell, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, *13*(2), 143–158. doi:10.1080/136588199241391

Syfert, M. M., Smith, M. J., & Coomes, D. a. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS ONE*, *8*(2), e55158. doi:10.1371/journal.pone.0055158

Vale, C. G., Tarroso, P., & Brito, J. C. (2014). Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara-Sahel transition zone. *Diversity and Distributions*, *20*(1), 20–33. doi:10.1111/ddi.12115

Wang, S., & Xie, Y. (2004). China species red list. *China High Education Press*, *Vol.I*(Red list,Beijing).

Wiens, J. J., & Graham, C. H. (2005). Niche conservatism: Integrating Evolution, Ecology, and Conservation Biology. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 519–539. doi:10.1146/annurev.ecolsys.36.102803.095431

Wilson, K. A., Westphal, M. I., Possingham, H. P., & Elith, J. (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, *122*(1), 99–112. doi:10.1016/j.biocon.2004.07.004

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. doi:10.1111/j.1472-4642.2008.00482.x

APPENDIX

Important environmental variables for giant panda

As we know, doing research on effects of environmental variables is essential to giant panda conservation. However, this study does not focus on the analysis of environmental variables. Although, we can also achieve some useful information from MaxEnt running. The Jackknife test was applied to determine the relative importance of environmental variables for generating the models in MaxEnt (Prates-Clark etc., 2008). The contribution of environmental variables to the giant panda distribution is demonstrated in Figure. The Jackknife of regularized training gain figures below show how much better the MaxEnt distribution fits the presence data in different scenarios. In addition, a model was created using each variable in isolation to determine contribution of variables.

This study selected eight Jackknife figures which according to the five scenarios proved to have highest accuracy in Result Chapter. Figure 26 to Figure 33 tell us the environmental variable with highest gain when used in isolation wettest quarter (bio 8), max temperature of warmest month (bio 5), precipitation of driest quarter (bio 17), coldest month (bio 6), precipitation of coldest quarter (bio 19) and altitude in the eight scenarios with highest accuracy respectively. Those are the most important contributing to the predicted giant panda distribution according to this study.

