Application of random forest algorithm for studying habitat selection of colonial herons and egrets in human-influenced landscapes

Luis Carrasco · Miyuki Mashiko · Yukihiko Toquenaga

Graduate School of Life and Environmental Sciences, University of Tsukuba, Ibaraki 305-8572, Japan

E-mail: luis@pe.ska.life.tsukuba.ac.jp

Tel: +81-20-853-6657

Fax: +81-20-853-6657

Abstract Understanding the mechanisms of habitat selection is fundamental to the construction of proper conservation and management plans for many avian species. Habitat changes caused by human beings increase the landscape complexity and thus the complexity of data available for explaining species distribution. New techniques that assume no linearity and capable to extrapolate the response variables across landscapes are needed for dealing with difficult relationships between habitat variables and distribution data. We used a random forest algorithm to study breeding-site selection of herons and egrets in a human-influenced landscape by analyzing land use around their colonies. We analyzed the importance of each land-use variable for different scales and its relationship to the probability of colony presence. We found that there exist two main spatial scales on which herons and egrets select their colony sites: medium scale (4 km) and large scale (10-15 km). Colonies were attracted to areas with large amounts of evergreen forests at the medium scale, whereas avoidance of high-density urban areas was important at the large scale. Previous studies used attractive factors, mainly foraging areas, to explain bird-colony distributions, but our study is the first to show the major importance of repellent factors at large scales. We believe that the newest non-linear methods, such as random forests, are needed when modelling complex variable interactions when organisms are distributed in complex landscapes. These methods could help to improve the conservation plans of those species threatened by the advance of highly human-influenced landscapes.

Keywords Breeding-site selection · Colonial birds · Habitat selection · Landscape ecology · Predictive models

Introduction

Understanding the mechanisms of habitat selection is fundamental to the construction of proper conservation and management plans for many avian species. Choosing breeding sites is a crucial task for avian species, but it is still not clear how they undertake it. Approximately 13% of birds breed in spatially packed colonies (Gill 2007). Colony site selection is a more difficult problem than choosing an individual nest site because the site selection affects the fate of all members of the colonies.

Identifying those scales at which certain distribution pattern occur can help to clarify what mechanisms are involved in habitat selection. However, many studies on colonial birds have been conducted at a single spatial scale, so results about colony site selection and explanations for the mechanisms involved have been widely diverse. Most of them used linear models or simple correlations between landscape variables and the presence of the species for explaining the colony distribution (Fasola and Canova 1991, Tourenq *et al.* 2004). The interaction between the explanatory variables are very intricate, and a high correlation among scales makes this analysis even more complicated, specially when studying mixed species colonies, when differential habitat selection among species could add more complexity.

Some authors created habitat suitability models for colonial birds (Kelly *et al.* 2008, Parkes *et al.* 2012), but their methods assumed linear responses between the dependent variable and the explanatory variables. All of these methods are generally appropriate when studying relatively simple variable interactions and when responses to the explanatory variables are linear. However, widely used methods such as logistic regression are often misapplied. In many cases, applying a logistic regression does not guarantee maximum-likelihood estimates and the odd ratios are not always proportional to the probability of presence of the species (Keating and Cherry 2004). For this reason, new methodologies that

can successfully incorporate non-linear and complex-variables' relationships are needed to analyze differences in site selection for each scale.

In the last decades, the high human impact on natural landscapes has challenged scientists to improve their predictive models in order to create effective conservation plans for bird species that share habitat with human beings. The complexity of the optimization problem in ecological models increases with spatial complexity (Seppelt and Voinov 2002), so including a higher fragmentation of agricultural landscapes may add difficulty to the analysis of the relationships between the habitat variables and the colony locations. Moreover, landscape complexity can affect the ability of the species to assess the habitat and for the detection of resources (Wiens and Milne 1989). Furthermore, for some agricultural landscapes affected by urban development, the explanatory data are too complex, and it is necessary to use other techniques without assuming linearity, such as classification trees or machine learning methods. These techniques are better tools for extrapolating the response variables across landscapes and for analyzing the importance of the predictors than are other methods such as linear regressions (Prasad *et al.* 2006). The random forest (RF) technique (Breiman 2001) does not need to assume linearity. It allows for the modelling of complex interactions among predictor variables and is becoming widely used due to its predictive power (in comparison with normal decision trees) and its capacity to measure variable importance (Cutler *et al.* 2007).

Our objective was to detect the factors that affect, at different scales, breeding site selection of colonial birds in a human-influenced landscape. Japan is a good example of a highly human-influenced complex landscape where we can still find birds breeding in mixed-species colonies, and where we can obtain precise data of land uses and breeding locations distribution. We used location data for heron and egret colonies distributed in the fairly complex agricultural landscape of Ibaraki and surrounding prefectures in Japan in 2011, and compared the land types surrounding the colonies with those around unoccupied

sites using geographic information systems (GIS) techniques. Then we applied a RF algorithm to analyze the importance of the different land-use variables at different scales for establishing a colony. Our results showed that there were two main scales at which herons and egrets select colony sites: medium scale (4 km) and large scale (10-15 km). Evergreen forest was the most important land type for explaining colony distribution at the medium scale. Conversely, disturbance factors, such as urban areas, were determining factors for colony locations at larger scales. Foraging habitats were revealed to be unimportant predictors at all scales, especially at the smaller scales.

Methods

Study area and species

The study area was the central and southern regions of Ibaraki Prefecture and some bordering regions of Tochigi, Gunma, Saitama and Chiba prefectures in central Japan (Fig. 1). The region is limited by mountains to the north-west, by the Pacific Ocean to the east and by the Tone River to the south, with a total area of approximately 10 022 km. It is mainly a low altitude plain and its main geological feature is the presence of Lake Kasumigaura. The predominant human-influenced land use is agricultural, rice fields being the dominant cultivation (8.5% of the study area). There are residential areas of various sizes and forest patches spread all around the region. Six species of herons and egrets, Grey Heron (*Ardea cinerea*), Great Egret (*A. alba*), Little Egret (*Egretta garzetta*), Intermediate Egret (*E. intermedia*), Cattle Egret (*Bubulcus ibis*) and Black-crowned Night Heron (*Nycticorax nycticorax*), breed mainly in mixed-species colonies every year in the study area. They build their nests on conifers, broad-leaf trees and in bamboo thickets (Environmental Agency of Japan 1994).

