UDS International Journal of Development [UDSIJD]
Volume 5 No. 2, 2018
http://www.udsijd.org
ISSN: 2026-5336

THE EFFECT OF SPATIAL THINNING ON THE POTENTIAL DISTRIBUTION OF 10 AFRICAN INDIGENOUS VEGETABLES

***Nyarko, G. and Bayor. H.**

Department of Horticulture, University for Development Studies, Tamale, Ghana.

**Corresponding author's email: gnyarko@uds.edu.gh*

Abstract

Species distribution modelling is important in conservation planning and many other fields of study. It is however often fraught with bias in the location data used to develop the models. Spatial thinning is one of the bias correction methods. It has been reported to be superior to the background correction method in modelling experiments. However, the effect of spatial thinning on predicted areas and model assessment characteristics are unreported. We examined the effects of spatial thinning on the potential distribution of 10 African indigenous vegetables (AIV). The aims of our study were to investigate the effect of different spatial thinning distances on (1) the potential predicted areas (present and future 2070) of 10 species of AIV and (2) model evaluation statistics. We applied spatial thinning to the location data using the R package 'spThin' at distances of 0, 10, 20, 40, 60, 80 and 100km. For each species MaxEnt was used to run 10 replicate models with cross-validation and a threshold of 10% training presence. There were between 54 and 564 location data points a species after cleaning of GBIF data and 153-25 after thinning at a spatial resolution of 100km. The area under the curve (AUC) of the receiver operating characteristic, Boyce Index and the true skill statistic (TSS) decreased with increasing spatial thinning distance but sensitivity remained relatively constant. There was consistency in the direction of prediction for eight of the 10 species while spatial thinning influenced the direction of prediction for two species. Future 2070 suitable climatic envelope may be larger than the present for six species, remain the same as present for three species and become smaller for one species. We concluded that while spatial thinning may be useful in correcting for under-estimation caused by clustered data, it might also lead to incompleteness in environmental space leading to unexpected results if not done with caution. Although the differences in the extent of suitable climatic envelope may imply reduction of overall biodiversity, no species was under serious threat of complete loss of suitable environment in the future.

Keywords: Area under the curve (AUC), Boyce Index, Sensitivity, Spatial Thinning, Specificity, True Skill Statistic, Vegetables

Introduction

Species distribution modelling is important in the study of many areas including ecology, conservation planning, species invasion, and evolution (Guisan, Thuiller and Wilfried, 2005; Kramer-Schadt *et al.*, 2013). The accuracy of models is important for decision making. Model inputs quality (location data and environmental variables) and processing (software algorithms, parameter setting and whether sufficient precautions are taken) affect the accuracy of the model (Beale & Lennon, 2012; Soultan & Safi, 2017).

Large amounts of geographic data are available in the form of electronic databases that are easily accessible, for example Global Biodiversity Information Facility (GBIF, 2018). Other records exist in natural history museums, herbaria, and published literature on collections. These records are often presented as presences without indication of

where the species was searched but not found. Errors are often associated with such data and include location errors, identification errors and un-even sampling effort over the species ranges (Bloom, Flower, & DeChaine, 2017; Newbold, 2010).

To build reliable models, the number of location records should form a representative sample of species environment (Hernandez et al., 2006; Wisz et al., 2008). Modelling experiments have shown that even small samples as low as 10 can produce reliable models when MaxEnt is used (Hernandez et al., 2006; van Proosdij et al., 2016). However, additional records from un-sampled areas of the species range always improve model performance irrespective of records already available (Feeley & Silman, 2011). Therefore, the maximum number of records should be used but clustered records should be avoided since clustering results in poor models (Kramer-Schadt et al., 2013; Stolar & Nielsen, 2015; Syfert, Smith, & Coomes, 2013).

Minor location errors may have little effect on model accuracy (Soultan & Safi, 2017). Spatial bias on the other hand, appears to lead to both local over prediction and local under prediction of prevalence in different grid cells (Syfert et al., 2013). Location data bias also increases both false positive and false negative (Kramer-Schadt et al., 2013) and may result in under prediction of the species range (Bayor, 2012; Beale & Lennon, 2012). However, most of the readily available data are from records collected and cumulated over a period of time (Gomes et al., 2018). These types of data are often spatially biased (Anderson, 2012; Beck, Böller, Erhardt, & Schwanghart, 2014; Gomes et al., 2018; Hortal, Jiménez-Valverde, Gómez, Lobo, & Baselga, 2008). The nature of the bias in such data is often unknown and therefore cannot be readily corrected (Bayor, 2012; Yesson et al., 2007). However, such data form important sources for research to the extent that they cannot be ignored (Newbold, 2010) and have been widely used because it is expensive and sometimes impractical to gather complete location data for any single species (Benito et al., 2013; Gomes et al., 2018). When un-accounted for in modelling biased data, results are often poor models (Fourcade, Engler, Rödder, & Secondi, 2014; Gomes et al., 2018; Hortal et al., 2008). Therefore, it is important to deal with bias in species distribution modelling (Aiello-Lammens, Boria, Radosavljevic, Vilela, & Anderson, 2015; Fithian, Elith, Hastie, & Keith, 2015; Phillips et al., 2009). How effective different methods are, have been under study (Fourcade et al., 2014; Kramer-Schadt et al., 2013). The background species method (Phillips et al., 2009) and the spatial thinning method (Aiello-Lammens et al., 2015) are widely regarded as effective (Fourcade et al., 2014; Kramer-Schadt et al., 2013). However, the extent of thinning required to make bias correction effective is unknown.

Spatial thinning has been reported as one of the best methods to deal with spatial bias of geographic data, a common feature of databases (Boria, Olson, Goodman, & Anderson, 2014; Fourcade et al., 2014). Representative information of the geographic distribution of the species and also of environmental data are necessary for building an accurate model (Tsoar et al. 2007; Kadmon, Farber, & Danin, 2003). Spatial thinning removes some of the location data leading to information loss. This loss of information however, may be compensated for by removal of the overweighting of some environmental variables (Boria et al., 2014).

There are many software programmes and algorithms used for modelling species distribution (Ahmed et al., 2015; Elith et al., 2006; Wisz et al., 2008). Of these MaxEnt is probably the most commonly used because it is simple, has less stringent data requirements (presence only data is used), able to accommodate minor spatial errors of location data and performs well when the sample size is small (Ahmed et al., 2015; Fourcade et al., 2014; Kramer-Schadt et al., 2013; Wisz et al., 2008).

African indigenous vegetables (AIV) form an important source of food and nutrition for a large number of people on the continent (Achigan-Dako et al., 2011; Maundu, Achigan-Dako, & Morimoto, 2009). Unfortunately, Africa still remains an under nourished continent and many households depend on wild and semi-wild vegetables for both food and nutrition (Achigan-Dako, Sogbohossou & Maundu, 2014; Achigan-Dako et al., 2011). Climate change presents a challenge particularly to wild plants.

Africa is regarded as a food insufficient continent (Luan, Cui, & Ferrat, 2013). Families and households have supplemented this with gathering vegetables and other food items from the wild (Achigan-Dako et al., 2011; National Research Council, 1996, 2008).

Vegetables from the wild also form an important income generating activity for many rural households (Weinberger & Pichop, 2009). Climate change may negatively impact on food production in agricultural systems (Ramirez-Cabral, Kumar, & also give them a high potential for future cultivation in the face of climate change (Ebert, 2014). It is therefore important to develop conservation strategies for edible wild vegetables for which climate change may affect adversely. However, the impacts of climate change on these vegetables are an important prelude to conservation planning.

Methods to evaluate species distribution models are many (Liu, White, & Newell, 2009; Liu, White, & Newell, 2011). Some of them include the area under the curve (AUC) of the receiver operating characteristic curve and the true skill statistic but these methods were developed for presence-absence data (Liu, White, & Newell, 2009). Their application to presence only data has been questioned (Jiménez-Valverde, 2012; Leroy et al., 2017; Lobo, Jiménez-Valverde, & Real, 2008; Raes & Ter Steege, 2007). It has been shown that biased location data often gives inflated values for AUC making models appear better than they actually are (Hijmans, 2012; Radosavljevic & Anderson, 2014). Also, species with narrow distributions generally give higher AUC values than widely distributed species (Raes & Ter Steege, 2007; van Proosdij et al., 2016). Hirzel et al., (2006) proposed the Continuous Boyce Index (Boyce Index) as a presence-only model evaluation method, a modified form of Boyce et al., (2002) index.

AUC values range from 0-1 with 0.5 representing models with no better than random predictions and 1 representing perfect model prediction success (Fielding & Bell, 1997). AUC values >0.7 are often regarded to have good discriminatory power (Raes & Shabani, 2017; Wheeler et al. 2000). Food insufficiency may probably get worse and contribution from non-agricultural food sources including gathering from the wild may increase in importance. The resilience of wild vegetables may Ter Steege, 2007). The true skill statistic ranges from -1 to +1 with values around zero and below indicating models with no better than random prediction success (Allouche, Tsoar, & Kadmon, 2006). Positive values indicate better than random predictions with +1 being a perfect model fit. The Boyce index ranges from -1 to +1. Negative values indicate wrong model, zero values indicate models with discriminatory power no better than random and +1 indicates a perfect prediction ability (Hirzel et al., 2006).

In this paper we explore the effects of spatial thinning on the present and future predicted areas of 10 species of wild and semi-wild AIV. Our aims were to investigate (1) the effects of the distance of spatial thinning on the potential predicted areas of 10 AIV (2) the effects of spatial thinning on AUC, true skill statistic (TSS), sensitivity, specificity and Boyce Index of distribution models.