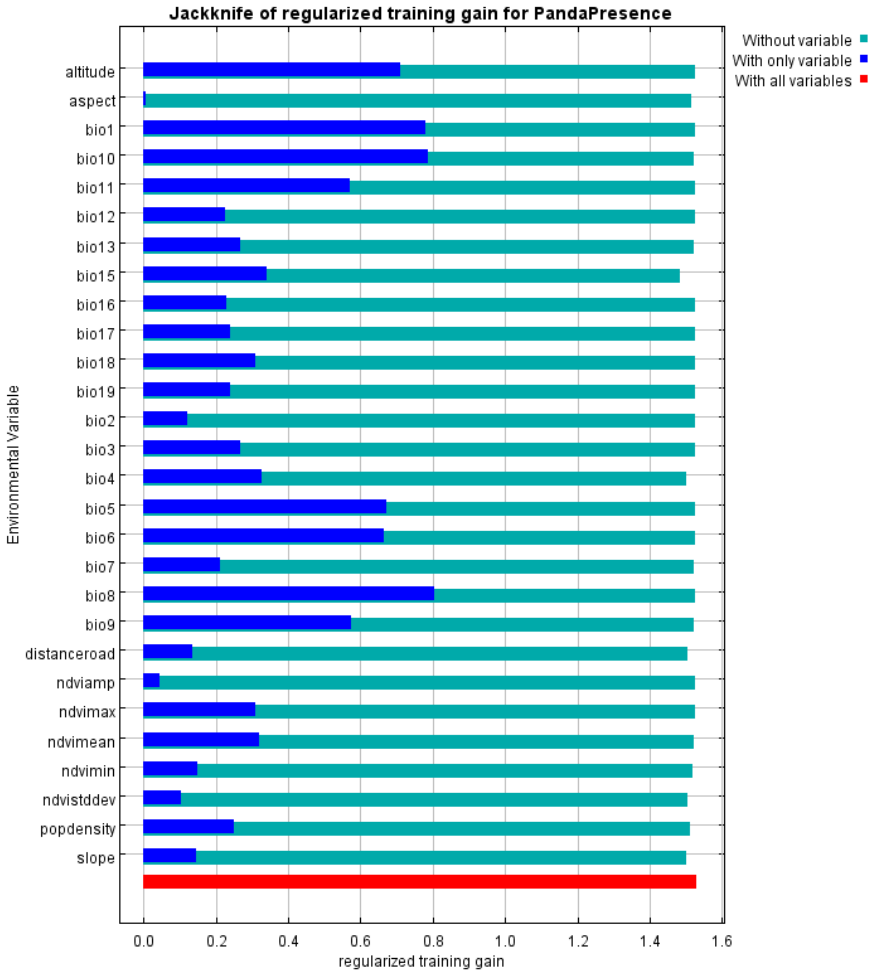


Figure 26. Importance of environmental variables in modelling the distribution at county level with 10% presences

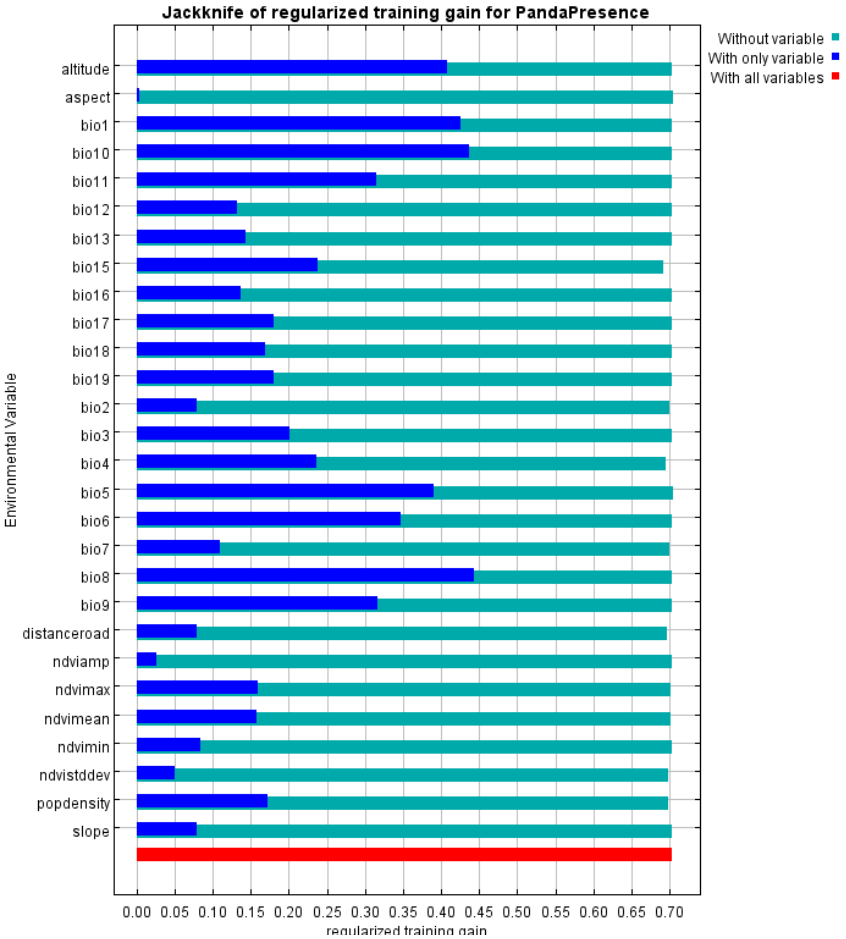


Figure 27. Importance of environmental variables in modelling the distribution at county level with full presences

