iForest Biogeosciences and Forestry

Research Article
doi: 10.3832/ifor2767-012
vol. 12, pp. 418-426

Essential environmental variables to include in a stratified sampling design for a national-level invasive alien tree survey

Johann DF Kotze [(1-2)], **Hein B Beukes** [(1)], **Thomas Seifert** [(2-3)]

There is a direct relationship between the abundance of biological invasions and their impact, which means that it is important to capture spatial patterns in their abundance and use this information to focus management actions. However, protocols to objectively determine invasive alien plant (IAP) distributions and abundance are lacking at a national level, resulting in the inability to determine and monitor changes in spatial extent and density over time. A complete inventory of IAP spatial distribution across an extensive area such as South Africa is not possible and so requires an efficient sampling approach. A simple random sampling design would not be efficient, so monitoring of IAP species at a national level requires an appropriate sampling design such as a stratified sampling. The selection of environmental variables to be included in such a stratification should be based on the relationship between IAP species and their physical environment to successfully summarize variance in their abundance within the different strata. A further objective is to obtain all possible combinations of environmental variables or a full rank design in the stratification to allow for the comparison of different strata based on actual field sampled data. This raises the question of which predictive environmental variables as well as how many to include in the stratification. For this purpose, three invasive tree species, namely *Acacia cyclops*, *Acacia mearnsii* and *Prosopis glandulosa* were selected as they cover the maximum possible area at the highest density with the least amount of geographic overlap. A total of 26 environmental variables that included climatic, soil and topographic type variables were tested with linear regressions against correlations with the abundance of those tree species. The results showed that a combination of average precipitation, soil depth, clay content in the B-horizon and terrain morphological units will serve as a suitable stratification at a national level to explain IAP abundance variation sufficiently well whilst retaining a full rank design. These results will be applied as the first phase in the formation of a regional level IAP monitoring programme for South Africa on a scientific basis.

Keywords: Invasive Alien Plant (IAP) Species, Monitoring, Sampling Design, Stratification, Environmental Variables

Introduction

Alien plant invasions are known to have severe disruptive impacts on biodiversity, ecosystems, plant and animal populations, ecosystem services, agriculture, forestry, the economy and human welfare (Jeschke et al. 2014, Vilà & Hulme 2017). One of the most important attributes of biological invasions in terms of impact is invasive species abundance (Kumschick et al. 2015). In other words, the more there is of an invasive alien plant (IAP) species, whether number of individual plants or biomass, the greater the impact. Thus in particular invasive tree species with their large biomass and their ability to change their local environment substantially have an impact on ecosystems and ecosystem services (Le Maitre et al. 2016).

Mitigation strategies to deal with alien plant invasions have been implemented across the world with noted successes in the control of invasive species (Simberloff et al. 2011). In South Africa, the long-term Working for Water Programme was initiated in 1996 as an IAP control programme sponsored by government (Van Wilgen et al. 2012). South Africa, with its rich biodiversity (Driver et al. 2012), has been invaded by many different IAP species, especially tree species (Nel et al. 2004), and the ecological and economic impacts of these invasions have been well documented (De Lange & Van Wilgen 2010, Le Maitre et al. 2016). Further to this, South Africa hosts three of the 35 current biodiversity hotspots in the world (Mittermeier et al. 2011). This makes the threat of IAP species to this region an international concern (Mitter-

(1) Institute for Soil, Climate and Water, Agricultural Research Council, Private Bag X79, Pretoria, 0001 (South Africa); (2) Stellenbosch University, Department of Forest and Wood Science, Faculty of AgriSciences, Private Bag X1, Matieland, 7602 (South Africa); (3) Chair of Forest Growth, Albert-Ludwigs-University Freiburg, Tennenbachstraße 4, 79106 Freiburg (Germany)

@ Johann DF Kotze (kotzei@arc.agric.za)

Received: Feb 23, 2018 - Accepted: Jun 12, 2019

Citation: Kotze JDF, Beukes HB, Seifert T (2019). Essential environmental variables to include in a stratified sampling design for a national-level invasive alien tree survey. iForest 12: 418-426. – doi: 10.3832/ifor2767-012 [online 2019-09-01]

Communicated by: Francisco Lloret Maya

meier et al. 2011). Species abundance data is essential in the effective management of such control programmes and serves as an important indicator in the measurement of their success (Wilson et al. 2018). To prioritise intervention or mitigation strategies at a national level, it is important to achieve IAP distribution and abundance data at this scale. Despite the success of many of these initiatives, they still lack sound protocols for objectively determining IAP distribution at a national level, with the obvious result of not being able to measure and monitor actual IAP spatial extent and abundance changes over time (Dehnen-Schmutz et al. 2018).

