GEOG71922 Assignment 2:

Species distribution modelling for Sciurus vulgaris

11155827

Word count: 1809

1. Introduction

This study is to develop a species distribution model for Sciurus vulgaris (red squirrel) in terms of scale and inter-specific interactions. A species distribution model refers to a tool to predict the distribution of a specific species based on occurrence observations (Elith and Leathwick, 2009).

Meanwhile, scale in spatial analysis can be referred to as the artificially defined spatial extent rather than the ecological processes (Jelinski and Wu, 1996). The term 'scale' has different concepts according to discipline. However, for this study, it can be considered a spatial resolution that affects the granularity of the data, leading to different analysis results (Sheppard and Mcmaster, 2008). The scale or spatial resolution that brings the highest model accuracy will be chosen for this study and be used for model evaluation, tuning, and selection.

Inter-specific interactions or associations (ISA) can provide a comprehensive understanding of species distribution by considering co-occurrences, segregations or attractions among species (Keil et al., 2021). The spread of grey squirrels is considered to be one reason for the decline in the number of red squirrels (Flaherty et al., 2012). Therefore, the occurrence data of grey squirrels will be used for the study.

The study area extends to about 10,000km2 around the Nethy Bridge in the Scottish Highlands (57°14'54.6 "N 3°36'22.7 "W). The Land Cover Map 2015 (LCM2015) from CEH (2017) provides the land cover type information as the base map. The National Biodiversity Network Atlas provides the occurrence data for red and grey squirrels. The data includes the year, verification status, and coordinates of the records.

The models are set to be classification models to categorise the presence between 0 (none), 1 (red squirrels), and 2 (grey squirrels) according to the land cover and number of grey squirrels within the range of the points in the study area. Randomly generated points across the study area were used as non-presence points, along with the occurrence points of squirrels. Random forest and Support Vector Machine (SVM) models were chosen for classification, commonly used algorithms for classification (IBM, 2023 and Pisner and Schnyer, 2020). Both models can reduce the risk of overfitting; the random forest model uses multiple decision trees (IBM, 2023), and SVM allows misclassifications by margins around the decision boundary (Pisner and Schnyer, 2020). Other classification models, such as KNN and the logistic regression model, do not appear suitable for this study. KNN does not take into account any dynamics between the spatial components because it classifies the points only by the distances between points. And the logistic regression is appropriate for binary classification, but this study requires multi-classification for three types of presences (none, red squirrels, and grey squirrels).

The core spatial components of the study are study area, coordinates, occurrence points of red squirrels and grey squirrels, and land cover types. Using the components, the random forest and SVM models will be evaluated by both conventional cross-validation and spatial cross-validation in terms of classification accuracy. Spatial cross-validation is to consider any potential spatial dependencies within the study area.

2. Methodology

2.1. The data and study area

The LCM2015 from CEH (2017) consists of 21 land cover types in 25m-sized rasters. Figure 1 below shows the LCM2015 within the study area.

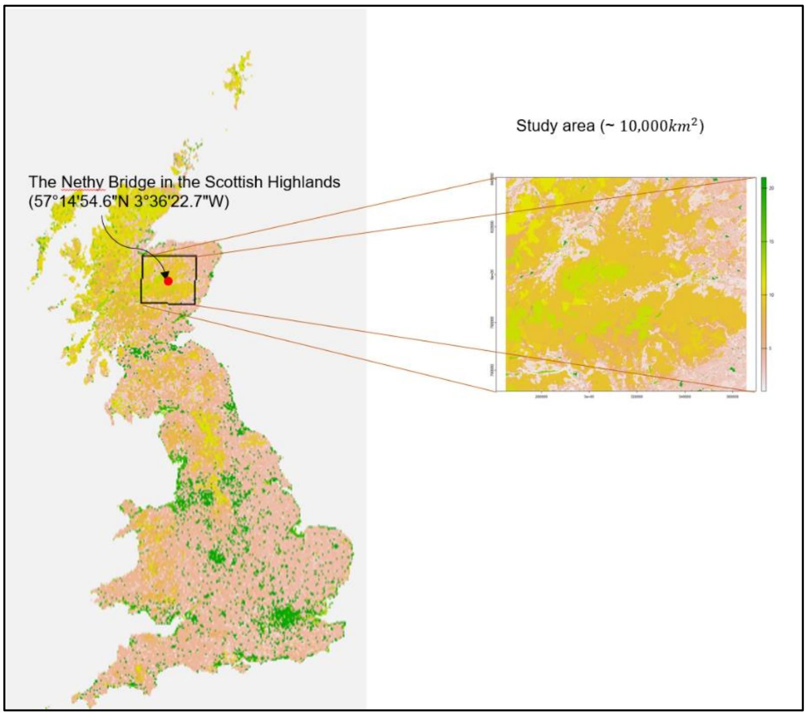

Study area (~ $10{,}000km^2$)
The Nethy Bridge in the Scottish Highlands
(57°14'54.6"N 3°36'22.7"W)