Colony locations

Twenty colony locations were recorded by ground surveys (Mashiko and Toquenaga 2013) in the study area during the breeding season, from March to August, of 2011. In our study, we aimed to analyze the land use surrounding the colonies, so we referred to the colony data for 2011 due to the limited availability of land-use maps with sufficient resolution up to this year. The site selection model for this study was based on differences in the areas surrounding colonies and the those surrounding unoccupied sites. For the statistical model to be consistent, we needed to compare the same number of colonies and unoccupied sites, so we randomly chose 20 locations, which corresponds with the number of observed colonies in this study, where a colony could, potentially, be formed. Locations available for colonization were defined as follows. First, a rectangular area of the study area was arbitrarily delimited (35°52’32”N - 36°35’43”N, 139°35’36”E - 141°00’00”E). Second, forest areas below an altitude of 100 m were selected, as colonies are seldom found at higher elevations in this area of Japan (Fig. 1). Unoccupied sites were then randomly selected among the potentially available regions. Because the random selection of sites could lead to slightly different results, we created 30 different data sets of 20 points each, and analyzed the data for each set.

Landscape variables

To analyze the information on land use in the areas surrounding the colonies and unoccupied sites, we used a land-use map of Japan provided by the Japanese Aerospace Exploration Agency (JAXA). This map was created with multi-satellite imagery from 2011. The final map had an approximately 45-m pixel size. The processing and classification details are explained at http://www.eorc.jaxa.jp/ALOS/lulc/lulc_jindex.htm.

Eight relevant land-use variables for herons and egrets were identified as follows: bare land, evergreen forest, deciduous forest, grassland, crop land, paddy field, urban area and body of water (Fig. 2). The selection of the variables was based on previous knowledge of the ecology of the heron species (Tojo 1996, Lane and Fujioka 1998) and examination of the satellite images. Layers of circular buffer zones were created around the colonies and the 20 randomly selected points, and areas of the eight selected land-use variables were identified within each. The radii of the buffer zones were 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, and 40 km, distances that cover the inter-distance of colonies with a high resolution as well as areas at coarser scales (Fig. 1). The raster package (Hijmans and van Etten 2012) in R 2.15.2 (R Development Core Team 2011) was used to extract the land-use information.

Statistical analyses

Random forest model

A colony site selection model was created using a random forests (RF) algorithm (Breiman 2001), which relies on the ideas of classification and regression trees (CART) (Breiman *et al.* 1984), and on bagging methods (Breiman 1996). For input data against models, we used the area of each landscape variable surrounding the 20 colonies and the 20 randomly chosen unoccupied sites. One model for each scale around the colonies (each buffer radius) was built. For example, for creating the 3-km scale model the eight land use variables measured at a buffer radius of 3 km were used. Randomized models were created to compare the predictive power of our colony distribution models with the predictive power of the models when the outputs of the training sets (presence and absences) were randomly permuted. One third of the data was left out for each bootstrap sample (the out-of-bag data, OOB) for each tree and 500 trees were created for each classification model. The process was repeated for each of the 30 random points data sets.

Predictive accuracies

We calculated the accuracy of the model as 1 − OOB error estimate (Breiman 2001). The results of the RF model can be slightly different each time it is performed, even when using the same parameters, so ten models were built for each scale. The mean value of the OOB error for the ten repetitions was used to measure the accuracy for one model. Disparate accuracy results are also obtained for the randomized models for different permutations of the presence-absence values, so the mean value of the accuracy out of 100 repetitions was used. The prediction accuracy of models tended to vary among random sets of hypothetical unoccupied colony locations. Each random set of hypothetical colonies plays a role similar to that of the data for supervised learning in a neural network model. Bad and good data could, respectively, cause low and high accuracies. So we calculated not only means but also maximum accuracies of the RF models for each scale. Mean accuracies represent the overall tendency of model performance across the scales, but the best performance of a RF model at each scale should be evaluated by the maximum accuracy values. This is because we wanted to use the set that best explained the colony distribution, not being interested in one "average model" for analyzing the important explanatory variables.

Variable importance

For the scales with the best accuracy models, the importance of each land use variable was analyzed using the mean decrease accuracy index (Breiman 2001). To do this, we selected the data-set with the maximum accuracy for a specific scale among all sets. As we found some variance each time the RF was constructed, the average of the mean decrease accuracy index over ten models was used.

Variable effect on colony presence

Partial dependence plots (Friedman 2001, Hastie *et al.* 2005) were used to graphically characterize relationships between individual predictor variables and predicted probabilities of colony presence (Cutler *et al.* 2007). The vertical axis of this plot is a measure of the marginal effect of a certain explanatory variable on the class probability. In our study, the vertical axis represents the effect of the area of each land use on the probability of colony presence. The horizontal axis represents the value of the variable for which partial dependence is sought. We interpreted that a colony "preferred" certain land use when the effect on the probability of colony presence was higher for higher areas of that land type. A colony, therefore, "avoided" a land type when the effect on the probability of colony presence decreased as that land-use area increased. The randomForest package (Liaw and Wiener 2002) in R was used for model creation and analysis.

Results

The accuracy values of the randomized models strongly depended on the random permutation of the output values (colony presence or absence) but, on average, all the scales showed an accuracy of between 47% and 48% (Figure 3). The 4-km scale model had the best accuracy on average, followed by the 15-, 30-, 10- and 1-km scales. The 4-km model had a relatively high accuracy for all the sets but, in most cases, it was not the scale with the best accuracy within the set. In contrast, 10-km or 15-km scales performed the best in many of the sets.