The spatial extent of the study area (South Africa covers approximately 122 million hectares), the environmental and ecological heterogeneity (Driver et al. 2012), and limited resources to conduct surveys (Ricciardi et al. 2017), make a complete IAP inventory not feasible to carry out (Webster & Lark 2013). The best alternative for providing unbiased and reliable quantitative information is a partial estimation based on sampling (Gitzen et al. 2012). An example of this is the statistical or sample based surveys which have been applied for many years in most large scale forestry surveys in many countries (Ståhl et al. 2016), and are based on strict design-based principles (Naesset et al. 2011). The success of such monitoring programmes is determined by the underlying sampling design (Gitzen et al. 2012). Ideally, the sampling strategy should effectively represent the variability of the entire target population with as few as possible sample points. The simple random sampling design is known to be inefficient in providing an even representative coverage of a study area, due to the tendency of sample point locations to cluster at low sampling intensities, resulting in large undetected areas (Webster & Lark 2013). The result is that resource demands such as costs, manpower and time required for random sampling designs are high (Kalkhan 2011) if the aim is to ensure that the inherent variation in the target population is represented (Webster & Lark 2013). An alternative is to use a pre-stratified sampling design which improves the accuracy of the estimates and allows for a better efficiency (Webster & Lark 2013). The objective of stratification for vegetation surveys is to incorporate those habitat types that show the most meaningful association with the vegetation attribute of interest, and so to ensure that all the possible habitat specific variation that contributes to the target species range and abundance is included in the survey (Gitzen et al. 2012). The latter also provides well-defined strata, which allows for effective comparisons across strata for valid inference between field observations (Webster & Lark 2013). The challenge in this context is the selection of environmental variables which adequately define spatial units representing homogenous conditions or strata for species abundance. These strata should clearly reflect the relationship between IAP species and their physical environment and thereby summarize this underlying non-random relationship (Volis 2016). The main aim of stratification is thus to minimise the variance within the strata while maximising the variance between them. All of which leads to the main question: which predictive environmental variables and how many of them should be included for defining the strata while maintaining a full rank design. Such a design provides the most effective inference between species' observations obtained from actual field surveys.

Appropriate methods to model the correlation between species' occurrence and environmental variables such as climate, soil and terrain are predictive vegetation or species distribution models (SDM – Hageer et al. 2017). SDMs not only provide insights into the species-environment relationships, but they are also used to predict spatial distributions of target species by means of maps of the correlated environmental predictor variables (Elith & Franklin 2017). Multiple ways have been proposed to model species distribution and prominent examples include regression trees, boosted regression trees and random forests machine learning algorithms that are used to combine rules for species occurrence in an optimum way (Franklin 2010). Examples for rule-based systems are GARP (Stockwell 1999) that applies a genetic algorithm or MaxEnt (Anderson et al. 2003) that works on a maximum entropy optimisation. Other authors applied a traditional parametric algorithm such as regression analysis (Fahrmeir et al. 2013). Most methods are known to provide equally good results (Aitor & Garcia-Viñas 2011, Sahragard & Ajorloa 2018). Species distribution modelling has been widely applied in the field of invasion biology for a range of objectives (see Robinson et al. 2017 for a review). For instance, Rouget et al. (2015) used broad scale predictor variables that included climate, natural biomes and anthropogenic factors in relationship to the distribution of IAP species' assemblages in an effort to map wide-ranging alien plant biomes. Applications also include the use of models to support the development of appropriate sampling designs and includes the definition of appropriate strata (Särndal 2010).

In this study we assessed the extent to which the modelled associations between IAP species and environmental variables were meaningful based on repeated correlation patterns across extensive areas with high levels of environmental variation. We hypothesised that, although localized associations between IAP species and different environmental variables might vary, there would be constant regional correlation patterns with a limited number of specific variables.

The objective of the stratification process was to obtain all possible combinations or interactions between the different environmental variables. This full rank design is advantageous from a statistical point of view and easily obtained in a controlled environment or a planned experiment. The challenge is to obtain such a design within the natural environment at a national scale. Thus the aim of this study was to combine the unique and varying natural geographical distribution patterns of the underlying deterministic environmental variables to effectively summarise IAP species' abundance within these strata, whilst maintaining a complete full rank design.

This paper presents an approach to firstly filter and select environmental variables most suitable for such a national-level design-based stratification in South Africa and, secondly, to explore how many categories could realistically be included in such a stratification. The results represent the first phase in establishing a regional level IAP monitoring programme for South Africa on a scientific and statistically rigorous basis.

Materials and methods

Study area

The study area was the whole of South Africa and the focus was on undisturbed areas or rather natural and semi-natural areas or habitats as defined by Nel et al. (2004), namely: "natural and semi-natural ecosystems, that is, those that are still reasonably intact, having most of their biodiversity structure and functioning, and with primary driving forces operating within natural/evolutionary limits". These habitats are most threatened by IAP species by having the greatest impact on native biodiversity and ecosystem services (Nel et al. 2004).

IAP distribution records

The most comprehensive set of records of the spatial distribution of IAP species for the study area is the Southern African Plant Invaders Atlas (SAPIA) database that contains records for more than 500 different IAP species (Henderson & Wilson 2017). SAPIA observed IAP species with no underlying statistical basis along road transects of 5-10 km long and within the adjacent road area from a moving vehicle (Henderson & Wilson 2017). IAP species were mostly recorded per quarter degree square (QDS), a 15′ latitude × 15′ longitude square, therefore the exact location of species was related to a total area of approximately 25×27 km or 65,000 ha. As many as 120 different IAP species were recorded per QDS and often with repetitive observations per species. An abundance value is provided for each record based on the approximate number of actual plants observed per unique IAP species within a 10 km transect. A number of habitat classes are also provided per species record to allow for a species to be classified based on habitat preference. Many of these species have a limited distribution and abundance, overlap in distribution and are biased towards certain habitat classes, so the SAPIA database was filtered for the study using a stepwise rule-based approach (for further details on the species filter process, see Appendix 1 and Fig. S1 in Supplementary material). Species were firstly selected on the basis of having the maximum distribution range across South Africa at a high abundance. This captured the full environmental gradient contributing to a particular IAP species' observed distribution. Subsequently, species with minimum overlap in spatial extent with other species were identified to create mutually exclusive observations for each of the IAP species across geographic space. The combination of maximum spatial distribution with minimum overlap led to the selection of three tree species, namely *Acacia cyclops, Acacia mearnsii* and *Prosopis glandulosa* (Fig. 1). Matrices produced for each of the species consisted of the total abundance values for that particular species in a given location.