Figure 1. The land cover map and the study area (CEH, 2017)

The land cover types are as below:

1) Broadleaf woodland, 2) Coniferous woodland, 3) Arable, 4) Improved grassland, 5) Neutral grassland, 6) Calcarous gassland, 7) Acid grassland, 8) Marsh, 9) Heather, 10) Heather grassland, 11) Bog, 12) Inland rock, 13) Saltwater, 14) Freshwater, 15) Supra-littoral rock, 16) Supra-littoral sediment, 17) Littoral rock, 18) Littoral sediment, 19) Saltmarsh, 20) Urban, 21) Suburban.

As aforementioned, the occurrence data from the National Biodiversity Network Atlas contains the recording period in date, coordinates (latitude, longitude), uncertainty (distance range), verification status, and the data provider. Among them, coordinates, uncertainty, uncertainty, and verification status are important for this study. Any record with uncertainty≥1000m, verification status=unconfirmed, or missing coordinates will be removed from the dataset to prevent introducing uncertainties to the study. Table 1 summarises the number of records according to these categories.

	Missing coordinates	**uncertainty≥1000m**	**verification status=unconfirmed**
Red squirrel data	0	2,128	2,072
Grey squirrel data	5	38	5

Table 1. The number of squirrel occurrence records with quality issues

2.2. Analysis

The analysis consists of five steps – Data pre-processing, data scaling, model evaluation, model tuning, and model selection (Figure 2).

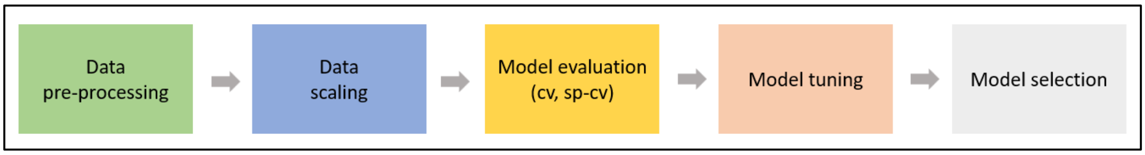


Figure 2. The analysis process of the study

The analysis starts with the data pre-processing, which includes treating missing and unconfirmed values, projection, cropping, and reclassifying. Projecting all the datasets into the same coordinate reference system (CRS) and cropping them according to the study area are also important groundwork.

Then, buffer sizes from 100 to 2000 metres will be tested and compared by the random forest model accuracy. Buffer size relates to the percentage of the broadleaf land cover and the number of grey squirrels, which leads to different model performances (accuracy). To compare accuracy values, a simple cross-validation method of three folds and three repetitions was used with the random forest model.

There are two reasons why we consider the broadleaf woodland land cover important. First, we can find that the histograms in Figure 3 show that the majority of both squirrel types live on the broadleaf land cover. Figure 4 also shows the occurrence points appearing along the broadleaf. In addition, Flaherty et al. (2012) state that both squirrel types prefer broadleaf.