The shape of the maximum accuracies graph was quite similar to that of the mean accuracies graph, but the 4-km models did not have the maximum values for accuracy (Figure 3). The scale with the highest maximum accuracy was the 10-km scale, with almost 78% accuracy, 30% more accurate than its equivalent randomized model.

Some scales were ineffectual for explaining colony distribution, even when nearby scales were important. This was the case for the 0.5-, 2-, 5-, 6- and 7-km scales, showing low mean accuracy levels (Figure 3). Variable importance was analyzed for the sets that showed maximum accuracies for the 4-, 10- and 15-km scale models (Figure 4). Evergreen forest was the most important variable for explaining colony distribution at the 4-km scale, followed by urban and crop areas (Figure 4A). The importance of the evergreen forests decreased as scale size increased, being very low for 15 km. For the 10- and 15-km scales, urban areas and bare land were the most important variables for the model (Figure 4B and C). Paddy fields were revealed to have low importance for the model, although their importance increased with scale.

Partial dependence plots of each landscape variable were very similar for the different scales, so we chose the most important ones of the best explanatory scales to analyze the relationship between the area of each landscape predictor and the probability of colony presence. Evergreen forest was an attractive land type for establishing a colony (Figure 5). For the 4-km scale, regions made up of less than 7% evergreen forest were strongly avoided. On the other hand, areas made up of more than 35% evergreen forest were neither attractive or repellant. We used the 10-km model to analyze the importance of urban areas, as this variable was the most important at this scale. Regions made up of at least 10% urban areas were strongly avoided (Figure 5). Colonies tended to be established where urban areas made up between 5% and 10% of the area within the 10-km radius. Bare soil produced similar results to those of urban land types. Crops and paddy fields were attractive land types when they were relatively important variables in some models.

Discussion

We obtained highly accurate colony site selection models for 4-, 10- and 15-km scales. At the 4-km scale, evergreen forest was the most important variable, being an attractive factor. At the 10- and 15-km scales, urban areas and bare land were the main variables for explaining the models, both of which were avoided when they were present in high ratios. Paddy fields, the main foraging habitat for all the species, was not a high-importance variable for any of the scales, although its importance increased with scale. The highest accuracy obtained was 78% for the 10-km model.

It was revealed that there are two very distinct general scales by which site selection is most affected: the 4-km range (medium scale) and the 10-15 km range (large scale). The 4-km scale had the highest average accuracy for all sets, the 10-km scale had the highest maximum accuracy values and the 15-km scale had high average and maximum values. The 30-km scale also exhibited good performance on average and when analyzing the maximum accuracy. The variables that best explained the distribution of the colonies at 30 km were very similar to those at 15 km, so we consider this scale to be highly correlated with the 15-km scale. Herons and egrets could be mainly using these two scales to decide where to establish their colonies. Scales in between these two and also small scales were of very low importance for colony site selection.

Scale dependence could explain the diversity of results obtained in previous studies on the accuracy of various models used for predicting colony sites of herons and egrets, and could also provide a different explanation for which factors affect colony distribution. For example, Gibbs and Kinkel (1997) used a 15-km scale to explain colony distribution. On the other hand, Fasola and Alieri (1992) and Tourenq *et al.* (2004) used a 5-km scale, while Boisteau and Marion (2007) used 25 km. In these cases, foraging habitats were used as predictors, so the scales were justified by the observed foraging ranges. Our study reveals that studying one single scale could lead to models with low predictive power, and with

potentially fatal consequences when erroneously considering explanatory variables as key factors, while other variables could be much more important at different scales.

Most of the past studies on herons and egrets analyzed areas not highly inhabited by humans. Scale is a crucial factor when studying colony site selection, and it could depend strongly on the landscape configuration. Among the studies that have considered different scales, Kelly *et al.* (2008) showed that the 1-km scale was best for explaining the colony distribution of herons and egrets in tidal marshes. Parkes *et al.* (2012) also found that small scale (below 1.5 km) was very important for the sites of Cattle Egret colonies in upland residential areas (although their model did not consider interaction between the explanatory variables). Bigger scales (from 1 to 10 km), however, were more important in rice field related colonies in France (Tourenq *et al.* 2004), depending on the study species. The scale at which the colony site selection is performed could strongly depend on the land types surrounding the colonies. Landscape configuration and different levels of fragmentation of important habitats could be crucial factors for determining which scales allow the species to assess the surroundings and choose optimal colony sites.

Evergreen forest, an attractive land type for herons and egrets, was the most important variable at the medium scale, while repelling factors such as urban areas and bare land were most important at large scales. Evergreen forest includes bamboo thickets, the most important nest substrate for herons and egrets, followed by trees, which explains the higher importance of this land use over deciduous forest. Identifying regions with greater amounts of bamboo thickets, places available for the establishment of a colony, could be one of the most important steps of habitat assessment after the arrival of the individuals. The importance of the medium scale for colony site selection might reflect an effect of the study area landscape patterns, where high densities of evergreen forest could be easily detected at the 4-km scale. At different scales, the distribution patterns of this land use could make the habitat assessment more difficult for herons and egrets, being unable to discriminate high

density forest regions at small or large scales. Colonies demonstrated, however, a preference for lower urban- and bare-area densities (the latter being highly related to human-influenced land types) at large scales. When food and forest availability does not determine colony locations, mechanisms such as avoiding disturbances become more important. There is evidence of heron and egret colonies avoiding urban areas at small scales (Fasola and Alieri 1992) but our current study is the first study to show this effect on the distribution of colonies at large scales. Avoiding large urban areas could be advantageous in terms of lower levels of disturbances for the colonies. In Saitama Prefecture, located in the south-west of our study area and in the northern suburbs of Tokyo, there has been a great deal of urban development since 1960; population sizes of herons and egrets have decreased and some colonies have disappeared, even where paddy fields and forest patches that are available for the establishment of colonies remain (Narusue 1992). The effect of the high density of urban areas at the large scale could have been the main reason for the extinction of heron and egret colonies in central Japan.