Environmental variables

A set of physiologically relevant environmental variables that have been shown to correlate with species abundance were included, namely climatic, topographic (terrain) and soil related variables (Williams et al. 2012, Hageer et al. 2017, Fois et al. 2018). The climatic variables were obtained from the WorldClim2 dataset (Fick & Hijmans 2017). Soil variables were extracted from the South African Land Type Survey database, which is based on detailed field surveys published at a 1:250,000 scale (Land Type Survey Staff 2006). Terrain variables such as aspect were derived from the Shuttle Radar Topographic Mission (SRTM) digital elevation data at the 90 m resolution (Farr et al. 2007).

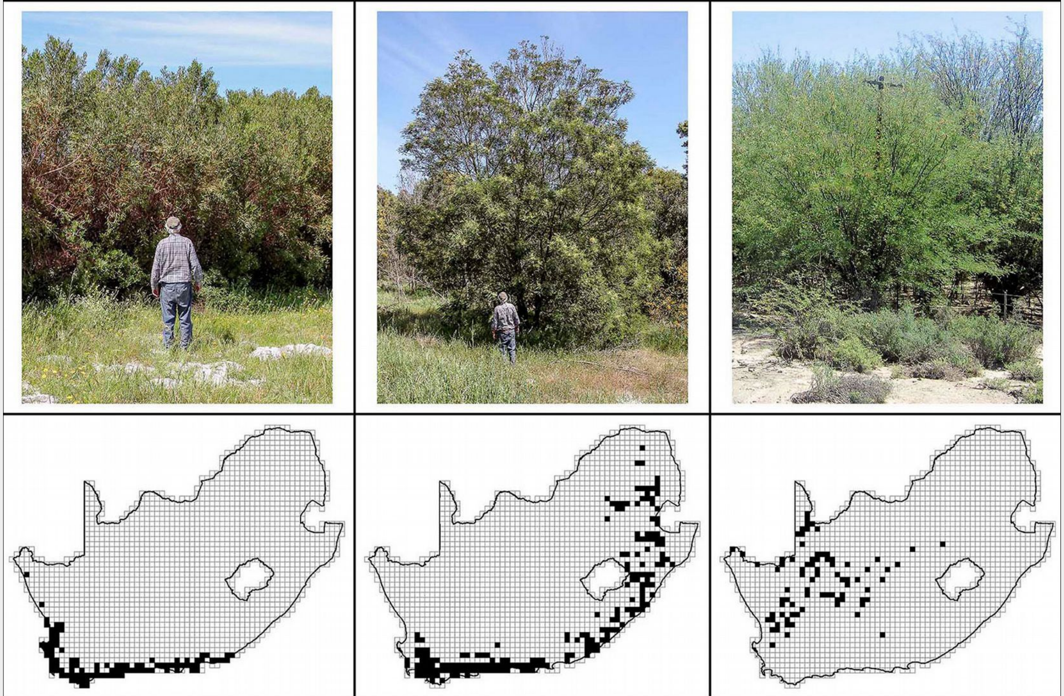
Environmental variables were resampled to a 400 × 400 m spatial resolution where required (Tab. 1). Multicollinearity (Franklin 2010) amongst predictor variables was assessed by means of the pair-wise correlation coefficient between variables (Williams et al. 2012). Pairwise correlation exceeding a threshold collinearity of more than 0.75 (Dormann et al. 2013) was used to exclude variables.

Spatial combination of species presence with environmental variables

Each of the three species' layers was spatially intersected with the overlapping environmental variable matrices to create three unique species/environmental datasets by means of ArcGIS® Desktop software (ESRI 2017). The application of the South African tertiary catchment delineation as an aggregation unit supplied replications across geographic space for each of these three layers. Catchment delineation was applied as an aggregation unit for it is defined by topography. Catchments therefore captures the full range of terrain morphological units which ensures that soil and climatic gradients are included. Adjacent catchments have closely matching gradients of these variables which makes them reasonable replicates for determining strata. Further to this, catchment delin-

Fig. 1 - A description and distribution of the three identified species, namely *Acacia cyclops, Acacia mearnsii* and *Prosopis glandulosa*.

Acacia cyclops	***Acacia mearnsii***	***Prosopus glandulosa***



Introduced from Australia, it forms dense evergreen shrubs to an eight metre tall tree with an average height of three metres. Plants carry lots of dead wood and old seed pods. It grows mostly in the winter rainfall areas or areas that receives rainfall throughout the year and typically below 300 metres above sea level. This species occurs on acidic and calcareous sands and often dominates low lying coastal flats.	This evergreen tree originating from Australia reaches heights of between 10 to 30 metres. Trees carry a fair amount of old wood and pods as well as large amounts of conspicuous pale yellow flowers during late winter and early spring. It occurs from sea level to high altitude areas across a broad rainfall region. Fairly drought and frost resistant, but prefers more temperate climates and lower lying valley bottoms. It occurs over a broad spectrum of soils.	This species originates from the South-western United States and Northern Mexico. It typically occurs in the more arid regions of South Africa. Favours habitats with deep soils and where ground-water is available which includes riparian zones, seasonal watercourses, pans and depressions. It usually grows as multi-stemmed shrubs to small trees averaging from 2 to 4 metres. Can tolerate extended drought periods.

Tab. 1 - Environmental variables used in the analysis.