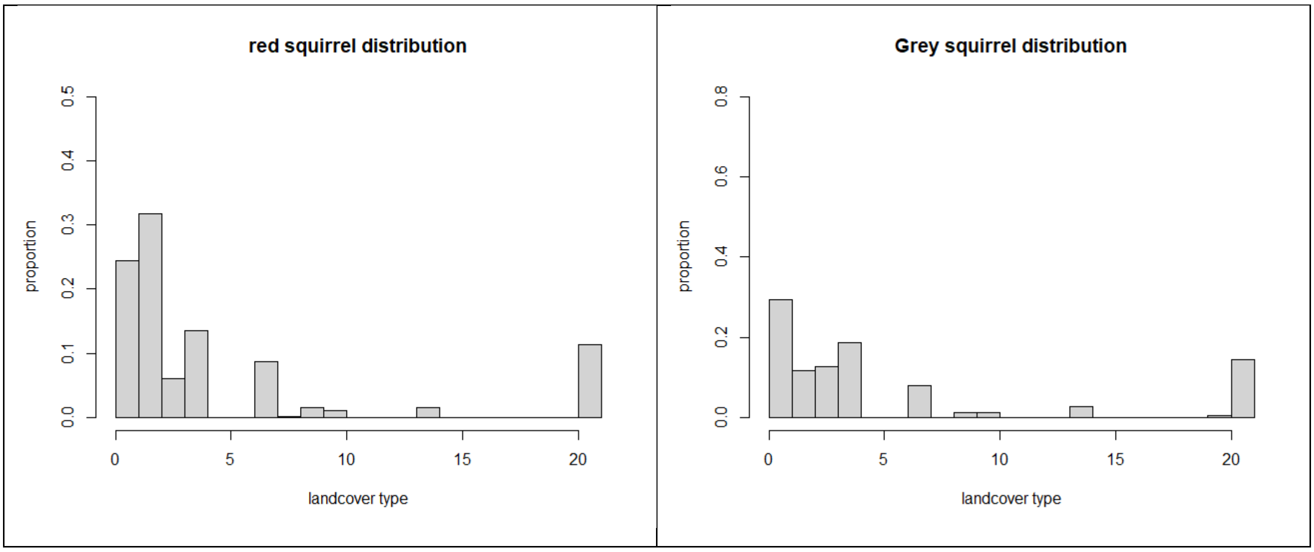


Figure 3. The distribution histograms of squirrels according to the landcover type

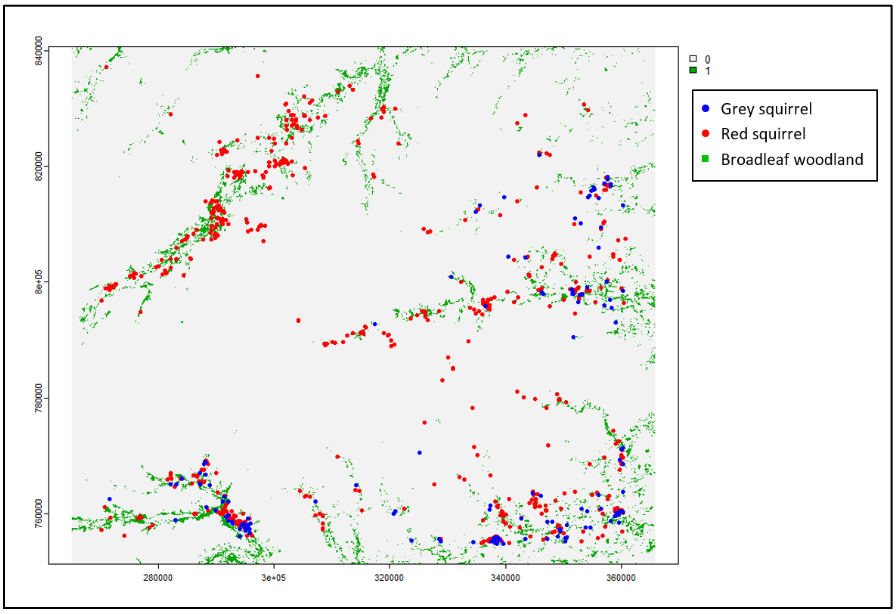


Figure 4. The occurrence points of squirrels on the broadleaf landcover map

After choosing the dataset of the scale with the highest accuracy, the random forest model and SVM model will be evaluated by conventional cross-validation (cv) and spatial cross-validation (sp-cv). Each method tries five-fold and ten repetitions, unlike the method for scaling. As aforementioned, comparing these two validation methods will give us an idea of how strong the spatial dependence is in the study area.

The two classification models have room for parameter tuning. Thus, we will tune each model to improve the model performance (accuracy). After model tuning, we will compare all the models and choose one for the study (i.e., the species distribution model for Sciurus vulgaris).

3. Results

The highest accuracy of the random forest model was obtained with the buffer size of 1200m (accuracy=0.7594, Figure). The accuracy ranges from 0.7566 to 0.7594 by buffer sizes.