Paddy fields, the main foraging areas for the study species, was revealed to be unimportant for predicting colony site, in contrast to results of past studies on herons and egrets. However, its importance seemed to increase with scale. Previous studies showed weak relationships between food habitats and the colony distribution of these species (Fasola and Alieri 1992, Boisteau and Marion 2007) in agricultural landscapes, but they included only attractive factors, and no repelling factors, in their models. Fasola and Canova (1991) had the same results for mixed-species colonies of gulls and terns, where foraging sites were not an important factor in colony location, allowing us to infer that this could also be true for other wading bird families. Landscape complexity in developed urban regions, as exist in the present study area, could lead to difficulties in the assessment of the quality of food habitats for herons and egrets. Difficulties for many species on the evaluation of food availability for the breeding period have been widely discussed (Orians

and Wittenberger 1991, Fuller 2012). Also, the capacity for evaluating the amount and quality of foraging habitats could be damaged by surrounding urban landscapes (Battin and Lawler 2006). Our study area is highly affected by urban development and some regions are experiencing very rapid landscape changes, so even resident species could have problems assessing habitat quality, despite that their evaluation process continues even during non-breeding seasons.

The study of positively and negatively associated factors on different scales revealed two main characteristics of colony distribution of herons and egrets: colony sites were established near large amounts of evergreen forests at medium-scale distances and where there are less urban areas at higher-scale distances. Including avoidance factors could improve the performance of predictive bird colony distribution models, especially when analyzing at large scales. Conservation and management of colonial wading birds living in human-influenced landscapes should not only focus on the maintenance of the available foraging and breeding habitats, but also on controlling urban development around the colonies.

In the present study, the predictive accuracy for the best models was higher than that of a predictive colony distribution model for mixed-species herons and egrets by Kelly *et al.* (2008) and similar to the predictive model for Cattle Egret by Parkes *et al.* (2012). This shows that the RF method can be a good tool for predicting the colony sites of herons and egrets, and that it can handle the complexity of human-influenced landscapes. However, the predictive power of the RF model was lower than that of those using other colonial but non-wading and single-species data (Bustamante 1997, Lauver *et al.* 2002, Heinänen *et al.* 2008). This is probably due to the relative simplicity of those data, where linear statistical tools could be sufficient for explaining the distribution data.

The use of GIS in combination with the newest classification techniques, such as RF, seems to be an appropriate analysis method for complex ecological data that includes the

complexities of human-influenced agricultural landscapes. Logistic regressions are often powerless, especially with complicated land-use patterns. Resultant logistic models are often very specific to data sets, and the same model cannot be applied to other similar data sets. Neural networks and decision trees are alternatives for such complicated land use patterns, but our approach of using RF has two prominent advantages against neural networks and other decision trees. The first is that we can evaluate the relative importance of competing variables, or land-use types. The second is that RF can avoid over specialization and remain generalized for similar problems (Breiman 2001).

In summary, we applied a RF algorithm for analyzing the distribution of herons and egrets colonies in a strongly human-influenced landscape in Japan, and we were able to clarify some important characteristics of the colony site selection strategies of these species. We strongly believe that non-linear methods as RF are more appropiate when dealing with predictive suitability models for birds living in highly human-influenced landscapes than classic linear methods. These methodologies could be a big help to rethink and improve the conservation plans of those species threatened by the advance of the agricultural and urban landscapes.

Acknowledgements We thank S. Ikeno, M. Seido, and K. Takeda for supplying information about the location of some colonies. We also thank K. Ohashi and members of the Population Ecology laboratory for helpful discussions. This study was supported in part by Grant-in-Aids for Scientific Research (13740433 and 19570014) to YT from the MEXT and JSPS. Additional financial support was provided through a Monbukagakusho scholarship to L. Carrasco from MEXT.

References

Battin J, Lawler JJ (2006) Cross-scale correlations and the design and analysis of avian habitat selection studies. Condor 108:59–70

Boisteau B, Marion L (2007) Habitat use by the grey heron (*Ardea cinerea*) in eastern France. Comptes rendus biologies 330:629–34

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth International Group, Belmong California

Bustamante J (1997) Predictive models for lesser krestel *Falco naumanni* distribution, abundance and extintion in southern Spain. Biol Conserv 80:153–160

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

Environmental Agency of Japan (1994) Distribution and population status of colonies and communal roosts of 22 bird species from 1990 to 1992. Wild Bird Society of Japan and the Environmental Agency of Japan, Tokyo

Fasola M, Alieri R (1992) Conservation of heronry Ardeidae sites in North Italian agricultural landscapes. Biol Conserv 62:219–228

Fasola M, Canova L (1991) Colony site selection by eight species of gulls and terns breeding in the <<Valli di Comacchio>> (Italy). Italian Journal of Zoology 658:261–266

Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Fuller RJ (2012) Birds and habitat: relationships in changing landscapes. Cambridge University Press, Cambridge

Gibbs J, Kinkel L (1997) Determinants of the size and location of great blue heron colonies. Colonial Waterbirds 20:1–7

Gill F (2007) Ornithology. Freeman and Company, New York

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27:88–85

Heinänen S, Rönkä M, Numers MV (2008) Modelling the occurrence and abundance of a colonial species, the arctic tern *Sterna paradisaea* in the archipelago of SW Finland. Ecography 31:601–611

Hijmans RJ, van Etten J (2012) Raster: geographic analysis and modeling with raster data. R package version 2.1-25

Keating K, Cherry S (2004) Use and interpretation of logistic regression in habitat selection studies. J Wild Manage 68:774–789