Type	Description	Resolution (m)	Source
Climate	Annual Mean Temperature	1000 × 1000	Fick & Hijmans 2017
	Mean Diurnal Range (Mean of monthly [Max Temp - Min Temp])	1000 × 1000	
	Isothermality (Mean Diurnal Range / Temperature Annual Range) (×100)	1000 × 1000	
	Temperature Seasonality (standard deviation ×100)	1000 × 1000	
	Max Temperature of Warmest Month	1000 × 1000	
	Min Temperature of Coldest Month	1000 × 1000	
	Temperature Annual Range (Max Temperature of Warmest Month - Min Temperature of Coldest Month)	1000 × 1000	
	Mean Temperature of Wettest Quarter	1000 × 1000	
	Mean Temperature of Driest Quarter	1000 × 1000	
	Mean Temperature of Warmest Quarter	1000 × 1000	
	Mean Temperature of Coldest Quarter	1000 × 1000	
	Annual Precipitation	1000 × 1000	
	Precipitation of Wettest Month	1000 × 1000	
	Precipitation of Driest Month	1000 × 1000	
	Precipitation Seasonality (Coefficient of Variation)	1000 × 1000	
	Precipitation of Wettest Quarter	1000 × 1000	
	Precipitation of Driest Quarter	1000 × 1000	
	Precipitation of Warmest Quarter	1000 × 1000	
	Precipitation of Coldest Quarter	1000 × 1000	
Soil	Soil depth (mm)	400 × 400	Land Type Survey Staff 2006
	Percentage clay in the A-horizon	400 × 400	
	Percentage clay in the B-horizon	400 × 400	
Terrain	Terrain morphological units (valley bottom, footslope, midslope, scarp and crest)	90 × 90	Land Type Survey Staff 2006
	Elevation (m a.s.l.)	90 × 90	Farr et al. 2007
	Aspect	90 × 90	
	Slope (%)	90 × 90	

eation is also applied to define management units in the Working for Water Programme.

The most detailed or highest order catchment delineation for the country is represented by quaternary catchments of which there are approximately 1845. Most of these quaternary catchments were simply too small to include a sufficient species abundance gradient. For this reason, the next level of catchment delineation was chosen, namely tertiary catchment delineation. There are 274 tertiary catchments which provided a sufficient species abundance gradient due to a much larger surface area (mean area: 455,520 ha).

Environmental modelling

Species abundance served as the response variable, whilst the environmental variables were applied as predictor variables. Relationships between the response variables and each predictor variable were investigated by means of visual inspection of the resulting graphs to determine the type of correlation models to use from the wide range of techniques available for modelling species-environment associations. Data distribution guides the selection of the type of modelling approach to apply (Dormann 2011). The relationships were linear, resulting in opting for a more traditional modelling approach, namely linear regression models (Dormann 2011). These were developed based on the generalized linear model (GLM) framework (Fahrmeir et al. 2013). GLM's are extensively applied in species distribution modelling due to their strong statistical foundation and ability to realistically model species-environment associations (Elith & Franklin 2017).

Data outliers were identified and subsequently removed for each environmental variable per tertiary catchment based on set cut-off limits applied to the left and right of the normal distributions per variable. Cut-off limits were based on the variable's coefficient of variation (CV). For instance, for a CV<10 the applied cut-off z-value is 1.96, whilst for a CV<20 the applied cut-off z-value was decreased to 1.65.

The three IAP species were investigated and analysed independently per tertiary catchment. Statistical analysis was conducted by means of Matlab® software (MathWorks 2017). The environmental predictor variables were added one at a time to the model, and all possible combinations of the number and type of variables were explored for their effect on model performance. This resulted in a multitude of models per tertiary catchment for each species. Akaike's Information Criterion (AIC) was used to evaluate models per catchment and select the most appropriate model with its associated predictor environmental variables (Symonds & Moussalli 2011).

Stratification simulation

Environmental variables were reclassified into three classes each based on a gradient ranging from low to medium and finally high for two aggregation levels or spatial scales, namely the complete study area (South Africa) and the tertiary catchment delineation. Interaction classes between variables were created by intersecting the different variables in geographic space for these two aggregation levels by progressively increasing the number of environmental variables in subsequent intersections (see Appendix 2 in Supplementary material for an explanation on the stratification procedures). The number of interaction classes created at the two aggregation levels at each intersection level were compared with the number of classes required for an ideal theoretical full rank design. This comparison provided an indication of an appropriate number of variables to be included in such a stratification exercise to achieve a design as close as possible to, if not a complete factorial design.

The three IAP species were then combined with the created strata at the maximum identified intersection level before actual stratification started to deviate from the ideal full rank design. The effectiveness of the identified environmental variables and the subsequent stratification to reduce the variation of IAP distribution and abundance within strata was compared to the overall variation without any form of stratification at a tertiary catchment level. Should the stratification be meaningful, IAP abundance variation at a stratum level would be significantly less than at an unstratified level on a repetitive basis across tertiary catchments. An analysis of variance (ANOVA) was then applied to the data to simulate a data analysis using the data as if it came from an actual IAP survey to see if there was a significant association between IAP distributions and the respective strata. Should there be no association between IAP distribution and respective strata, therefore IAP abundance varied at random across strata, the use of strata as categories to describe IAP distributions as a response variable within them would be meaningless.

Results

Environmental modelling

Fourteen variables from the original 26 remained after testing for multicollinearity among predictors. A threshold was applied to select environmental variables most frequently associated with the three IAP species, namely those variables repetitively observed more than 75% per species across tertiary catchments. In the case of *A. cyclops* these variables included soil depth, percentage clay in the A-horizon, percentage clay in the B-horizon, slope and the terrain morphological units. Variables most frequently associated with *A. mearnsii* were the terrain morphological units, percentage clay in the A-horizon, percentage clay in the B-horizon, soil depth, long-term mean annual precipitation and isothermality. *P. glandulosa* was mostly associated with clay in the B-horizon, soil depth and long-term mean annual precipitation (Tab. 2). The total percentage association between environmental variables and the three species combined was then determined to provide an overall indication of association per variable across all species. (Fig. 2). Further filtering of variables was based on a combination of ecological reasoning (Dormann et al. 2013) and the frequency of occurrence of variables for all three IAP species.