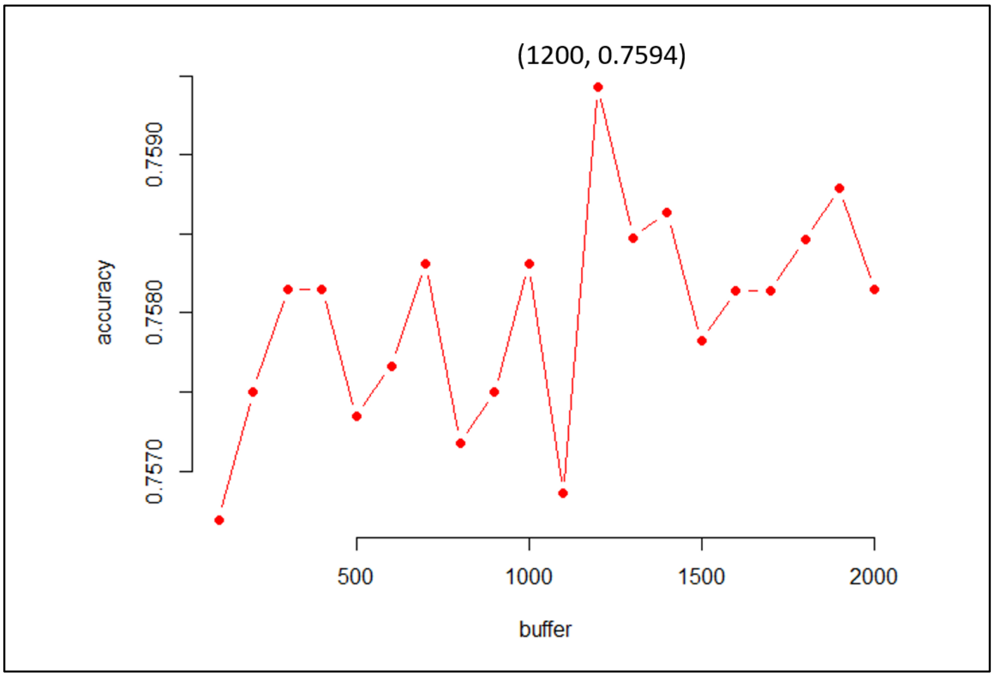


Figure 5. The accuracy values according to buffer sizes

Table 2 below shows the accuracy values of each model with each cross-validation method. Using spatial cross-validation, both models showed lower accuracy values than conventional cross-validation. The decrease in accuracy of the random forest model was bigger than the SVM model (0.2110 > 0.0688).

Model	Conventional cross-validation	Spatial cross-validation
Random forest	0.8086	0.5976
SVM	0.7686	0.6998

Table 2. The accuracy of models with each cross-validation method

The result of hyperparameter tuning of each model are below (Table 3).