Kelly J, Stralberg D, Etienne K, McCaustland M (2008) Landscape influence on the quality of heron and egret colony sites. Wetlands 28:257–275

Lane S, Fujioka M (1998) The impact of changes in irrigation practices on the distribution of foraging egrets and herons (Ardeidae) in the rice fields of central Japan. Biol Conserv 83:221–230

Lauver CL, Busby WH, Whistler JL (2002) Testing a GIS model of habitat suitability for a declining grassland bird. Environmental management 30:88–97

Liaw A, Wiener M (2002) Classification and regression by randomForest. R news 2:18–22

Mashiko M, Toquenaga Y (2013) Increasing variation in population size and species composition ratio in mixed-species heron colonies in Japan. Forktail 29:71-77

Narusue M (1992) Changes in the distribution and extent of breeding colonies of egrets in Saitama Prefecture. Strix 11:189–209

Orians G, Wittenberger J (1991) Spatial and temporal scales in habitat selection. Am Nat 137:S29–S49

Parkes ML, Mora MA, Feagin RA (2012) Using scale, cover type and GIS to evaluate nuisance cattle egret colony site selection. Waterbirds 35:56–63

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Seppelt R, Voinov A (2002) Optimization methodology for land use patterns using spatially explicit landscape models. Ecol Modell 151:125–142

Tojo H (1996) Habitat selection, foraging behavior and prey of five heron species in Japan. Jap J Ornithol 45:141–158

Tourenq C, Benhamou S, Sadoul N, Sandoz A, Mesleard F, Martin J, Hafner H (2004) Spatial relationships between tree-nesting heron colonies and rice fields in the Camargue, France. Auk 121:193–202

Wiens JA, Milne BT (1989) Scaling of 'landscapes' in landscape ecology, or, landscape ecology from a beetle's perspective. Landsc Ecol 3:87–96

FIGURE LEGENDS

Fig. 1 Locations of heron and egret colonies in 2011 in our study area. Each dot represents a colony location (20 colonies in total). Mean nearest neighbor distance between colonies was 9.97 km. Grey regions show an altitude greater than 100 m where the distribution of herons and egrets is much lower.

Fig. 2 Land-use map of the study area. Map provided by JAXA combining ALOS satellite imagery and ground surveys from 2011. BL: bare land, EF: evergreen forest, DF: deciduous forest, GL: grassland, CL: crop land, PF: paddy field, UA: urban area, and WB body of water.

Fig. 3 Model accuracies for each scale. The mean graph represents the average accuracy for all of the 30 random-points datasets. The maximum graph considers only the random set that provide the maximum accuracy for each scale. The randomized accuracy graph represents the average accuracy of the randomized models for all 30 random-points data sets.

Fig. 4 Variable importance of each land use for three scale models (4, 10 and 15 km) represented by the mean decrease accuracy index. Error bars represent a 95% confidence interval. EF: evergreen forest, UA: urban areas, CL: crop land, BL: bare land, DF: deciduous forest, PF: paddy field, WB: body of water, and GL grassland.

Fig. 5 Partial dependence plot for evergreen-forest land use for the 4-km scale model and urban land use for the 10-km scale model.

FIGURE 1:

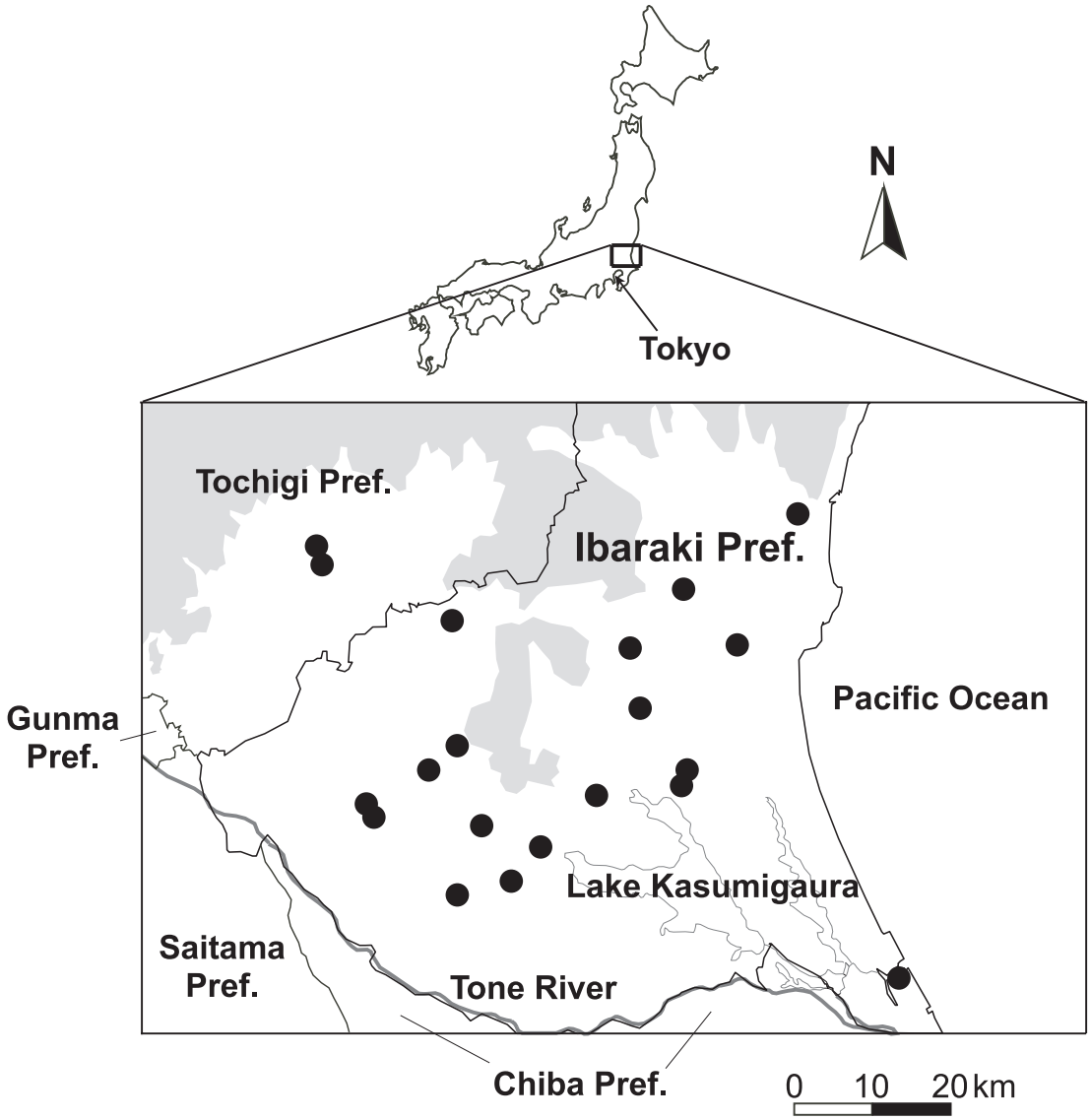
N
Tokyo
Tochigi Pref.
Ibaraki Pref.
Pacific Ocean
Gunma
Pref.
Lake Kasumigaura
Saitama
Pref.
Tone River
Chiba Pref.
0 10 20 km