Methods

Species Location Data

Ten AIV chosen on the basis of their usefulness scoring of at least a three star in their uses at the Prota4U database (PROTA4U, 2018) and they are wild or semi-wild. The species are shown in Table 1. Location data were downloaded from GBIF (GBIF, 2018) for all 10 species. These were supplemented with location descriptions found in the literature. These records were verified by plotting on the map of Africa. Ranges were checked with information on the literature for each species, especially using information from the Prota4U database.

Table 1: Species and number of location records used for modelling and the spatial thinning distances (km) applied. Thinning at 0km implies thinning was not done.

	Thinning Distance (km)						
Species	0	10	20	40	60	80	100
Commelina Africana L.	564	543	520	297	234	184	153
Cleome gynandra L.	324	297	278	220	186	165	148
Corchorus olitorius L.	287	243	209	152	130	119	105

Ceratotheca sesamoides Endl.	229	184	149	118	100	86	80
Gymnanthemum amygdalinum (Delile) Sch.Bip. ex Walp.	217	195	171	138	138	104	97
Hibiscus sabdariffa L.	131	114	101	84	76	66	60
Solanum aethiopicum L.	313	240	198	143	110	83	72
Solanum macrocarpon L.	139	120	106	93	76	68	60
Solanum scabrum Mill.	54	45	39	33	30	27	27
Talinum fruticosum (L.) Juss	100	67	52	38	30	27	25

Environmental Variables

Climate data were obtained from the WorldClim website (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) with additional soil data from the Harmonized World Soil Database (FOA/IIASA/ISRIC/ISS-CAS/JRC, 2012). Climate data were trimmed to match the map of Africa. All the 19 bioclimatic variables were used although issues with autocorrelation may arise (Dormann et al., 2013; Hijmans, 2012). We were of the opinion that the benefits derived from dropping some variables is minimal (Kramer-Schadt et al., 2013). Four soil variables taken from the top layer (namely, soil pH, soil texture, soil organic matter content and soil bulk density) were added to the bioclimatic variables giving the total number of variables used as 23. All climate and soil variables were prepared at 2.5 arc minutes equivalent to 4.5km at the equator.

Spatial Filtering

Location data from each of the species were filtered at six distances (10, 20, 40, 60, 80, and 100km) using the package 'spThin' (Aiello-Lammens et al., 2015). In addition, the un-thinned data sets were added (designated 0km) for the modelling. Therefore, for the ten species we had 70 data sets (including non-thinned data from GBIF).

Software

Each of these data sets was run on MaxEnt (ver. 3.4.0) using the default setting. These included a background pseudo-absence records of 10,000, and output format of Cloglog with 10 replications of cross-validation. A threshold of 10% training presence was used. Seven hundred models (10 species x 7 thinning distances x 10 replicates) were built and for each a prediction was made for the present and the future (2070) using the Had85 representative concentration pathway. This path-way represents one of the severe situations that might arise (van Vuuren et al., 2011). Present and future predicted areas were calculated using ArcMap 10.4.1.

Model Assessment

Models were evaluated with the Area Under the Curve (AUC) of the receiver operating characteristic (ROC) curve (Fielding & Bell, 1997), and the True Skill Statistic (TSS) (Allouche et al., 2006). Sensitivity and Specificity were also calculated and examined (Fielding & Bell, 1997). Continuous Boyce index (Boyce index; Hirzel et al., 2006) was calculated with the R package 'ecospat' (Di Cola et al., 2017).

Results

Number of Location Records

Number of location records for the species ranged from a high of 564 to a low of 25 after thinning (Table 1). These were the records that MaxEnt actually used in building the models. Number of records within this range has often been considered to be adequate for building reliable models (Soultan & Safi, 2017; van Proosdij et al., 2016).

Model Assessment

The AUC, Boyce index, TSS, and Specificity all showed significant (Kruskal Wallis test $p<0.05$; mean ranks separated by Conover-Iman procedure) reduction with increase in spatial thinning distance of location data points (**Error! Reference source not found.**; Figure 1a-j; Figure 2a-j; Figure 3a-j; and Figure4a-j). The models built with GBIF data always give the highest values in all these model assessment parameters (Appendix 1) but in many instances, these were not significantly different from the next two or three thinning distances (0, 20 and 40km). The

widest spacing (100km) mostly gave the least values of these four parameters. Sensitivity, however, showed only a mild declining trend and did not differ significantly (Kruskal Wallis test p =0.05) with increasing thinning distance. The median values were generally high for all models (Appendix 1). The lowest medians of the model assessment statistics at the widest spacing (100km) were AUC (0.77 - 0.91), TSS (0.37 - 0.62), sensitivity (0.65 - 0.82), specificity (0.54 - 0.90) and Boyce index (0.56 - 0.86) (Appendix 1). This probably indicates a good enough discriminatory power even at the widest spacing for models.

The spread in AUC, Boyce Index, TSS and sensitivity increased with increasing thinning distance (Fig 1a-j; Figure 2a-j; Figure 3a-j; and Figure5a-j). This was particularly pronounced in the sensitivity measures.

Climate and Predicted Area

Future climate had different effects on potential predicted areas of the species (Figure 6). Based on paired t-test (at 5% probability) of present and future predicted areas of each species, one species (*Commelina Africana*) is likely to experience a significant reduction in suitable climatic envelope in the future 2070. Three other species (viz. *C. gynandra*, *C. oliturios* and *G. amygdalinum*) would experience no change while the rest are likely to experience various amounts of increases in climatic envelope. The magnitudes of changes vary widely among species ranging from -62% for *Commelina Africana* to 138% for *Talinum fruticom* (Figure7a-j).

Thinning and Predicted Area

Over all, thinning increased the potential predicted area. In all instances the un-thinned data had the least predicted area. As indicated in **Error! Reference source not found.**, the widest thinning also corresponds with fewest number of location records. The present predicted areas showed increasing trend with increasing distance of spatial thinning (Figure 7a-j). However, the widest spatial thinning (100km) does not always produce the largest predicted area except for three species (*C. Africana*, *C. gynandra* and *S. aethiopicum*). In the other instances, the 100km thinning distance produced predicted areas close to the highest and mostly not significantly different from the highest (one-way ANOVA; means separated by Tukey HSD).

Table 2: *Kruskal Wallis test values and levels of significance (p-values) to compare model assessment statistics (Test AUC, sensitivity, specificity, TSS and Boyce Index) among different thinning distances for each species. Thinning distances did not significantly affect sensitivity values but AUC, specificity, TSS and Boyce Index values decreased with increasing thinning distances. (AUC = area under the curve of the receiver operating characteristic curve; TSS = true skill statistic)*

Species	Test-statistic	Test AUC	sensitivity	specificity	TSS	Boyce Index
	Chi-square	26.8	1.8	61.6	27.6	40.1
Commelina africaa	Significance	<0.01	0.94	< 0.01	<0.01	<0.01
	Chi-square	30.2	1.3	57.6	22.5	13.9
Cleomen gynandra	Significance	<0.01	0.97	<0.01	<0.01	0.03
	Chi-square	31.1	7.0	55.4	16.9	28.6
Corchorus olitorius	Significance	<0.01	0.32	<0.01	0.01	<0.01
	Chi-square	29.1	12.5	52.9	20.6	28.1
Ceratotheca sesamoides	Significance	<0.01	0.05	<0.01	<0.01	<0.01
	Chi-square	22.8	10.5	29.4	15.1	31.7
Gymnanthemum amygdalinum	Significance	<0.01	0.10	<0.01	0.02	<0.01
	Chi-square	15.6	8.6	50.7	16.4	9.9
Hibiscus sabdariffa	Significance	0.02	0.20	<0.01	0.01	0.13

	Chi-square	14.2	5.6	62.1	12.9	30.2
Solanum aethiopicum	Significance	0.03	0.47	<0.01	0.04	<0.01
	Chi-square	15.4	4.1	49.0	7.9	7.3
Solanum macrocarpon	Significance	0.02	0.66	<0.01	0.24	0.29
	Chi-square	12.3	3.8	39.5	7.4	7.6
Solanum scabrum	Significance	0.06	0.71	<0.01	0.28	0.27
	Chi-square	14.4	3.8	47.2	6.5	1.1
Talinum fruticosum	Significance	0.03	0.70	<0.01	0.37	0.98

Most of the predicted future potential areas also followed this trend (Figure 7a-j). However, there were instances where the magnitude of the predicted areas of the present and future change sizes as a result of thinning. For eight out of the 10 species, spatial thinning increased the potential predicted areas of both the present and the future 2070. However, for *Corchorus olitorius* (Figure 7c) closer spacing of data points (0, 10, and 20km thinning) predicts a larger present climatic enveloped for the species than for the future 2070, while wider spacing (60, 80, and 100km) predicts a smaller present suitable climatic envelope than for the future. There can be only one 'true' climatic envelope for the species and it can either be larger or smaller than the present but not both. The reverse of this situation occurs in *G. amygdalinum* (Figure 7e). For these species either one or both observed situations may be wrong.