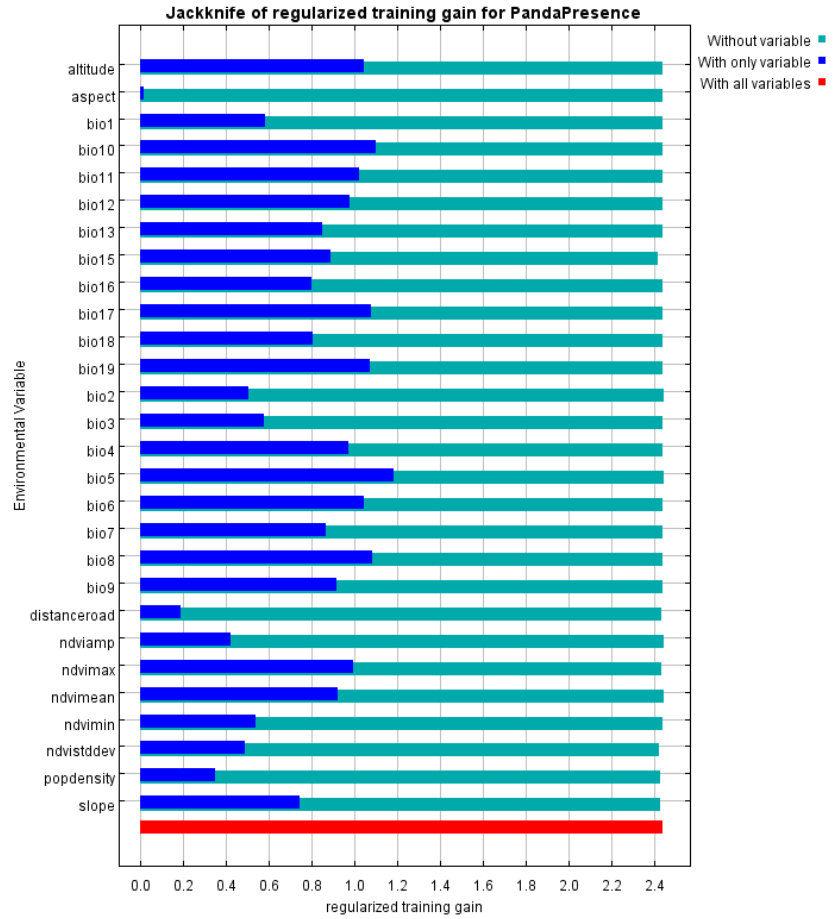


Figure 28. Importance of environmental variables in modelling the distribution at provincial level with 10% presences