FIGURE 2:

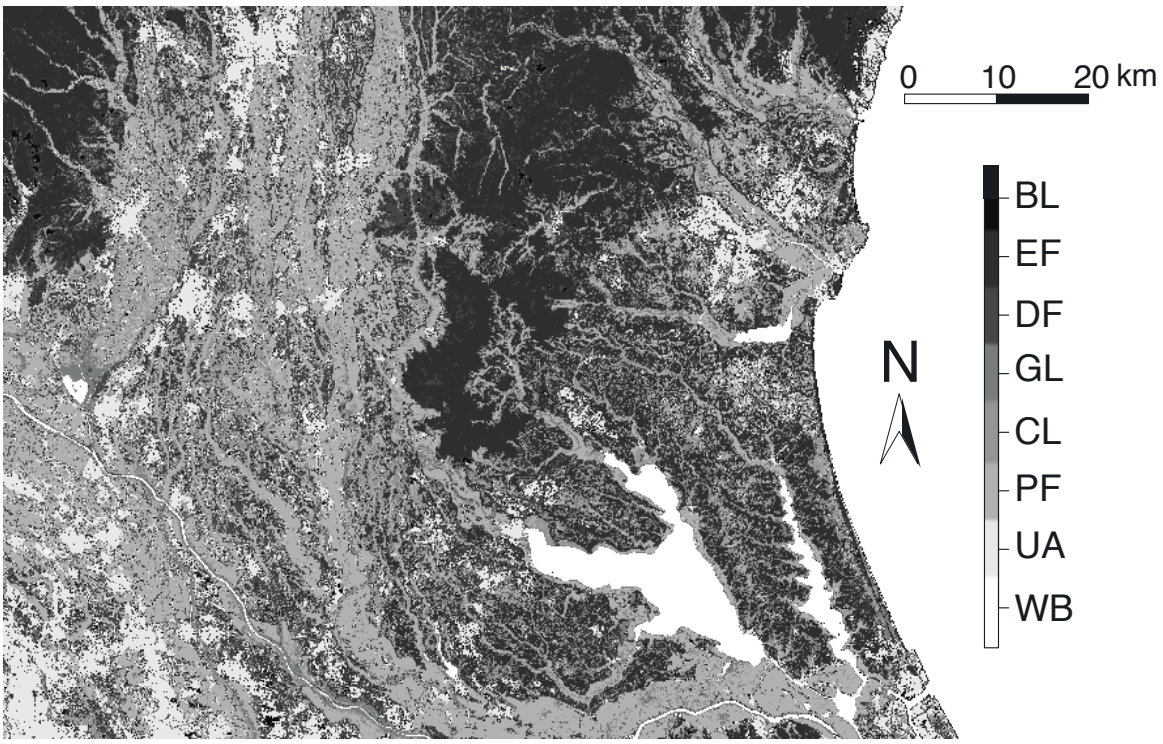

FIGURE 3:

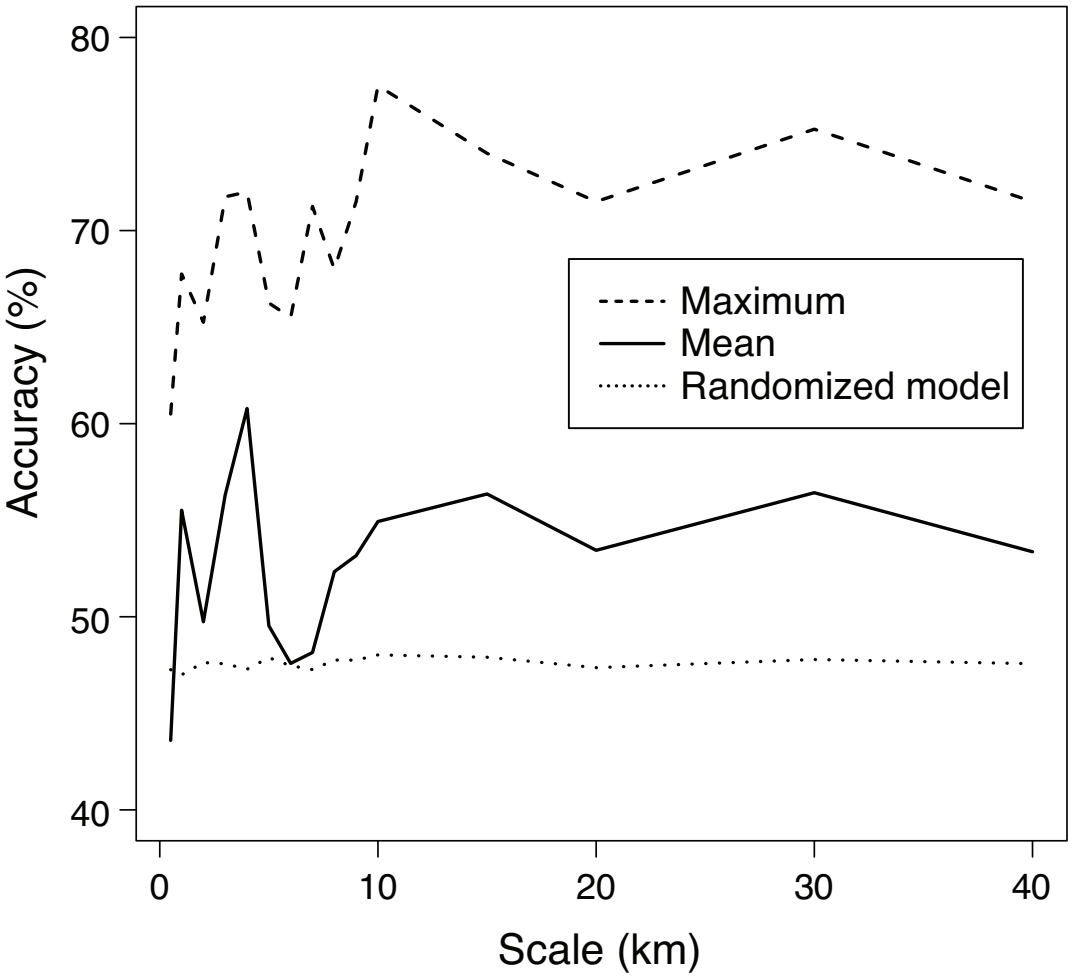
80
70
60
50
40
Accuracy (%)
0
10
20
30
40
Scale (km)
Maximum
Mean
Randomized model

FIGURE 4:

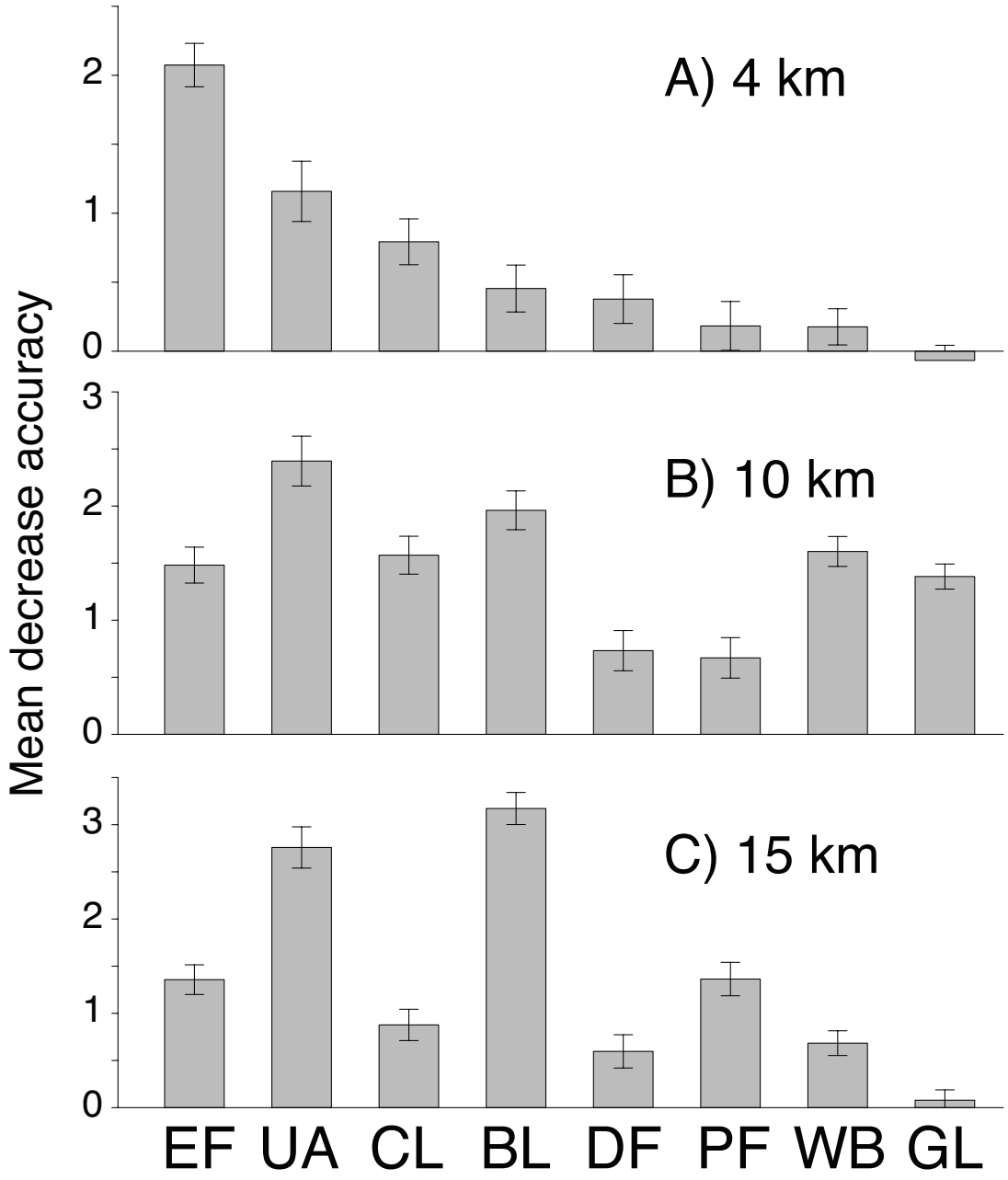
Mean decrease accuracy
A) 4 km
B) 10 km
C) 15 km
0
1
2
3
EF
UA
CL
BL
DF
PF
WB
GL