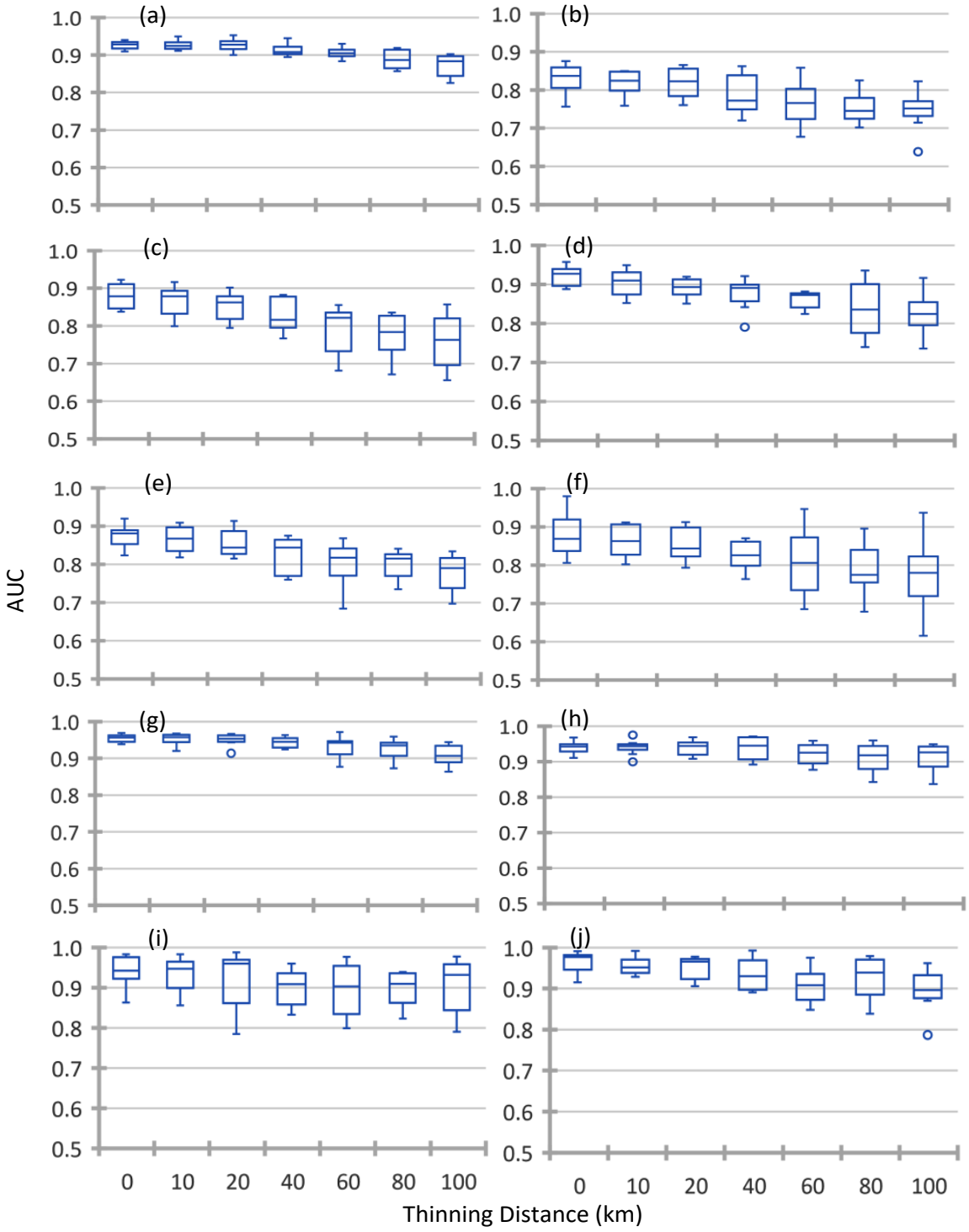

Figure 1: Box and whiskers plot of the area under the curve (AUC) and thinning distance (km). The box represents the inter-quartile range. The line in the box represents the median. (a) C. africana, (b) C. gynandra, (c) C. olitorius (d) C. sesamoides, (e) G. amygdalinum, (f) H. sabdariffa, (g) S. aethiopicum (h) S. macrocarpon (i) S. scabra (j) T. fruticosum.

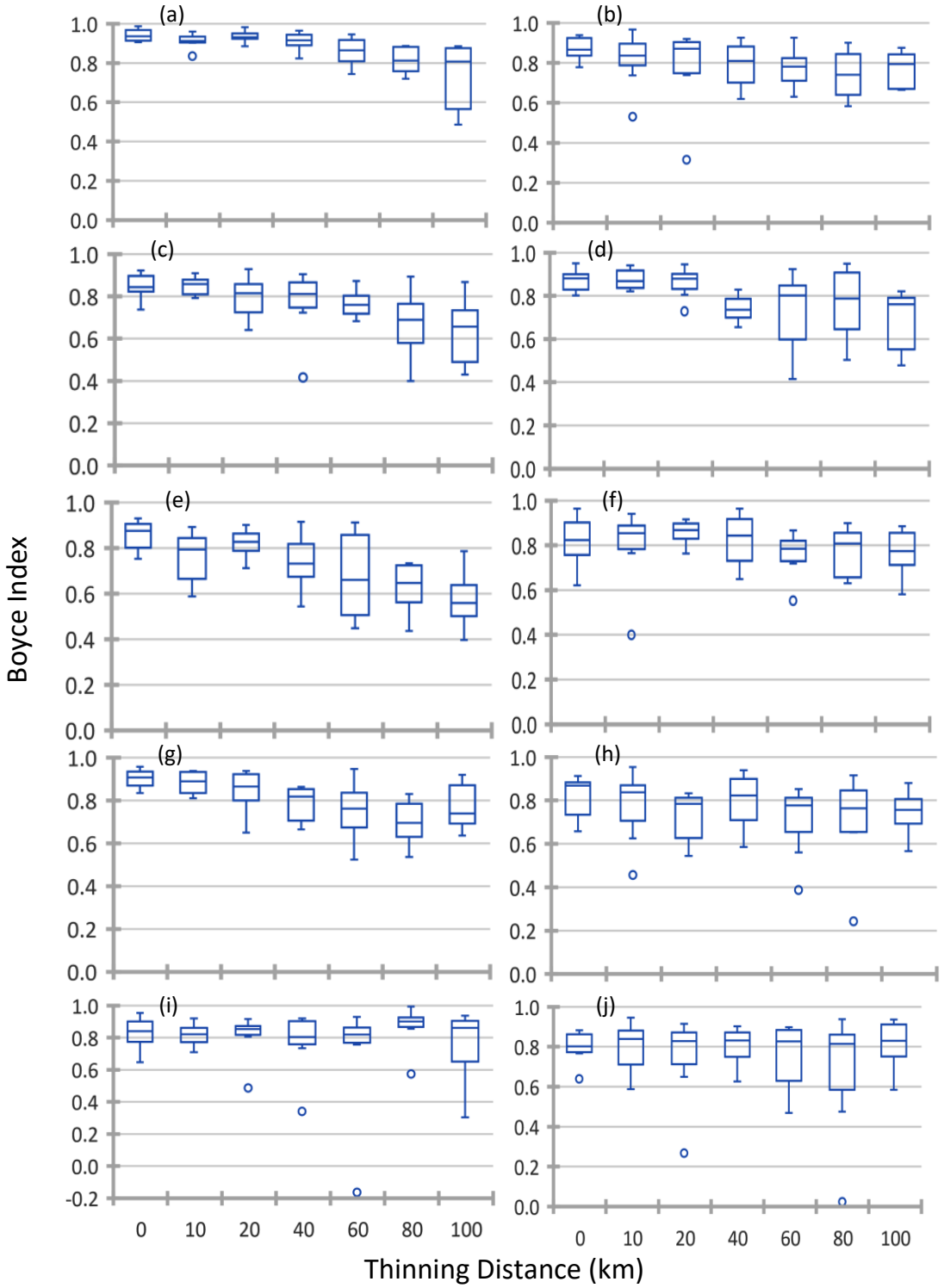

Figure 2: Box and whiskers plot of Boyce Index for each of the species shows a decrease with increasing thinning distance. (a) C. africana, (b) C. gynandra, (c) C. olitorius (d) C. sesamoides, (e) G. amygdalinum, (f) H. sabdariffa, (g) S. aethiopicum (h) S. macrocarpon (i) S. scabra (j) T. fruticosum