Stratification simulation

The stratification of the complete study area up to the spatial intersection of five environmental variables generated 243 classes, which was similar to the total number of possible classes at that level for a full rank design within a controlled experiment (Fig. 3). Stratification at the smaller aggregation level, namely the tertiary catchment delineation, started to deviate from a full rank design with the intersection of between three and four variables, therefore between 27 and 81 classes. When this was done with five or more variables with three levels each, the stratification at a tertiary catchment level started to deviate substantially from the total amount of all possible class combinations (Fig. 3). The

Tab. 2 - Environmental variables associated the most frequently with the different species (>75%).

Environmental Variables	IAP Species		
	A. cyclops	*A. mearnsii*	*P. glandulosa*
Annual precipitation	-	×	×
Percentage clay in the A-horizon	×	×	-
Percentage clay in the B-horizon	×	×	×
Soil depth	×	×	×
Isothermality	-	×	-
Slope	×	-	-
Terrain morphological units	×	×	-

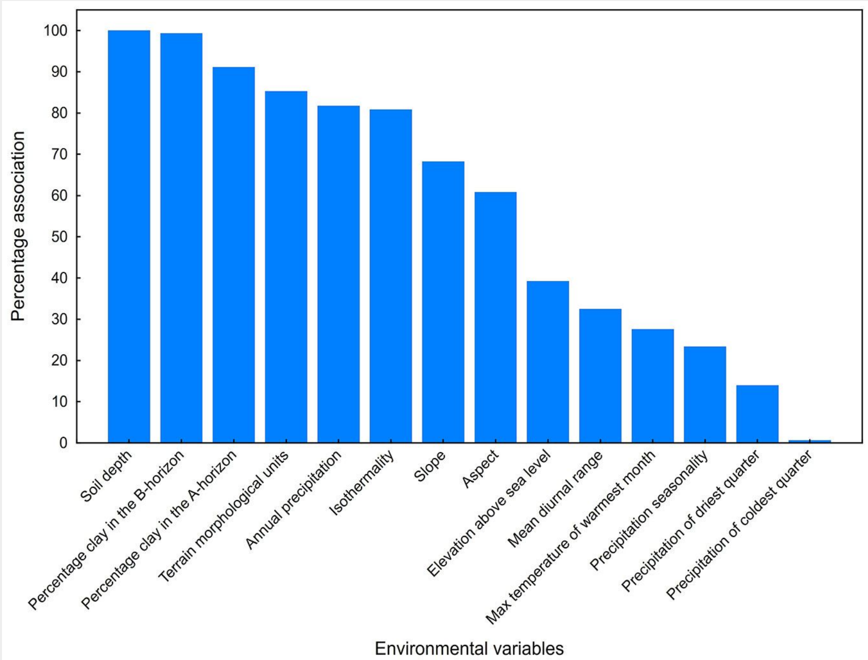


Fig. 2 - The total percentage association between the specific predictor environmental variables and the three tree species combined for all tertiary catchments.

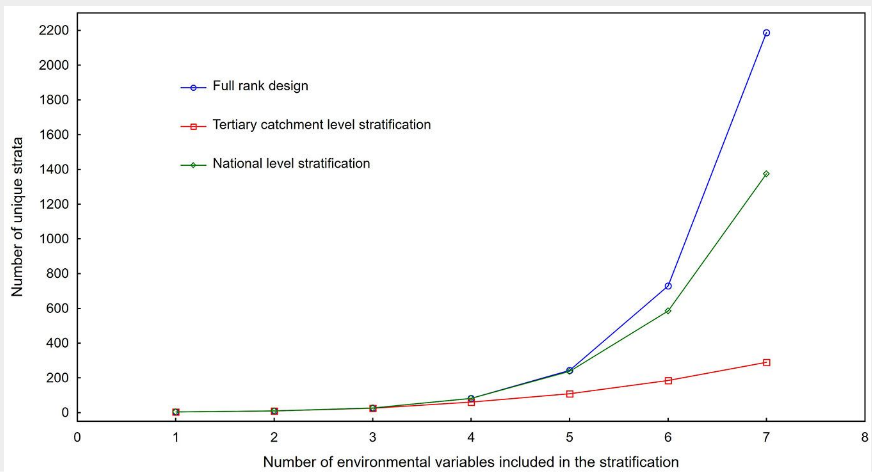


Fig. 3 - The number of unique strata created by means of intersecting environmental variables. The graph only includes up to the intersection of seven variables with three even area classes each for thereafter the difference in number of obtained intersection classes only increases.