Model	Tuning
Random forest	Mtry=1, nodesize=7
SVM	C=1790, sigma=0.00508

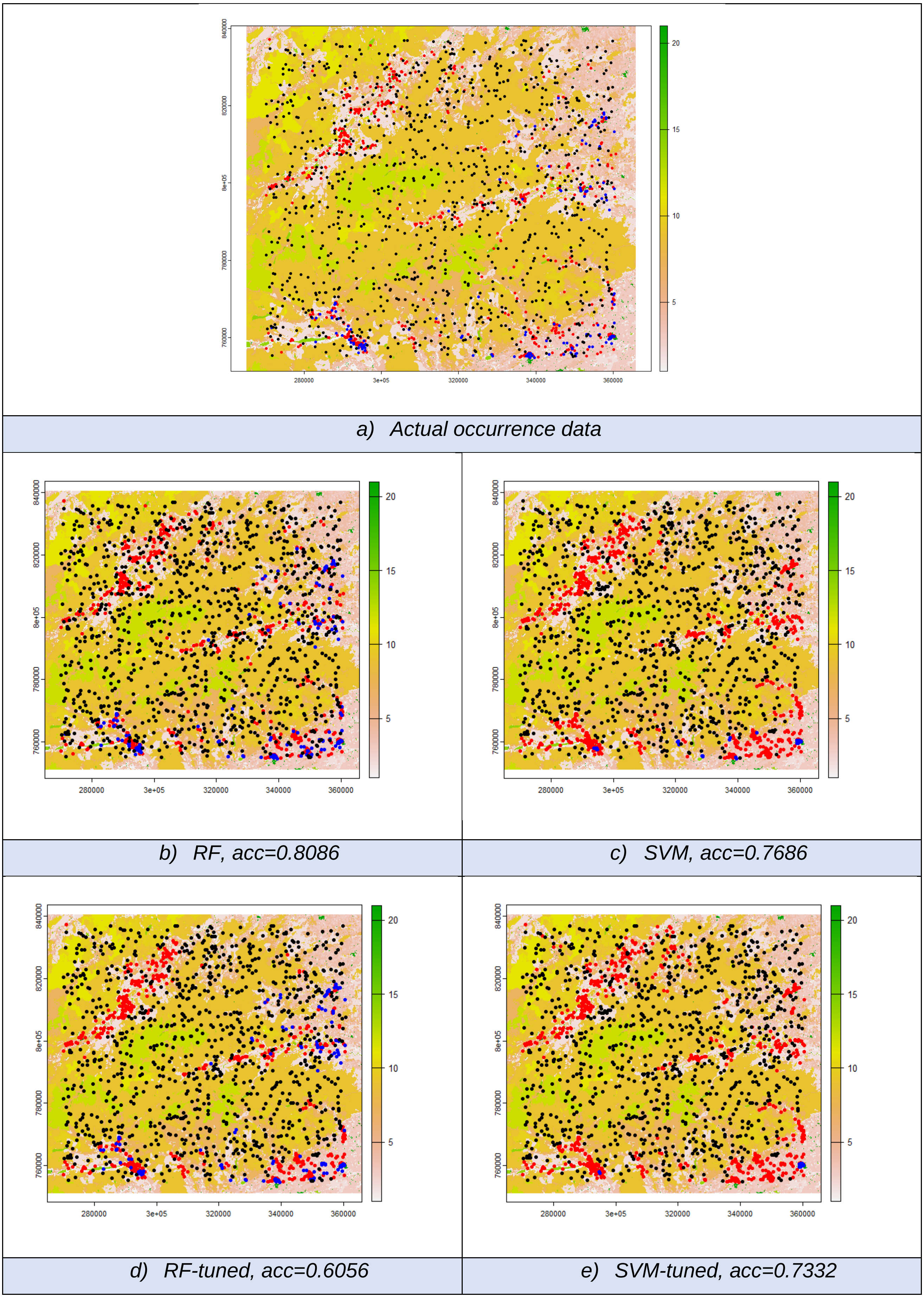
Table 3. The parameter tuning for each model

The improvement of model accuracy was bigger for the SVM model than the random forest model (0.0334 > 0.008, Table 4).

Model	Before tuning	After tuning
Random forest	0.5976	0.6056
SVM	0.6998	0.7332

Table 4. The accuracy improvement by tuning of each model

Figure 6 shows the prediction results of each model. In general, the random forest model predicted more points as grey squirrels than the SVM model. Even the RF-tuned model with the lowest accuracy (d, 0.6056) predicted more grey squirrel points than the SVM models. Compared to the actual data (a), the number of grey squirrel points predicted by the SVM model is also smaller.



a) *Actual occurrence data*

b) *RF, acc=0.8086*

c) *SVM, acc=0.7686*

d) *RF-tuned, acc=0.6056*

e) *SVM-tuned, acc=0.7332*

Figure 6. The prediction results of each model

4. Discussion

The point of this study is to see how accurately the model predicted the occurrence (or distribution) of the target species (i.e., red squirrels) using the core spatial components - study area, coordinates, occurrence points of red squirrels and grey squirrels, and land cover types. In addition to prediction accuracy, another important issue is how well this model can reflect the inter-specific dynamics and spatial dependencies.

First, the accuracy values of each scale (Figure 5) showed that the buffer size of 1200m can lead to the highest prediction accuracy than the other sizes. Thus, it can be said that this study proved the importance of scaling to spatial analysis. However, further studies need to discuss whether the accuracy value differences are significant enough to be meaningful. Moreover, different results from this study are possible considering the accuracy differences and the randomness in the evaluation method (i.e., model accuracy by cross-validation).

The accuracy difference between the conventional cross-validation method and the spatial method implies a considerable spatial dependency across the study area. The distribution of land cover type can be considered as the reason; landcover types are not randomly and evenly distributed but appear with a certain tendency. Species distribution can also be spatial dependent because it changes over time by reproduction and movement in space.

Meanwhile, as shown in Section 3, the random forest model got penalised more severely than the SVM model by the spatial dependency but got improved less than the SVM by parameter tuning. Further studies are required to figure out what differences between these two models led to this result. Or it may be due to the tuning method; this study tuned only two parameters (mtry, nodesize) for the random forest model so tuning more parameters may lead to better improvement for the model. One thing certain is that model tuning is essential to improve model performance.