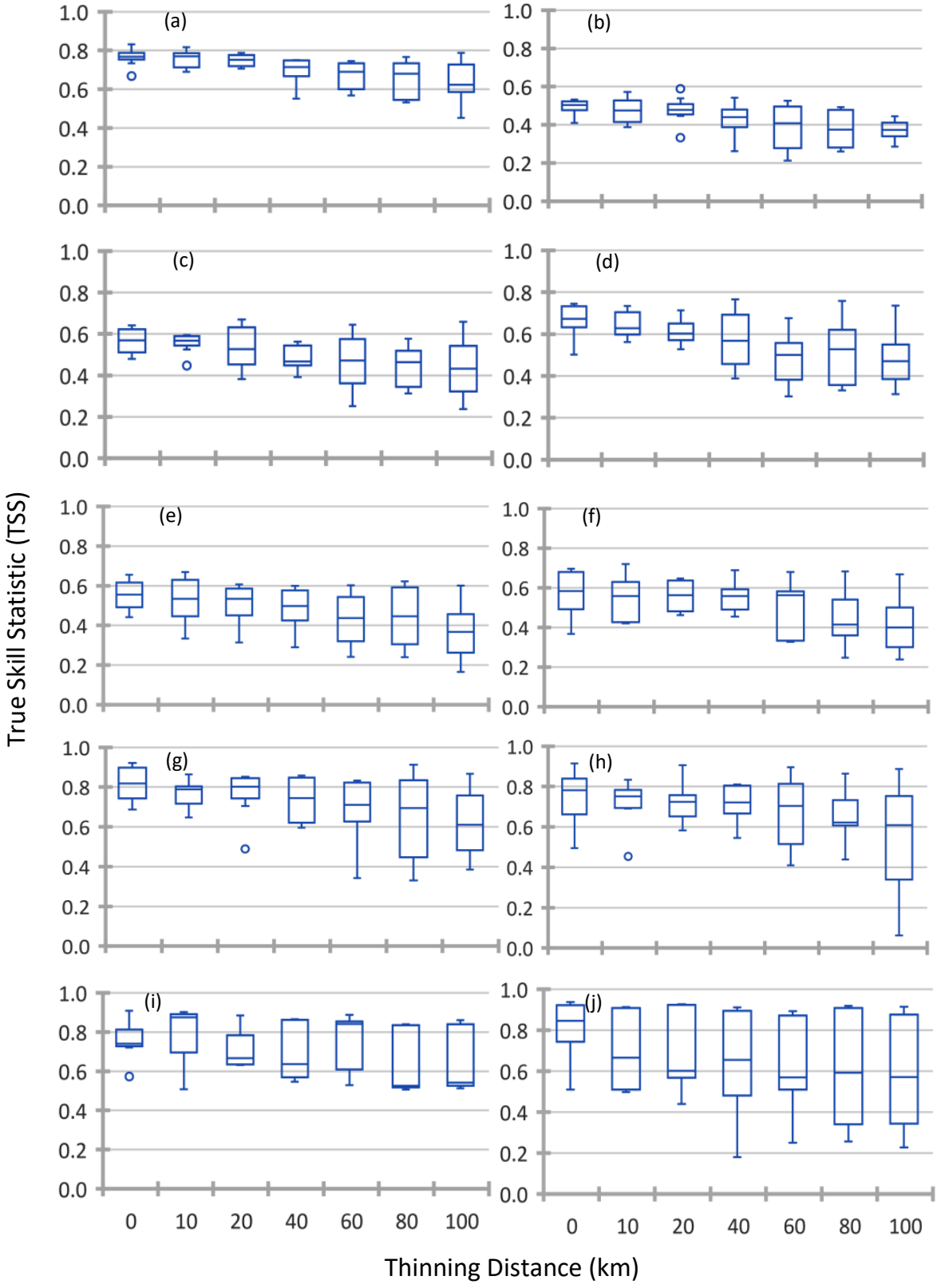

Figure 3: Box and whiskers plot of the true skill statistic (TSS) and thinning distance (km). The box represents the inter-quartile range. The line in the box represents the median. (a) C. africana, (b) C. gynandra, (c) C. olitorius (d) C. sesamoides, (e) G. amygdalinum, (f) H. sabdariffa, (g) S. aethiopicum (h) S. macrocarpon (i) S. scabra (j) T. fruticosum

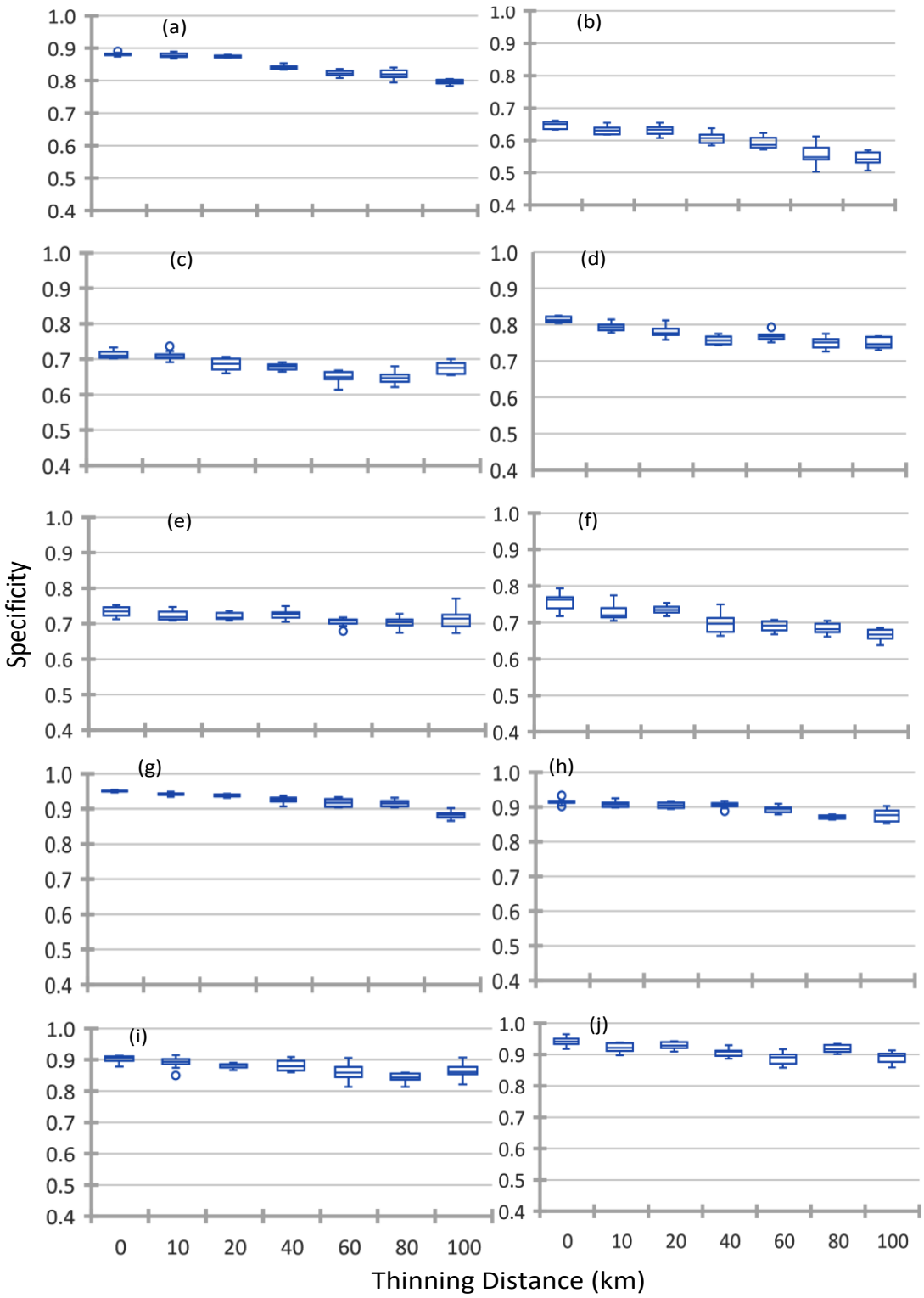

Figure 4: Box and whiskers plot of specificity and thinning distance (km) showing a decreasing trend with thinning distance. The box represents the inter-quartile range. The line in the box represents the median (a) C. africana, (b) C. gynandra, (c) C. olitorius (d) C. sesamoides, (e) G. amygdalinum, (f) H. sabdariffa, (g) S. aethiopicum (h) S. macrocarpon (i) S. scabra (j) T. fruticosum

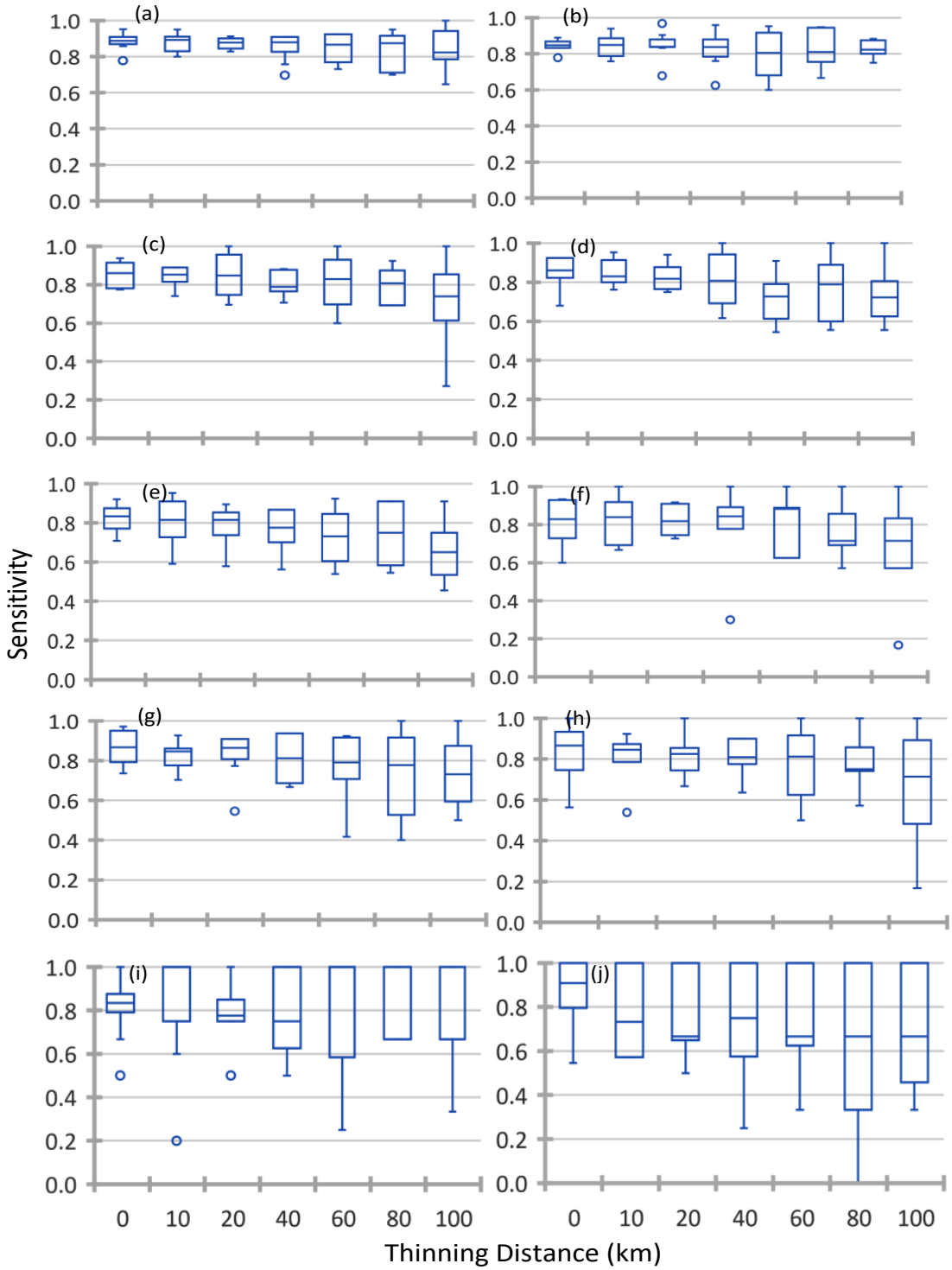

Figure 5: Box and whiskers plot of sensitivity and thinning distance (km). The box represents the inter-quartile range. The line in the box represents the median. (a) C. africana, (b) C. gynandra, (c) C. olitorius (d) C. sesamoides, (e) G. amygdalinum, (f) H. sabdariffa, (g) S. aethiopicum (h) S. macrocarpon (i) S. scabra (j) T. fruticosum