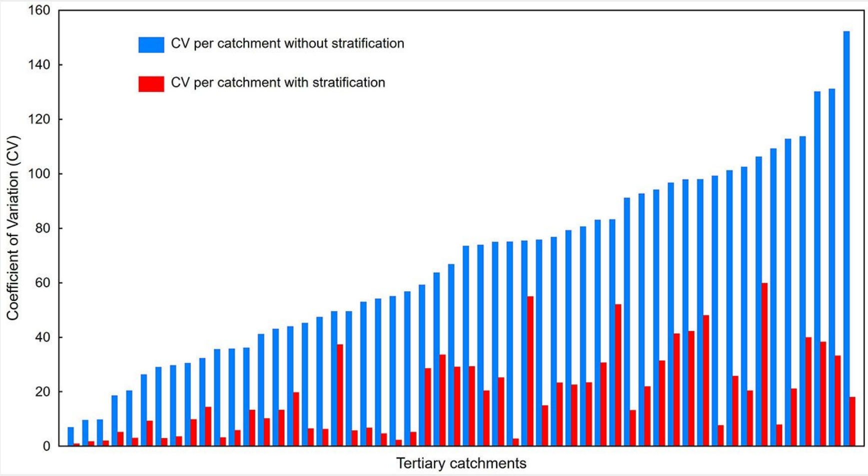


Fig. 4 - Comparison of IAP abundance variation, measured as coefficient of variation (CV) per tertiary catchment without stratification and thereafter with stratification (level of significance: $p<0.05$) for each of the three tree species associated with the respective tertiary catchments in which they occur.

further testing of the feasibility of the stratification was based on a stratification done at a tertiary catchment level by including four variables with three levels each and thereby 81 possible unique interaction combinations or strata per catchment. The variance in IAP abundance per stratified tertiary catchment was significantly lower than the variance in related tertiary catchments without any stratification across all tertiary catchments (Fig. 4), indicating that stratification had a substantial effect. The results of the analysis of variance applied to the same dataset showed significant differences in mean IAP abundance variation as summarized by the stratification for each of the three species (Tab. 3, Tab. 4).

Discussion

The results of this study are a method for stratified sampling as a base for a large scale inventory of invasive tree species or other invasive alien plants at a national level. The proposed coherent and objective method provides a means to use edaphic, climatic and geomorphologic variables to choose adequate strata in order to gain the necessary sampling efficiency for larger areas. It closes an obvious gap in IAP monitoring, where the current methods lack a statistical rigorous design-based approach and have mainly relied on either opportunistic recording of IAPs along accessible pathways such as roads (Henderson & Wilson 2017) or have been used to get presence/absence information based on expert knowledge, literature and herbarium records (Vinogradova et al. 2018).

Although species distribution modelling is a standard tool to predict potential IAP distribution (Robinson et al. 2017), the objective of this study was not to map potential

Tab. 3 - ANOVA summary including main effects and the intersection of the predictor environmental variables up to the 1st order. IAP species abundance served as the response variable (level of significance applied was $p<0.05$). (df): degrees of freedom.

Variables	Sum of squares	df	Mean square	F-value	P
Intercept	10211609823	1	10211609823	11883.68	<0.001
Rainfall	776469154	2	388234577	451.80	<0.001
Soil Depth	1680298283	2	840149141	977.72	<0.001
Clay B-hor	155313015	2	77656508	90.37	<0.001
Terrain morphology	1602866749	2	801433375	932.66	<0.001
Rainfall × Soil depth	1538399235	4	384599809	447.57	<0.001
Rainfall × Clay B-hor	232806813	4	58201703	67.73	<0.001
Rainfall × Terrain morphology	3132974317	4	783243579	911.49	<0.001
Soil depth × Clay B-hor	444560606	4	111140151	129.34	<0.001
Soil depth × Terrain morphology	259746046	4	64936512	75.57	<0.001
Clay B-hor × Terrain morphology	1227720920	4	306930230	357.19	<0.001
Error	91180877400	106111	859297	-	-

Tab. 4 - ANOVA table with all possible levels of intersection up to the 3rd order. IAP species abundance served as response variable and the environmental variables as predictor variables (level of significance applied was $p<0.05$). (df): degrees of freedom.

Variables	Sum of squares	df	Mean square	F-value	P
Intercept	10420581217	1	10420581217	12306.99	<0.001
Rainfall × Soil depth	570818379	4	142704595	168.54	<0.001
Rainfall × Clay B-hor	526171140	4	131542785	155.36	<0.001
Soil depth × Clay B-hor	458571226	4	114642806	135.40	<0.001
Rainfall × Terrain morphology	2993017343	4	748254336	883.71	<0.001
Soil depth × Terrain morphology	808312529	4	202078132	238.66	<0.001
Clay B-hor × Terrain morphology	959066426	4	239766606	283.17	<0.001
Rainfall × Soil depth × Clay B-hor	1096535396	8	137066925	161.88	<0.001
Rainfall × Soil depth × Terrain morphology	949042946	8	118630368	140.11	<0.001
Rainfall × Clay B-hor × Terrain morphology	884723287	8	110590411	130.61	<0.001
Soil depth × Clay B-hor × Terrain morphology	504441130	8	63055141	74.47	<0.001
Rainfall × Soil depth × Clay B-hor × Terrain morphology	2446333516	16	152895845	180.57	<0.001
Error	89812467726	106071	846720	-	-

IAP distribution, but rather to use modelling to support the development of a stratification that could be used in a sampling design (Särndal 2010) in order to quantify IAP abundance based on a representative grid of empirical sampling points. Similar approaches have been used to guide surveys for example where field surveys are limited due to a lack of resources (Fois et al. 2018). In these cases post-model field surveys were targeted on where a high probability of occurrence was predicted but without pre-model field data (Peterman et al. 2013). In other studies, this approach has been used to improve the assessment and verification of the distribution of scarce species and to optimise resources by focusing surveys on localities where a high probability of occurrence of such rare species was predicted (Peterman et al. 2013). This study, based on three invasive tree species of major ecological relevance, serves as the first step in the establishment of a scientifically-based regional level IAP monitoring programme for South Africa. Such a monitoring programme requires that actual IAP distribution and abundance data is sampled in the field and the resulting data should be used to iteratively refine and optimize future national level surveys (Volis 2016, Fois et al. 2018).