Regarding the inter-specific interactions between red squirrels and grey squirrels, this study tried to use the number of grey squirrels nearby for classification. It was possible to obtain an accuracy above 0.7 for the SVM model even when considering the spatial dependency, but it is notable that the model always predicted fewer grey squirrels than the actual data. This is the opposite of the tendency of grey squirrels to threaten the survival of red squirrels mentioned in Flaherty et al. (2012). Thus, the SVM model may not be appropriate for predicting the distribution of red squirrels even though its accuracy values.

Based on the points above, we may consider choosing the tuned random forest model with spatial dependency. The random forest model without spatial dependency shows higher accuracy than the tuned model (0.8086 > 0.5976), but this difference can be considered the result of over-optimistic validation. Therefore, the tuned random forest model can be chosen as the final model of this study. As mentioned above, the tuned model can be further improved by tuning parameters other than mtry or nodesize.

The overall approach of this study also has room for improvement. For example, Figure 3 shows that landcover type-2 (Coniferous woodland) and type-4 (Improved grassland) also take up the majority of the squirrels' distribution. Thus, another model that includes these landcover types as inputs can be trained for prediction. The cross-validation methods may also change regarding the number of folds or repetitions. Furthermore, the study area is another important scaling factor in spatial analysis (Sheppard and Mcmaster, 2008); we may change the extent or the coordinates of the area for an improved model.

It is also possible to make other fundamental changes to the model. The model can be more sophisticated using the Artificial Neural Network (ANN) algorithm with more independent variables. This study used classification algorithms rather than regression models and did not take into any changes over time. In particular, a time series model may effectively represent the dynamics between the two squirrel types mentioned in Flaherty et al. (2012). Point pattern analysis approaches can also be used to compare. Like the spatial cross-validation method, point pattern analysis also can be a way to deal with spatial dependencies. The essence is getting to the heart of the complex dynamics affecting species distributions to create more accurate but uncomplicated models.

REFERENCES

- CEH (2017). Land Cover Map 2015. [Online]. Available at: https://doi.org/10.5285/bb15e200-9349-403c-bda9-b430093807c7 (Accessed: 13 May 2023).
- Elith, J. and Leathwick, J. (2009). ‘Species Distribution Models: Ecological Explanation and Prediction Across Space and Time’, *Annual Review of Ecology, Evolution, and Systematics* 40, pp. 677-697. [Online]. Available at: https://doi.org/10.1146/annurev.ecolsys.110308.120159 (Accessed: 14 May 2023).
- Flaherty, S., et al. (2012). ‘Impact of Forest Stand Structure on Red Squirrel Habitat Use’, *Forestry: An International Journal of Forest Research* 85(3), pp. 437-444. [Online]. Available at: https://doi-org.manchester.idm.oclc.org/10.1093/forestry/cps042 (Accessed: 14 May 2023).
- IBM (2023). *What is random forest?.* Available at: https://www.ibm.com/topics/random-forest (Accessed: 16 May 2023).
- Jelinski, D. E. and WU, J. (1996). ‘The Modifiable Areal Unit Problem and Implications for Landscape Ecology’, *Landscape ecology* 11(3), pp. 129–140. [Online]. Available at: https://doi.org/10.1007/BF02447512 (Accessed: 15 May 2023).
- Keil, P., et al. (2021). ‘Measurement and Analysis of Interspecific Spatial Associations as a Facet of Biodiversity’, *Ecological monographs* 91(3), pp. 1-22. [Online]. Available at: https://doi.org/10.1002/ecm.1452 (Accessed: 14 May 2023).
- National Biodiversity Network Atlas (2019). Sciurus carolinensis Gmelin, 1788. [Online]. Available at: https://species.nbnatlas.org/species/NHMSYS0000332764 (Accessed: 15 May 2023).
- National Biodiversity Network Atlas (2015). Sciurus vulgaris Linnaeus, 1758. [Online]. Available at: https://species.nbnatlas.org/species/NBNSYS0000005108 (Accessed: 15 May 2023).
- Pisner, D. and Schnyer, D. (2020). ‘Chapter 6 – Support vector machine’, in Mechelli, A. and Vieira, S. (eds.) *Machine Learning: Methods and Applications to Brain Disorders*, London: Academic Press. pp. 101-121.
- Sheppard, E. and McMaster, R. (2004). ‘Scale and Geographic Inquiry: Contrasts, Intersections, and Boundaries’, in Sheppard, E. and McMaster, R. (eds.) *Scale and Geographic Inquiry: Nature, Society, and Method*. Malden, USA: Blackwell Publishing Ltd. pp. 256–267.