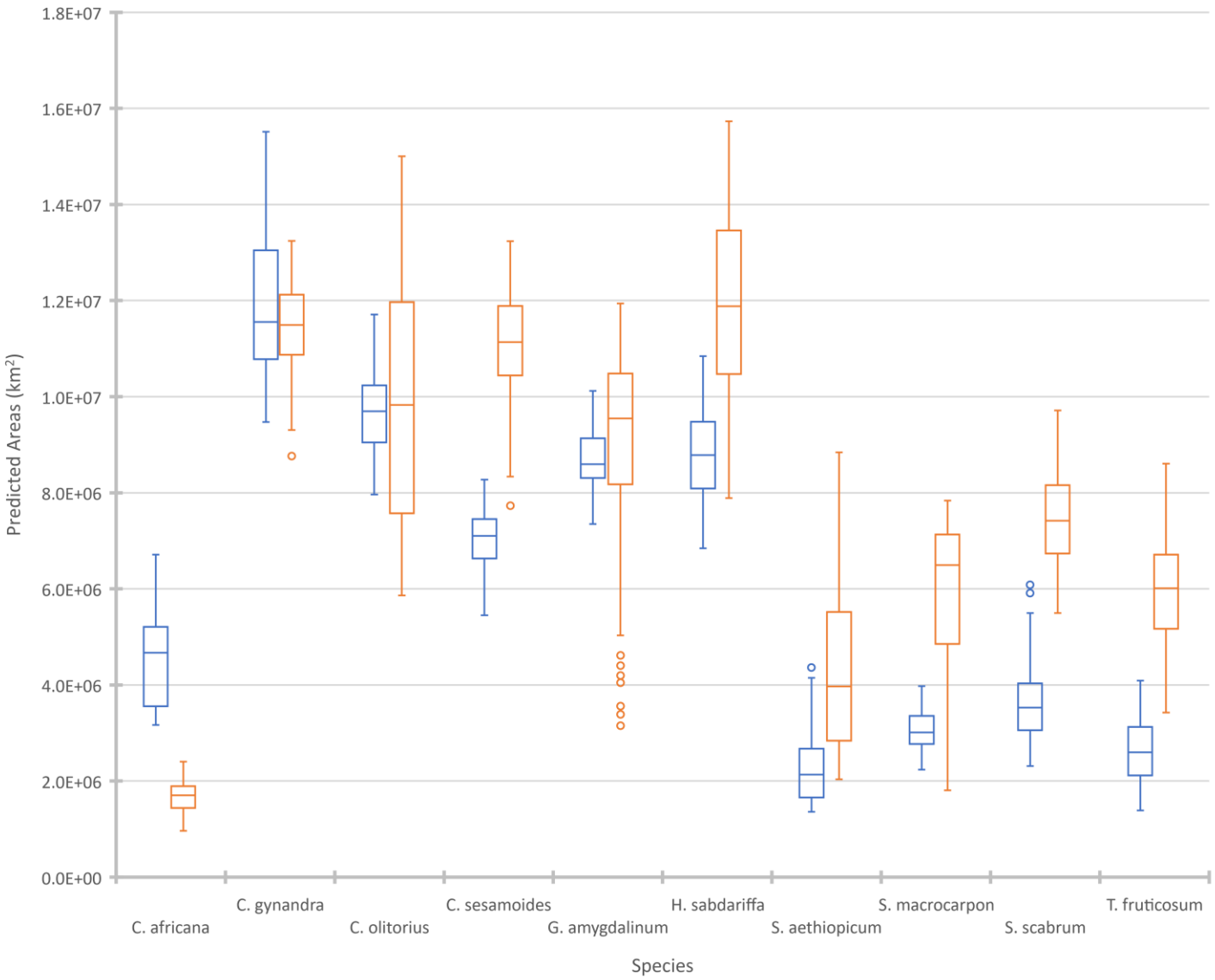

Figure 6: Box and whiskers plot of the predicted mean area of the present (1990) and the future (2070) of 10 indigenous vegetables. Blue represents the prediction for the present condition (mean of 1960-1990 climate data); brown represents future (2070) prediction.

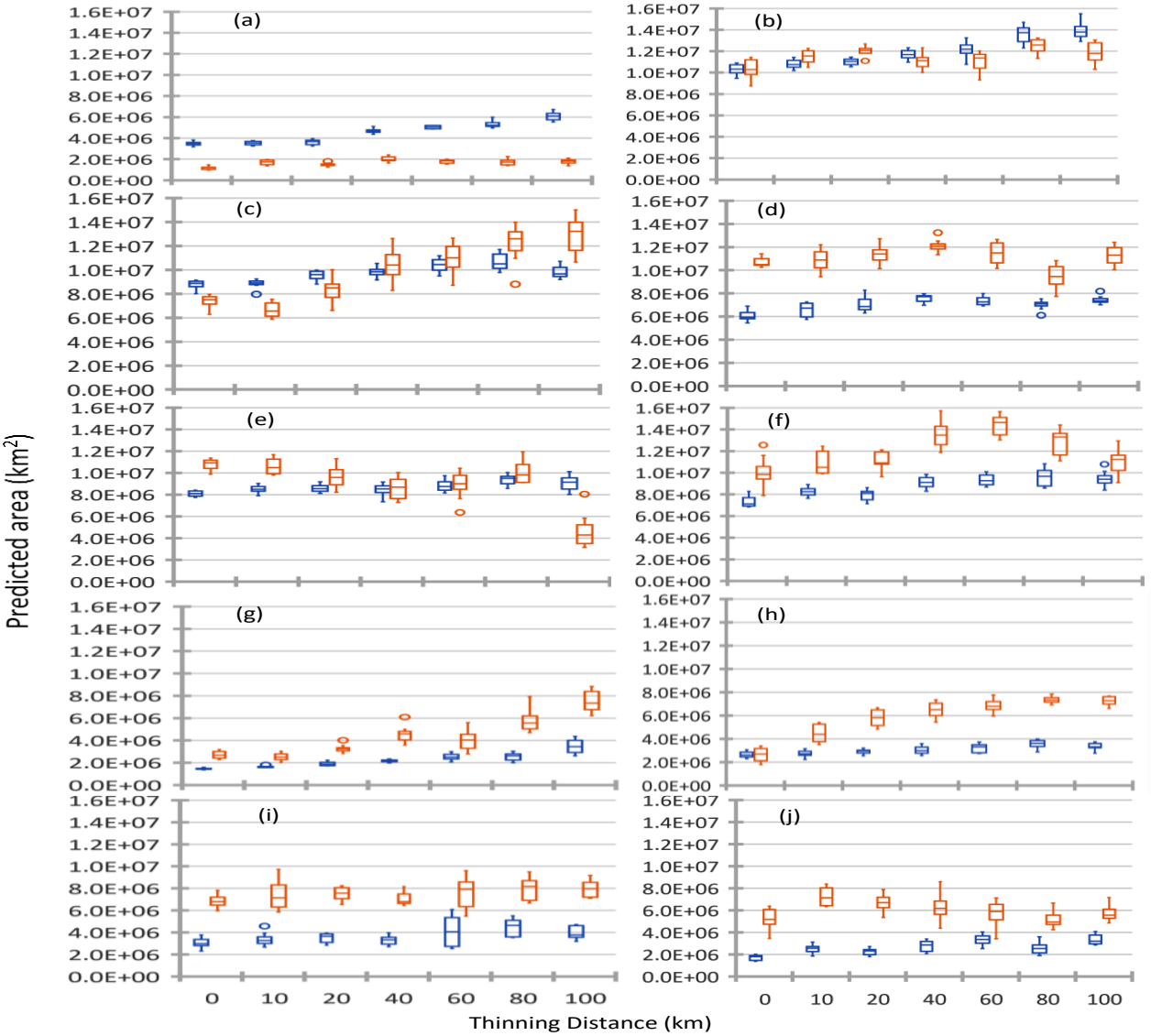

Figure 7: Box and Whiskers plot of present 1990 (blue) and future 2070 (brown) predicted areas (km^2) for different thinning distances for species of wild and semi-wild vegetables. Species are (a) C. africana; (b) C. gynandra; (c) C. olitorius; (d) C. sesamoides; (e) G. amygdalinum; (f) H. sabdariffa; (g) S. aethiopicum; (h) S. macrocarpon; (i) S. scabrum; and (j) T. fruticosum

Discussion

Global Biodiversity Information Facility (GBIF, 2018) is an important source of biodiversity information for research and probably is the largest single collection of digital species data (Hortal et al., 2008; Yesson et al., 2007). However, the data has been shown to be biased for some species examined (Bayor, 2012; Beck et al., 2014; Ruete, 2015). The magnitude of this bias may be unknown for any particular species (Yesson et al., 2007). However, use of these data for species distribution modelling must take into consideration the bias that may be inherent in the data (Fourcade et al., 2014; Phillips et al., 2009; Syfert et al., 2013). Spatial thinning (removing records such that the remaining are at least a specified distance apart) is one method that has been shown to be promising in correcting for spatial bias (Fourcade et al., 2014; Kramer-Schadt et al., 2013). In this experiment, clustering was not formally tested. However, visual plots (not shown) were examined and showed clustering to varying degrees among the various species.