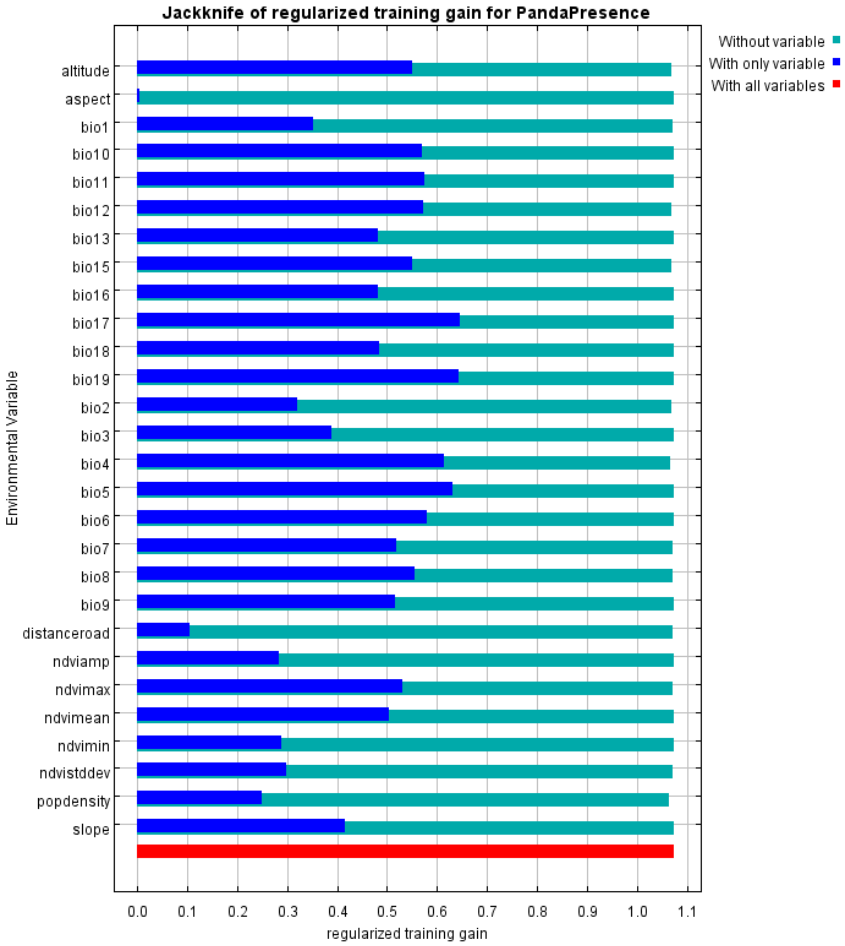


Figure 29. Importance of environmental variables in modelling the distribution at provincial level with full presences

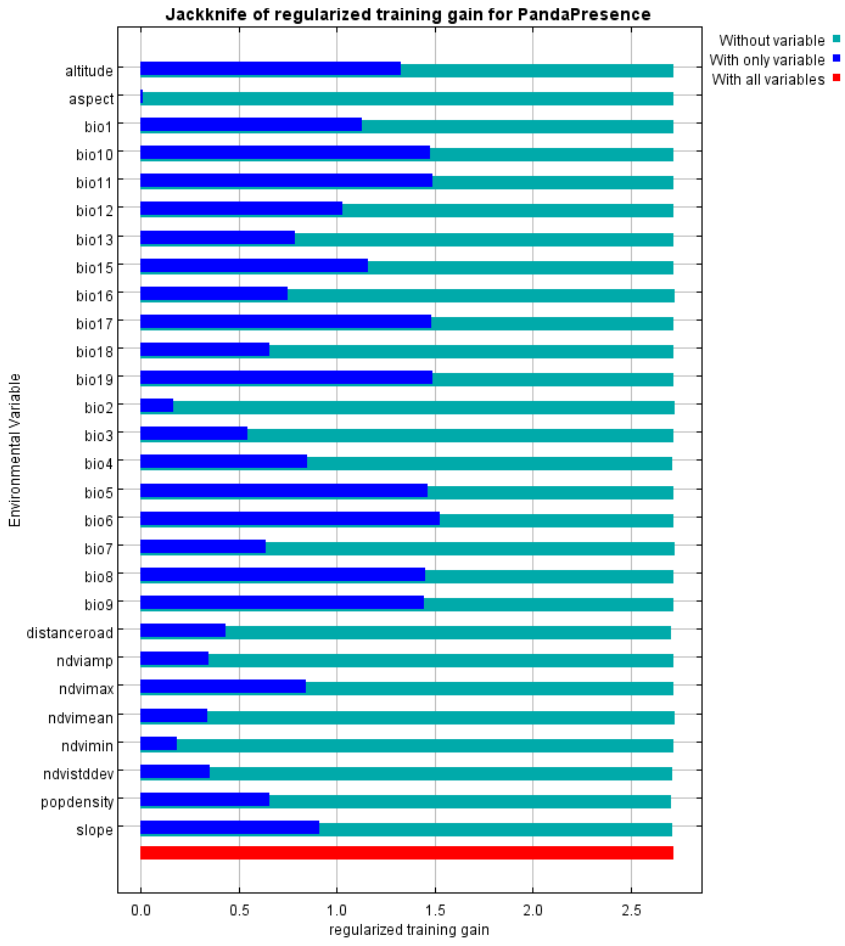


Figure 30. Importance of environmental variables in modelling the distribution at regional level with 10% presences