The results of this study revealed distinct species-specific differences in the occurrence patterns of the three invasive tree species under consideration, which may point to their ecological differences and optimum habitats. These three IAP tree species were introduced to South Africa with specific objectives and thereby established on a non-random basis. *Acacia mearnsii* was planted on a wide scale by the commercial forestry sector for its high tannin content in the bark. *Acacia cyclops* was used to stabilize drift sands along the coast and *Prosopis glandulosa* was planted extensively in the arid regions for animal fodder. Although all these species have had a residence time well in access of a 100 years in South Africa, it is possible that they have not reached their full geographic extent and their distribution is therefore not yet in equilibrium, which could cause problems for correlative models as pointed out by Robinson et al. 2017. This is an unknown and was mitigated for by selecting those species with the largest possible geographical extent.

Although a wide range of variables are available for such correlative investigations between species and the environment in which they occur, the emphasis of this study was on physical parameters, for instance soil depth and clay content. Therefore, chemical attributes such as soil pH that could significantly affect the distribution of IAP species (Soti et al. 2015) were not directly investigated. However, it can be reasoned that for instance soil clay content serves as a surrogate or indicator for soil pH. Sandy soils are usually more prone to acidic conditions due to leaching, whilst soils with a high clay content are typically more alkaline due to the combination of basic cations absorbed by clay particles and a lack of leaching (Cronin 2018). The size and geographic location of the aggregation area to be stratified plays an important role in the realisation of classes. The smallest aggregation unit showed that the intersection up to a maximum of four variables with three levels each does not deviate substantially from the maximum number of achievable combinations, hence the number of variables for stratification could be limited. Designs with six and more variables became impractical and this was also confirmed by other studies (Keppel 1982). It proved to be more effective to rather use fewer environmental variables within a stratification because this created the best opportunity to realise all possible classes, so four variables were finally selected. This selection was based on ecological reasoning and repetitive high associations between species and variables. Terrain morphological units were highly correlated with *A. cyclops* and *A. mearnsii. A. cyclops* is preferential to lower lying areas, especially coastal flats and seldom occurs within higher lying more mountainous areas. *A. mearnsii* is preferential to valley bottoms and foot slopes rather than higher lying landscape positions. Many pine species on the other hand are prolific invaders of mid-slopes and higher lying areas which further supports a distribution gradient between IAP tree species and the terrain morphological units. Tree species need soil of sufficient depth to establish and anchor their root systems to harvest nutrients and water which supports the high correlation with rainfall and soil depth. Clay content in the B-horizon was associated with all three species compared to clay content in the A-horizon, which was associated with only two species. *P. glandulosa* was preferential to clay in the B-horizon, that correlates with the typical high abundance of this species on alluvial soils in arid regions. The survival of trees in low rainfall areas is dependent on the water storage capacity of soils which is determined by clay content. Clay content in the B-horizon is the main water storage layer of soil, and some correlation between the occurrence of perennials and soils with a higher water storage capacity is expected in a predominantly low rainfall region such as South Africa, resulting in the B-horizon being more significant in the survival of evergreen trees. Collinearity between these four chosen variables was minimal. The analysis of variance (Tab. 3, Tab. 4) applied to the abundance of the respective species as response variable serves as confirmation of the viability of the selected environmental variables and their respective levels to be applied in a stratification and the resulting categories to reduce variation and distinguish between IAP abundance levels (Fig. 4).

Data availability and detail differs at national, continental and global levels. Some data sets, such as digital surface models and climate data that were also used in this study, are easily accessible at continental and global scales. Detailed data sets such as soil information are not available at continental and global scales because data acquisition standards differ largely between countries. Since decision making is typically conducted at the political entity of the national or regional scale, the use of national and regional data is an advantage since the level of inventory matches the level of decision making and thus makes use of the best available data set. By identifying significantly contributing factors from those national data sources, a need for a standardised assessment of those variables is highlighted to improve the inventory of invasive tree species also at larger scales, where invasive species have spread beyond national borders. Our methodology was set up specifically for South Africa, however, with small modifications it can be transferred to other countries.

Most studies carried out on IAPs in other countries rely on listing species, describing their occurrence in a geographic context and sometimes correlating species occurrence with further variables that can be derived from remote sensing sources or from ground borne data (Xu et al. 2012, Vinogradova et al. 2018). Concise large scale studies with a sound statistical sampling that enables an assessment not only of IAP occurrence but also an estimation of abundance remain the exception in applied IAP inventories. Only a few studies undertake it to establish a statistically sound and efficient sampling system as a base of a representative inventory of invasive trees. An example is the national forest inventory in the USA (Smith 2002). However, non-forested areas which are the vast majority in water limited countries such as South Africa, and might still host invasive trees, were not part of the inventory. Statistically derived habitat suitability models (HSMs) and species distribution models (SDMs) have been previously successfully applied to develop sampling designs that enable an efficient sampling of IAPs of large areas. Examples are Lemke & Brown (2012) and Wang et al. (2014). Our approach is similar in some ways since it is making use of environmental variables that correlate with the occurrence and abundance of IAPs but is also distinct in other ways since it focused on an optimisation of selecting the optimum spatial resolution for the stratification and not only selecting the best set of different influence variables. This provided a sound base for choosing the best IAP sampling design for South Africa.