Biased records when used to build a species distribution model results in inflated AUC values making models appear better than they really are (Bayor, 2012; Bean, Stafford & Brashares, 2012;

Hijmans, 2012). In this experiment the AUC values of the un-filtered data gave the highest AUC values probably confirming this issue.

Potential predicted areas were, however, the smallest. This probably suggests that, while the AUC values might be inflated, the areas predicted might be under-estimated (Bayor, 2012; Bean et al., 2012).

Although this experiment was not set-up to investigate sample size, it is known that insufficient sample size gives poor models (Wisz et al., 2008). The number of records required to build a reliable model depends on the characteristics of the species, whether it has wide range or is range restricted but in either case up to 20 or 25 records seems to be sufficient to build reliable models (Hernandez et al., 2006; Soultan & Safi, 2017). When records are spatially unbiased, even fewer records (less than 20) might still produce reliable models (Bean et al., 2012; van Proosdij et al., 2016). In our case the widest thinning (100km) left *Talinum fruticosum* (the species with the least records) with only 25 data points. Given the wide spacing applied to the data points, clustering may not be an issue, and this number of records is probably sufficient to give reliable models.

AUC, Boyce index, TSS and specificity all decreased with spatial thinning (Figure 1a-j; Figure 2a-j; Fig 3a-j; Figure 4a-j). This might suggest that, as the spatial thinning distance increased the ability of the model to discriminate between suitable and unsuitable areas reduced probably because fewer records were used for training and testing of the model. However, the purpose of the thinning is to improve accuracy by removing bias in the data. Although the validity of AUC and TSS as reliable measures of model accuracy with presence only data have been questioned many times (Jiménez-Valverde, Lobo, & Hortal, 2008; Leroy et al., 2017; Lobo et al., 2008; Somodi, Lepesi, & Botta-Dukát, 2017), the fact that there was consensus among all four measures might suggest in this instance the models are reliable. Thinning may therefore, involve sacrificing some level of model discriminatory power for improvement in predicted area.

The increase in spread of the model assessment parameters with thinning distance is also a pointer to the issue that when data points are lost, stability of the discriminatory power is lowered. It is therefore necessary to be mindful of the accuracy of the model assessment statistics when thinning is applied (Fourcade et al., 2014)

MAXENT (Phillips, Anderson, & Schapire, 2006) is a widely used species distribution modelling package (Guillera-Arroita, 2017) and may be robust even under small sample sizes (Hernandez et al., 2006) although Feeley and Silman (2011) suggest accurate models may require more data than previously thought. Graham et al. (2008) found MaxEnt gave accurate predictions even when location data were experimentally degraded probably suggesting MaxEnt is robust against locational data errors which are a common feature of databases such as GBIF (Beck et al., 2014; Hortal et al., 2008; Yesson et al., 2007). However, accounting for the bias in the use of GBIF data is important (Fourcade et al., 2014; Kramer-Schadt et al., 2013). Spatial thinning appears to be one of the high performing methods in correcting bias in location data during species distribution modelling (Aiello-Lammens et al., 2015; Kramer-Schadt et al., 2013). Our investigation in this experiment indicates that spatial thinning increases the predicted potential area and this may resolve the issue of under-prediction of the potential area (Bayor, 2012; Kramer-Schadt et al., 2013). For eight of the ten species modelled, the 100km distance thinning did not produce the largest predicted area, meaning that area does not continuously increase with increasing thinning. The thinning distance at which the maximum predicted area occurred varied with species and might also have to do with the nature (amount of clustering) of the data obtained. Optimal thinning distance, therefore, may depend on several factors and may have to be determined individually for each modelling situation.

The potential effect of climate change on the distribution of a species is influenced by the spatial distribution of the modelling data. Spatial thinning can alter the structure of the modelling data (Aiello-Lammens et al., 2015). This can have both beneficial and adverse effects on the final model. For example, in a rugged terrain, climatic variables may have narrow distribution ranges. Thinning may accidentally remove all records within a specific climatic range leading to climatic incompleteness resulting in a wrong model (Kadmon et al., 2003). In fact, this can apply to any environmental variable used in the modelling exercise. Over a range of thinning distances, it might probably cause incongruence between present and future prediction results of thinning at different spatial resolutions. Such situations can only be avoided by an understanding of the species range and characteristics.

Climatic completeness or having data that fully represent the climatic range of the species is a pre-requisite to a reliable and accurate model (Kadmon et al., 2003).

Conclusions

Future climate change would cause re-distribution of indigenous vegetables mostly by expansion of ranges but for *C. africana* there is going to be significant reduction. This may lead to unavailability of *C. africana* in some areas where it might have been used previously as a vegetable. We suggest cultivation and marketing as mechanisms to sustain availability where shortages may result. Spatial thinning increases potential predicted areas for both present and future. Thinning therefore appears to solve the important issue of under prediction as a result of spatial bias and can be used where data observation records are sufficient. Thinning may, however, inadvertently lead to variable incompleteness and may lead to unexpected results and therefore, should be used cautiously.

References

Achigan-Dako, E. G., N'Danikou, S., Assogba-Komlan, F., Ambrose-Oji, B., Ahanchede, A., & Pasquini, M. W. (2011). Diversity, geographical, and consumption patterns of traditional vegetables in sociolinguistic communities in Benin: implications for domestication and utilization. ***Economic Botany,*** *65*(2): 129–145. https://doi.org/10.1007/s12231-011-9153-4

Achigan-Dako, E. G., Sogbohossou, O. E. D., & Maundu, P. (2014). Current knowledge on Amaranthus spp.: Research avenues for improved nutritional value and yield in leafy amaranths in sub-Saharan Africa. ***Euphytica,*** *197*(3): 303–317. https://doi.org/10.1007/s10681-014-1081-9

Ahmed, S. E., Mcinerny, G., O'Hara, K., Harper, R., Salido, L., Emmott, S., & Joppa, L. N. (2015). Scientists and software - surveying the species distribution modelling community. ***Diversity and Distributions,*** *21*(3): 258–267. https://doi.org/10.1111/ddi.12305

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. ***Ecography,*** *38*(5): 541–545. https://doi.org/10.1111/ecog.01132

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. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Anderson, R. P. (2012). Harnessing the world's biodiversity data: Promise and peril in ecological niche modeling of species distributions. ***Annals of the New York Academy of Sciences,*** *1260*(1), 66–80. https://doi.org/10.1111/j.1749-6632.2011.06440.x

Bayor, H. (2012). *Diospyros in West Africa: Morphology, molecules and climate*. PhD Thesis. University of Reading, Reading.

Beale, C. M., & Lennon, J. J. (2012). Incorporating uncertainty in predictive species distribution modelling. ***Philosophical Transactions of the Royal Society B: Biological Sciences,*** *367*(1586): 247–258. https://doi.org/10.1098/rstb.2011.0178

Bean, W. T., Stafford, R., & Brashares, J. S. (2012). The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. ***Ecography,*** *35*(3): 250–258. https://doi.org/10.1111/j.1600-0587.2011.06545.x

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. ***Ecological Informatics***, *19*: 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Benito, B. M., Cayuela, L., & Albuquerque, F. S. (2013). The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: Guidelines to build better diversity models. ***Methods in Ecology and Evolution,*** *4*(4), 327–335. https://doi.org/10.1111/2041-210x.12022

Bloom, T. D. S., Flower, A., & DeChaine, E. G. (2017). Why georeferencing matters: Introducing a practical protocol to prepare species occurrence records for spatial analysis. ***Ecology and Evolution,*** (August 2017), 765–777. https://doi.org/10.1002/ece3.3516

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. ***Ecological Modelling,*** *275*: 73–77.

https://doi.org/10.1016/j.ecolmodel.2013.12.012

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. a. (2002). Evaluating resource selection functions. ***Ecological Modelling,*** *157*: 281–300.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., … Guisan, A. (2017). ecospat: an R package to support spatial analyses and modeling of species niches and distributions. ***Ecography,*** *40*(6): 774–787. https://doi.org/10.1111/ecog.02671

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carr, G., … Lautenbach, S. (2013). Collinearity : a review of methods to deal with it and a simulation study evaluating their performance. ***Ecography,*** *36*: 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Ebert, A. W. (2014). Potential of underutilized traditional vegetables and legume crops to contribute to food and nutritional security, income and more sustainable production systems. ***Sustainability (Switzerland)***, 6(1):

Elith, J., Graham, C. H., Anderson, R. P., 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. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Feeley, K. J., & Silman, M. R. (2011). Keep collecting: accurate species distribution modelling requires more collections than previously thought. ***Diversity and Distributions,*** 17(6): 1132–1140. https://doi.org/10.1111/j.1472-4642.2011.00813.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. https://doi.org/10.1017/S0376892997000088

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. ***Methods in Ecology and Evolution,*** 6(4): 424–438. https://doi.org/10.1111/2041-210X.12242

FOA/IIASA/ISRIC/ISS-CAS/JRC. (2012). Harmonized World Soil Database v1.21. Retrieved May 20, 2017, from http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/HWSD_Data.html?sb=4

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. ***PloS One***, *9*(5): e97122. https://doi.org/10.1371/journal.pone.0097122

GBIF. (2018). Global Biodiversity Information Facility. Retrieved January 15, 2018, from http://kr.mirror.gbif.org/portal/species/15461543/%3E

Gomes, V. H. F., Ijff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., Coelho, L. D. S., … Ter Steege, H. (2018). Species Distribution Modelling: Contrasting presence-only models with plot abundance data. ***Scientific Reports,*** *8*(1): 1–12. https://doi.org/10.1038/s41598-017-18927-1

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., … Zimmermann, N. (2007). The influence of spatial errors in species occurrence data used in distribution models. ***Journal of Applied Ecology,*** *45*(1): 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. ***Ecography,*** 40(2): 281–295. https://doi.org/10.1111/ecog.02445

Guisan, A., Thuiller, W., & Wilfried, T. (2005). Predicting species distribution : offering more than simple habitat models. ***Ecology Letters,*** 8: 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

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,*** 29*:* 773–785.