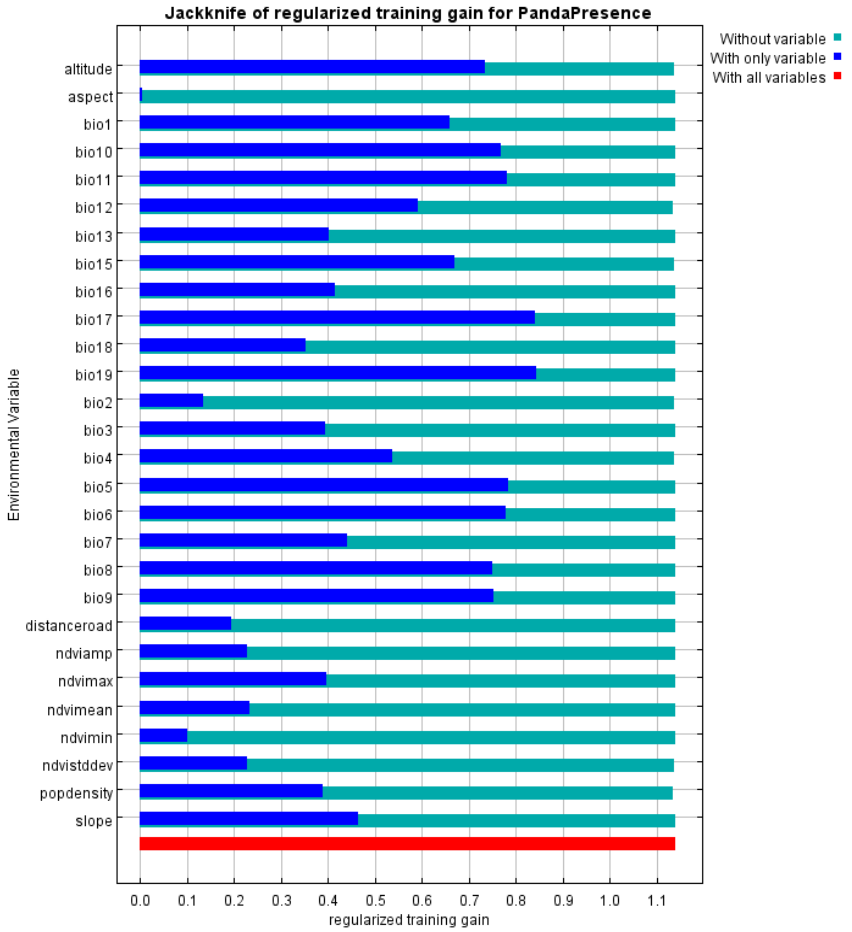


Figure 31. Importance of environmental variables in modelling the distribution at regional level with full presences