FIGURE 5:

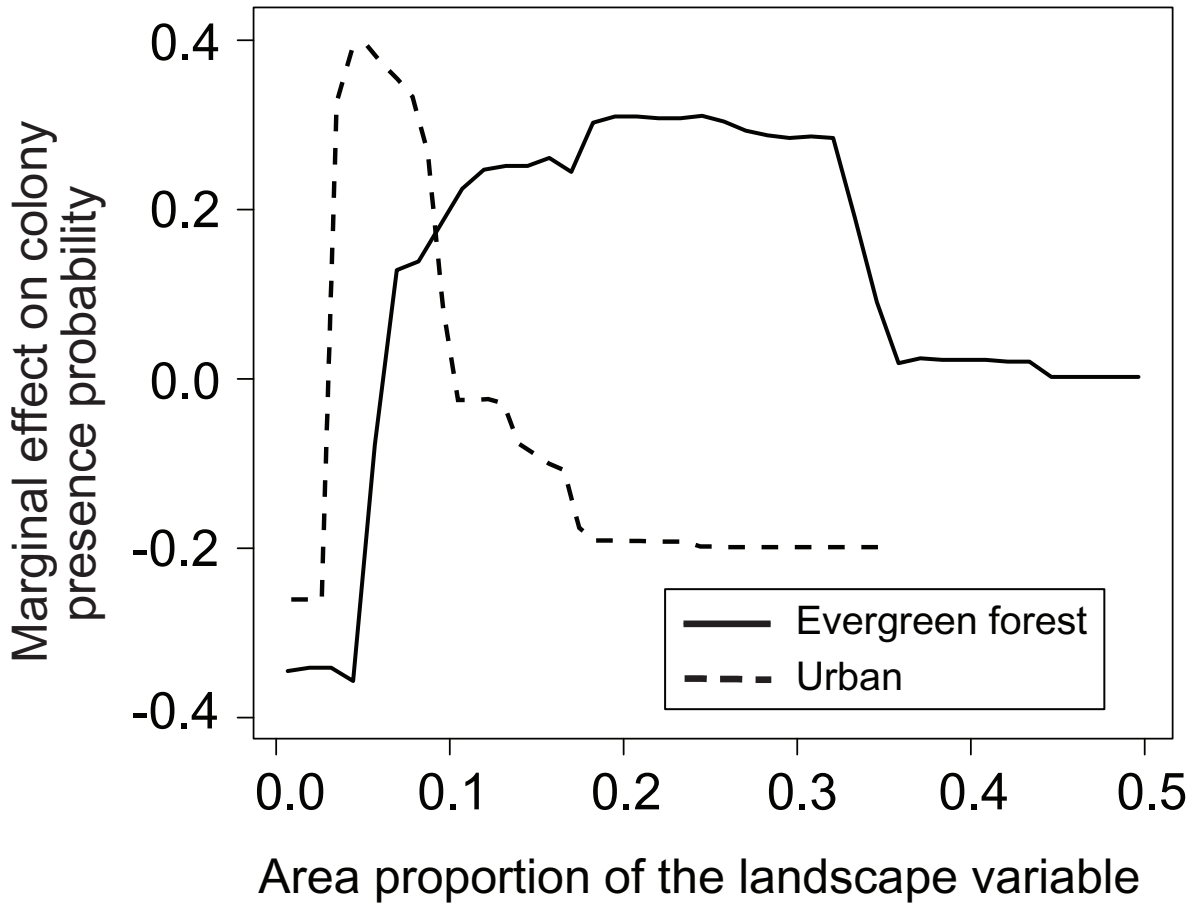
Marginal effect on colony presence probability
0.4
0.2
0.0
-0.2
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
Area proportion of the landscape variable
Evergreen forest
Urban

Table 1 Proportion of each land use (%) of the whole study area and of the average among all buffers surrounding the colonies for the two most important scales

Land use	Study Area	4-km Scale	10-km Scale
Body of water	3.8	2.7	4.9
Urban areas	11.7	11.2	8.8
Paddy field	22.8	21.8	22.7
Crop land	13.2	17.2	16.0
Grassland	6.7	6.3	6.7
Decidious forest	17.1	20.5	17.9
Evergreen forest	22.7	18.0	20.5
Bare land	2.1	2.2	1.9

Table 2 Correlation (Pearson's coefficient) between land use variables among all buffers of 4-km scale

	Urban areas	Paddy field	Crop land	Grassland	Decidious forest	Evergreen forest	Bare land
Body of water	-0.01	-0.46	-0.30	-0.18	-0.52	-0.06	0.42
Urban areas		0.07	-0.30	-0.43	0.01	-0.66	**0.71**
Paddy field			0.28	-0.11	0.06	-0.54	-0.15
Crop land				0.12	0.17	-0.19	-0.26
Grassland					-0.25	0.43	-0.38
Decidious forest						-0.09	-0.13
Evergreen forest							-0.58

Values higher than 0.7 are shown in *bold*.

Table 3 Correlation (Pearson's coefficient) between land use variables among all buffers of 10-km scale

	Urban areas	Paddy field	Crop land	Grassland	Decidious forest	Evergreen forest	Bare land
Body of water	-0.18	-0.44	-0.38	-0.60	**-0.77**	-0.06	0.13
Urban areas		0.02	-0.21	-0.47	-0.06	-0.61	**0.81**
Paddy field			0.42	-0.10	0.43	-0.53	0.07
Crop land				0.26	0.54	-0.26	-0.09
Grassland					-0.02	0.32	-0.46
Decidious forest						-0.10	-0.03
Evergreen forest							**-0.75**

Values higher than 0.7 are shown in *bold*.