Hijmans, R. J. (2012). Cross-validation of species distribution models : removing spatial sorting bias and calibration with a null model. ***Ecology,*** 93(3): 679–688. https://doi.org/10.1890/11-0826.1

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. https://doi.org/10.1002/joc.1276

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. ***Ecological Modelling,*** 199(2): 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. ***Oikos,*** 117(6): 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. ***Global Ecology and Biogeography,*** 21(4): 498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. ***Diversity and Distributions,*** 14(6): 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x

Kadmon, R., Farber, O., & Danin, A. (2003). A Systematic Analysis of Factors Affecting the Performance of Climatic Envelope Models. ***Ecological Society of America,*** 13(3): 853–867.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. ***Diversity and Distributions***, 19(11): 1366–1379. https://doi.org/10.1111/ddi.12096

Leroy, B., Delsol, R., Hugueny, B., Meynard, C. N., Barhoumi, C., Barbet-Massin, M., & Bellard, C. (2017). Theoretical solutions for the evaluation of discrimination capacity of species distribution models. ***The Reprint Server for Biology***. https://doi.org/https://doi.org/10.1101/235770

Liu, C., White, M., & Newell, G. (2009). Measuring the accuracy of species distribution models: a review. In R. S. Anderssen, R. D. Braddock, & L. T. H. Newham (Eds.), *The 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation* (pp. 4241–4247). Cairns, Australia: The Modelling and Simulation Society of Australia and New Zealand Inc. and the International Association for Mathematics and Computers in Simulation. https://doi.org/10.1007/s12524-009-0005-y

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence-absence data. ***Ecography,*** *34*(2): 232–243. https://doi.org/10.1111/j.1600-0587.2010.06354.x

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. ***Global Ecology and Biogeography,*** 17(2): 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Luan, Y., Cui, X., & Ferrat, M. (2013). Historical trends of food self-sufficiency in Africa. ***Food Security,*** 5(3): 393–405. https://doi.org/10.1007/s12571-013-0260-1

Maundu, P., Achigan-Dako, E., & Morimoto, Y. (2009). Biodiversity of African Vegetables. In C. M. Shackleton, M. W. Pasquini, & A. W. Drescher (Eds.), ***African Indigenous Vegetables in Urban Agriculture*** (pp. 65–104). London: Earthscan. https://doi.org/10.4324/9781849770019

National Research Council. (1996). ***Lost Crops of Africa*** (Vol. II). Washington, D.C.: National Academies Press. https://doi.org/10.17226/2305

National Research Council. (2008). ***Lost Crops of Africa*** (Vol. III). Washington, D.C.: National Academies Press. https://doi.org/10.17226/2305

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. ***Progress in Physical Geography,*** 34(1): 3–22. https://doi.org/10.1177/0309133309355630

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

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only

distribution models : implications for background and pseudo-absence data. ***Ecological Applications***, 19(1): 181–197. https://doi.org/wiley.com/10.1890/07-2153.1.

PROTA4U. (2018). Plant Resources of Tropical Africa. Retrieved March 15, 2018, from http://www.prota4u.org/search.asp

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. ***Journal of Biogeography,*** 41(4): 629–643. https://doi.org/10.1111/jbi.12227

Raes, N., & Ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. ***Ecography,*** 3*0*(5): 727–736. https://doi.org/10.1111/j.2007.0906-7590.05041.x

Ramirez-Cabral, N. Y. Z., Kumar, L., & Shabani, F. (2017). Global alterations in areas of suitability for maize production from climate change and using a mechanistic species distribution model (CLIMEX). ***Scientific Reports***, 7(1): 1–13. https://doi.org/10.1038/s41598-017-05804-0

Ruete, A. (2015). Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. ***Biodiversity Data Journal***, 3, e5361. https://doi.org/10.3897/BDJ.3.e5361

Somodi, I., Lepesi, N., & Botta-Dukát, Z. (2017). Prevalence dependence in model goodness measures with special emphasis on true skill statistics. ***Ecology and Evolution,*** 7(3), 863–872. https://doi.org/10.1002/ece3.2654

Soultan, A., & Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. ***PLoS ONE***, 12(11): 1–19. https://doi.org/10.1371/journal.pone.0187906

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. ***Diversity and Distributions,*** 21(5): 595–608. https://doi.org/10.1111/ddi.12279

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. https://doi.org/10.1371/journal.pone.0055158

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. ***Diversity and Distributions,*** 13(4): 397–405. https://doi.org/10.1111/j.1472-4642.2007.00346.x

van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. ***Ecography,*** 39(6): 542–552. https://doi.org/10.1111/ecog.01509

van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., … Rose, S. K. (2011). The representative concentration pathways: An overview. ***Climatic Change,*** 109(1): 5–31. https://doi.org/10.1007/s10584-011-0148-z

Weinberger, K., & Pichop, G. N. (2009). Marketing of African Indigenous Vegetables along Urban and Peri-Urban Supply Chains in Sub-Saharan Africa. In C. M. Shackleton, M. W. Pasquin, & A. W. Drescher (Eds.), ***African Indigenous Vegetables in Urban Agriculture*** (pp. 225–244). London: Earthscan. Retrieved from https://www.researchgate.net/profile/Ray_Yu_Yang/publication/240098514_Nutritional_Contributions_of_Important_African_Indigenous_Vegetables/links/02e7e52fb2d04c6fd0000000.pdf#page=138

Wheeler, T. R., Craufurd, P. Q., Ellis, R. H., Porter, J. R., & Vara Prasad, P. V. (2000). Temperature variability and the yield of annual crops. ***Agriculture, Ecosystems and Environment,*** *82*(1–3): 159–167. https://doi.org/10.1016/S0167-8809(00)00224-3

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. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., … Culham, A. (2007). How global is the global biodiversity information facility? ***PLoS ONE,*** 2(11): e1124. https://doi.org/10.1371/journal.pone.0001124

APPENDICES

Appendix 1: Median (minimum and maximum in parenthesis) values of model assessment statistics for 10 species of African Indigenous vegetables at different spatial thinning. AUC= area under the curve of the receiver operating characteristic curve;TSS = True skill statistic.

	Thinning Distance (km)						
	0	10	20	40	60	80	100
Statistic/Species				*Commenlina africana*			
AUC	0.93 (0.9-0.95)	0.93 (0.91-0.94)	0.93 (0.91-0.94)	0.92 (0.88-0.93)	0.9 (0.87-0.94)	0.9 (0.82-0.93)	0.86 (0.85-0.92)
Sensitivity	0.89 (0.78-0.95)	0.89 (0.8-0.95)	0.88 (0.83-0.91)	0.88 (0.7-0.91)	0.87 (0.73-0.92)	0.88 (0.7-0.95)	0.82 (0.65-1)
Specificity	0.88 (0.87-0.89)	0.88 (0.87-0.89)	0.88 (0.87-0.88)	0.84 (0.83-0.85)	0.82 (0.81-0.84)	0.82 (0.79-0.84)	0.8 (0.78-0.81)
TSS	0.77 (0.67-0.83)	0.77 (0.69-0.82)	0.75 (0.71-0.79)	0.71 (0.55-0.75)	0.69 (0.57-0.75)	0.68 (0.53-0.77)	0.62 (0.45-0.79)
Boyce Index	0.94 (0.91-0.99)	0.91 (0.84-0.96)	0.93 (0.89-0.98)	0.92 (0.82-0.97)	0.87 (0.74-0.95)	0.81 (0.72-0.89)	0.81 (0.49-0.89)
				Cleome gynandra			
AUC	0.84 (0.79-0.85)	0.82 (0.79-0.88)	0.82 (0.79-0.87)	0.79 (0.71-0.87)	0.78 (0.7-0.84)	0.76 (0.7-0.84)	0.77 (0.68-0.81)
Sensitivity	0.85 (0.78-0.89)	0.85 (0.76-0.94)	0.84 (0.68-0.97)	0.84 (0.63-0.96)	0.8 (0.6-0.95)	0.81 (0.67-0.95)	0.82 (0.75-0.88)
Specificity	0.65 (0.63-0.66)	0.63 (0.62-0.65)	0.63 (0.61-0.65)	0.61 (0.58-0.64)	0.59 (0.57-0.62)	0.55 (0.5-0.61)	0.54 (0.51-0.57)
TSS	0.5 (0.41-0.53)	0.48 (0.39-0.57)	0.48 (0.33-0.59)	0.44 (0.26-0.54)	0.41 (0.21-0.53)	0.38 (0.26-0.49)	0.37 (0.29-0.45)
Boyce Index	0.87 (0.78-0.94)	0.84 (0.53-0.97)	0.87 (0.32-0.92)	0.81 (0.62-0.93)	0.78 (0.63-0.93)	0.74 (0.58-0.9)	0.79 (0.67-0.88)
				Corchorus olitorius			
AUC	0.87 (0.84-0.91)	0.88 (0.8-0.89)	0.84 (0.78-0.93)	0.83 (0.78-0.87)	0.8 (0.73-0.88)	0.79 (0.74-0.88)	0.77 (0.62-0.89)
Sensitivity	0.86 (0.77-0.94)	0.85 (0.74-0.89)	0.85 (0.7-1)	0.79 (0.71-0.88)	0.83 (0.6-1)	0.81 (0.69-0.92)	0.74 (0.27-1)
Specificity	0.71 (0.70-0.73)	0.71 (0.69-0.74)	0.69 (0.66-0.71)	0.68 (0.66-0.69)	0.65 (0.61-0.67)	0.65 (0.62-0.68)	0.68 (0.66-0.7)
TSS	0.57 (0.48-0.64)	0.57 (0.45-0.59)	0.53 (0.38-0.67)	0.47 (0.39-0.56)	0.47 (0.25-0.64)	0.46 (0.31-0.58)	0.41 (-0.03-0.66)
Boyce Index	0.84 (0.74-0.92)	0.86 (0.79-0.91)	0.82 (0.64-0.93)	0.81 (0.42-0.91)	0.76 (0.68-0.87)	0.69 (0.4-0.89)	0.66 (0.43-0.87)
				Ceratotheca sesamoides			
AUC	0.92 (0.86-0.94)	0.91 (0.85-0.96)	0.89 (0.86-0.94)	0.86 (0.82-0.95)	0.85 (0.77-0.91)	0.85 (0.77-0.93)	0.83 (0.74-0.91)
Sensitivity	0.86 (0.68-0.92)	0.83 (0.76-0.95)	0.82 (0.75-0.94)	0.81 (0.62-1)	0.73 (0.55-0.91)	0.79 (0.56-1)	0.72 (0.56-1)