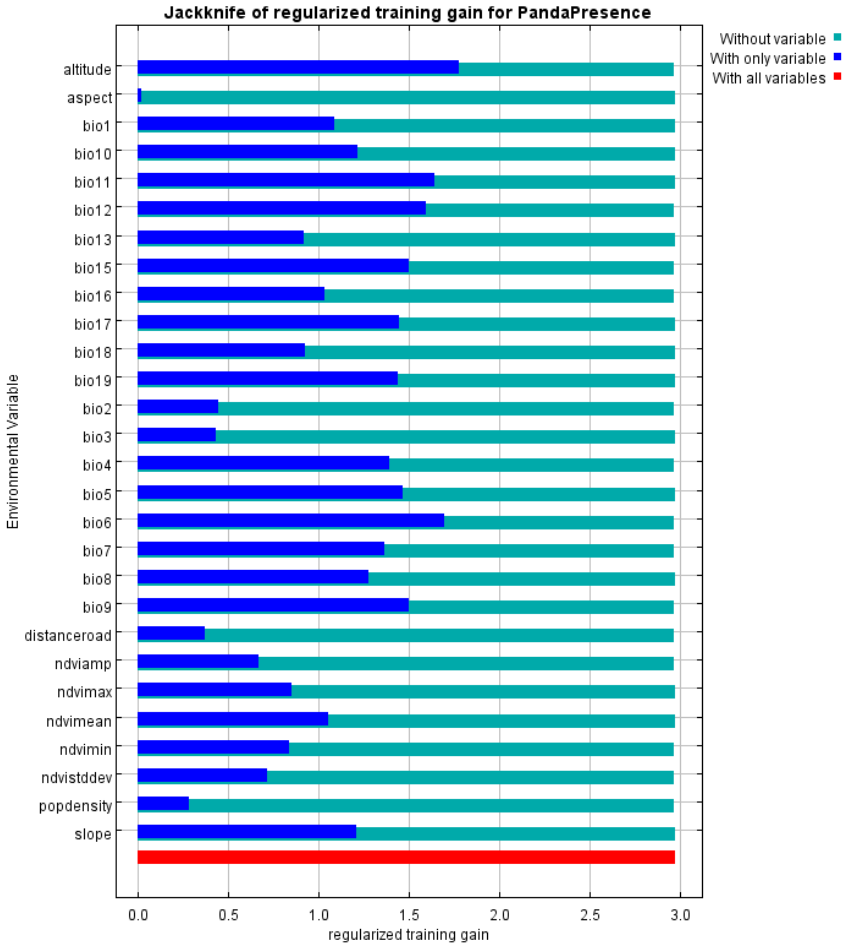


Figure 32. Importance of environmental variables in modelling the distribution at national level with 10% of presences

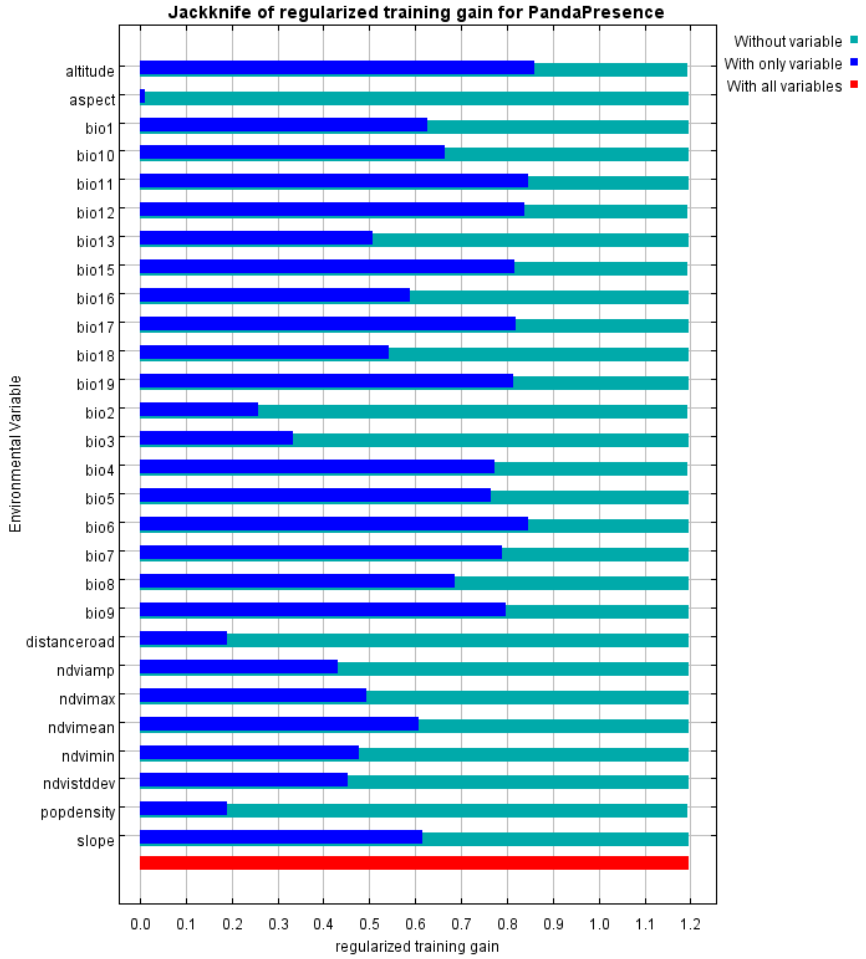


Figure 33. Importance of environmental variables in modelling the distribution at national level with full presences