Specificity	0.81 (0.80-0.82)	0.79 (0.78-0.81)	0.78 (0.76-0.81)	0.76 (0.74-0.78)	0.77 (0.75-0.79)	0.75 (0.73-0.77)	0.75 (0.73-0.77)
TSS	0.67 (0.50-0.74)	0.63 (0.56-0.73)	0.6 (0.53-0.71)	0.57 (0.39-0.77)	0.5 (0.3-0.68)	0.53 (0.33-0.76)	0.47 (0.31-0.74)
Boyce Index	0.88 (0.80-0.95)	0.87 (0.82-0.94)	0.88 (0.73-0.95)	0.74 (0.66-0.83)	0.8 (0.42-0.92)	0.79 (0.5-0.95)	0.76 (0.48-0.82)
	Gymnanthemum amygdalinum						
AUC	0.88 (0.83-0.9)	0.87 (0.8-0.93)	0.86 (0.77-0.91)	0.83 (0.77-0.87)	0.79 (0.72-0.9)	0.8 (0.75-0.89)	0.79 (0.7-0.88)
Sensitivity	0.83 (0.71-0.92)	0.81 (0.59-0.95)	0.82 (0.58-0.89)	0.78 (0.56-0.87)	0.73 (0.54-0.92)	0.75 (0.55-0.91)	0.65 (0.45-0.91)
Specificity	0.73 (0.71-0.75)	0.72 (0.71-0.75)	0.72 (0.71-0.74)	0.73 (0.7-0.75)	0.71 (0.68-0.72)	0.7 (0.67-0.73)	0.71 (0.67-0.77)
TSS	0.56 (0.44-0.66)	0.53 (0.33-0.67)	0.53 (0.31-0.61)	0.5 (0.29-0.6)	0.44 (0.24-0.6)	0.45 (0.24-0.62)	0.37 (0.17-0.6)
Boyce Index	0.88 (0.75-0.93)	0.79 (0.59-0.89)	0.83 (0.71-0.9)	0.73 (0.54-0.91)	0.66 (0.45-0.91)	0.65 (0.44-0.73)	0.56 (0.4-0.79)
	Hibiscus sabdariffa						
AUC	0.87 (0.72-0.95)	0.89 (0.75-0.97)	0.86 (0.78-0.9)	0.86 (0.77-0.92)	0.82 (0.67-0.91)	0.8 (0.61-0.89)	0.79 (0.54-0.91)
Sensitivity	0.83 (0.6-0.93)	0.84 (0.67-1)	0.82 (0.73-0.92)	0.8 (0.3-0.9)	0.89 (0.63-1)	0.71 (0.57-1)	0.71 (0.17-1)
Specificity	0.76 (0.72-0.79)	0.72 (0.71-0.77)	0.74 (0.72-0.75)	0.7 (0.66-0.75)	0.69 (0.67-0.71)	0.68 (0.66-0.71)	0.67 (0.64-0.7)
TSS	0.58 (0.37-0.7)	0.56 (0.42-0.72)	0.56 (0.46-0.65)	0.5 (0.05-0.6)	0.57 (0.33-0.69)	0.4 (0.25-0.68)	0.39 (-0.15-0.67)
Boyce Index	0.82 (0.62-0.96)	0.85 (0.4-0.94)	0.87 (0.76-0.92)	0.83 (0.65-0.96)	0.81 (0.55-0.91)	0.81 (0.63-0.9)	0.77 (0.58-0.89)
	Solanum aethiopicum						
AUC	0.95 (0.93-0.97)	0.95 (0.93-0.96)	0.95 (0.92-0.96)	0.94 (0.91-0.97)	0.93 (0.87-0.96)	0.93 (0.86-0.97)	0.89 (0.85-0.97)
Sensitivity	0.87 (0.74-0.97)	0.85 (0.7-0.93)	0.86 (0.55-0.91)	0.81 (0.67-0.94)	0.79 (0.42-0.92)	0.78 (0.4-1)	0.73 (0.5-1)
Specificity	0.95 (0.95-0.95)	0.94 (0.93-0.95)	0.94 (0.93-0.94)	0.93 (0.91-0.94)	0.92 (0.9-0.93)	0.91 (0.9-0.93)	0.88 (0.87-0.9)
TSS	0.82 (0.69-0.92)	0.79 (0.65-0.86)	0.8 (0.49-0.85)	0.74 (0.6-0.86)	0.71 (0.34-0.83)	0.69 (0.33-0.91)	0.61 (0.39-0.87)
Boyce Index	0.89 (0.82-0.96)	0.89 (0.81-0.94)	0.87 (0.65-0.94)	0.82 (0.66-0.87)	0.78 (0.52-0.95)	0.69 (0.54-0.83)	0.76 (0.64-0.92)
	Solanum macrocarpom						
AUC	0.94 (0.9-0.97)	0.94 (0.9-0.96)	0.94 (0.89-0.96)	0.94 (0.91-0.95)	0.92 (0.88-0.97)	0.91 (0.83-0.97)	0.91 (0.8-0.96)
Sensitivity	0.87 (0.56-1)	0.85 (0.54-0.92)	0.83 (0.67-1)	0.81 (0.64-0.9)	0.81 (0.5-1)	0.75 (0.57-1)	0.71 (0.17-1)
Specificity	0.92	0.91	0.91	0.91	0.89	0.87	0.88

	(0.9-0.93)	(0.9-0.92)	(0.89-0.92)	(0.89-0.92)	(0.88-0.91)	(0.86-0.88)	(0.85-0.9)
TSS	0.78	0.75	0.72	0.72	0.7	0.62	0.61
	(0.49-0.91)	(0.45-0.83)	(0.58-0.91)	(0.54-0.81)	(0.41-0.9)	(0.44-0.86)	(0.06-0.89)
Boyce Index	0.87	0.79	0.79	0.78	0.78	0.79	0.76
	(0.66-0.91)	(0.46-0.95)	(0.56-0.83)	(0.54-0.92)	(0.39-0.94)	(0.24-0.92)	(0.57-0.88)
	Solanum scabrum						
AUC	0.95	0.94	0.93	0.91	0.93	0.9	0.89
	(0.87-0.98)	(0.84-0.99)	(0.81-0.98)	(0.8-0.96)	(0.74-0.96)	(0.8-0.95)	(0.8-0.96)
Sensitivity	0.83	1.00	0.78	0.75	1.00	0.67	0.67
	(0.5-1)	(0.2-1)	(0.5-1)	(0.5-1)	(0.25-1)	(0.67-1)	(0.33-1)
Specificity	0.91	0.89	0.88	0.88	0.86	0.84	0.86
	(0.88-0.91)	(0.85-0.91)	(0.87-0.89)	(0.86-0.91)	(0.81-0.91)	(0.81-0.86)	(0.82-0.91)
TSS	0.74	0.86	0.65	0.63	0.82	0.53	0.54
	(0.41-0.91)	(0.11-0.9)	(0.37-0.88)	(0.4-0.87)	(0.12-0.89)	(0.51-0.84)	(0.24-0.86)
Boyce Index	0.84	0.82	0.85	0.82	0.82	0.9	0.86
	(0.65-0.95)	(0.69-0.92)	(0.49-0.92)	(0.73-0.92)	(-0.16-0.93)	(0.57-0.99)	(0.3-0.91)
	Talinum fruticosum						
AUC	0.97	0.94	0.95	0.93	0.9	0.92	0.91
	(0.9-0.99)	(0.9-1)	(0.89-0.99)	(0.88-0.99)	(0.84-0.99)	(0.88-0.96)	(0.83-0.97)
Sensitivity	0.91	0.73	0.67	0.75	0.67	0.67	0.67
	(0.55-1)	(0.57-1)	(0.5-1)	(0.25-1)	(0.33-1)	(0-1)	(0.33-1)
Specificity	0.94	0.92	0.93	0.91	0.89	0.92	0.90
	(0.92-0.96)	(0.9-0.94)	(0.91-0.94)	(0.89-0.93)	(0.86-0.92)	(0.9-0.93)	(0.86-0.91)
TSS	0.85	0.67	0.6	0.65	0.57	0.58	0.57
	(0.51-0.94)	(0.50-0.91)	(0.44-0.93)	(0.18-0.91)	(0.25-0.89)	(-0.09-0.92)	(0.23-0.91)
Boyce Index	0.82	0.84	0.83	0.81	0.83	0.84	0.83
	(0.64-0.94)	(0.59-0.95)	(0.65-0.92)	(0.27-0.87)	(0.47-0.90)	(0.02-0.94)	(0.59-0.94)