Conclusion

The study resulted in a stratified sampling procedure as a base for an invasive tree inventory at a national scale. Through detecting and minimising the full range of environmental variability within the defined population by means of grouping a continuous varying landscape into discrete classes or strata of similar variability, the sample variance was significantly reduced and sampling efficiency was increased to a level where large scale inventories are viable. The objective of this study was to determine which environmental variables most effectively summarize invasive tree abundance variability as well as to determine the number of strata to be included in such a stratification. These variables are to be applied in a future national level stratification by demarcating habitat types contributing the most to IAP occurrence in South Africa. This will ensure that all different habitat types are sufficiently included in a national level survey, as well as an optimized sample point allocation. It was shown that ideally not more than 81 unique strata should be created to obtain a stratification that does not deviate significantly from a statistically desirable full rank design. The number of variables included is obviously related to their levels and in this case it was shown that four variables at three different levels each can be used. Selected variables were identified based on a combination of correlation with species, replication across species as well as geographic space, and finally explained by means of biological reasoning. These variables included average rainfall, soil depth, clay content in the B-horizon and a form of landscape position such as terrain morphological units.

Acknowledgements

The National Invasive Alien Plant Survey (NIAPS) project is funded by the Working for Water Programme that resides with the Department of Environmental Affairs' Natural Resource Management Programme. This work originated from the NIAPS project and the authors acknowledge the financial contribution made by the Working for Water Programme. The last author acknowledges the contribution made by the "Care4C" project, grant no. 778322, in the EU Horizon 2020 Marie Sklodowska-Curie program.

References

Aitor GJ, Garcia-Viñas JI (2011). Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models. Ecological Modelling 222 (13): 2037-2041. - doi: 10.1016/j.ecolmodel.2011.04.015

Anderson RP, Lew D, Peterson AT (2003). Evaluating models of species' geographic distributions: criteria for selecting optimal models. Ecological Modeling 162: 211-232. - doi: 10.1016/S0304-3800(02)00349-6

Cronin D (2018). Handbook of soil fertility. Callisto Reference, New York, USA, pp. 240.

De Lange WJ, Van Wilgen BW (2010). An economic assessment of the contribution of biological control to the management of invasive alien plants and to the protection of ecosystem services in South Africa. Biological Invasions 12 (12): 4113-4124. - doi: 10.1007/s10530-010-9811-y

Dehnen-Schmutz K, Boivin T, Essl F, Groom QJ, Harrison L, Touza JM, Bayliss H (2018). Alien futures: what is on the horizon for biological invasions? Diversity and Distributions 24: 1149-1157. - doi: 10.1111/ddi.12755

Dormann CF (2011). Modelling species' distributions. In: "Modelling Complex Ecological Dynamics: An Introduction Into Ecological Modelling for Students, Teachers and Scientists" (Jopp F, Reuter H, Breckling B eds). Springer, Berlin, Germany, pp. 3-22. - doi: 10.1007/978-3-642-05029-9_13

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, García Marquéz JR, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, Mcclean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 35: 1-20. - doi: 10.1111/j.1600-0587.2012.07348.x

Driver A, Sink KJ, Nel JL, Holness S, Van Niekerk L, Daniels F, Jonas Z, Majiedt PA, Harris L, Maze K (2012). National Biodiversity Assessment 2011: an assessment of South Africa's biodiversity and ecosystems. Synthesis Report. South African National Biodiversity Institute and Department of Environmental Affairs, Pretoria, South Africa, pp. 24. [online] URL: http://opus.sanbi.org/handle/20.500.12143/786

Elith J, Franklin J (2017). Species distribution modeling. In: "Reference Module in Life Sciences". Elsevier, Amsterdam, Netherlands, pp. 15.

ESRI (2017). ArcGIS Desktop Release 10.6. Environmental Systems Research Institute, Redlands, CA, USA.

Fahrmeir L, Kneib T, Lang S, Marx B (2013). Regression: models, methods and applications. Springer-Verlag, Heidelberg, Germany, pp. 698. [online] URL: http://books.google.com/books?id=EQxU9iJtipAC

Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J, Werner M, Oskin M, Burbank D, Alsdorf D (2007). The Shuttle Radar Topography mission. Reviews of Geophysics 45: 1-33. - doi: 10.1029/2005RG000183

Fick SE, Hijmans RJ (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37: 4302-4315. - doi: 10.1002/joc.5086

Fois M, Cuena-Lombraña A, Fenu G, Bacchetta G (2018). Using species distribution models at local scale to guide the search of poorly known species: review, methodological issues and future directions. Ecological Modelling 385: 124-132. - doi: 10.1016/j.ecolmodel.2018.07.018

Franklin J (2010). Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK, pp. 340. - doi: 10.1017/CBO9780511810602

Gitzen RA, Millspaugh JJ, Cooper AB, Licht DS (2012). Design and analysis of long-term ecological monitoring studies. Cambridge, Cambridge University Press, UK, pp. 560. [online] URL: http://books.google.com/books?id=5Np9rU_0pPEC

Hageer Y, Esperón-Rodríguez M, Baumgartner JB, Beaumont LJ (2017). Climate, soil or both? Which variables are better predictors of the distributions of Australian shrub species? PeerJ 5: e3446. - doi: 10.7717/peerj.3446

Henderson L, Wilson JRU (2017). Changes in the composition and distribution of alien plants in South Africa: an update from the Southern African Plant Invaders Atlas. Bothalia 47 (2): 1-26. - doi: 10.4102/abc.v47i2.2172

Jeschke JM, Bacher S, Blackburn TM, Dick JTA, Essl F, Evans T, Gaertner M, Hulme PE, Kühn I, Mrugala A, Pergl J, Pyšek P, Rabitsch W, Ricciardi A, Richardson DM, Sendek A, Vilà M, Winter M, Kumschick S (2014). Defining the impact of non-native species: resolving disparity through greater clarity. Conservation Biology 28: 1188-1194. - doi: 10.1111/cobi.12299

Kalkhan MA (2011). Spatial statistics: geospatial information modelling and thematic mapping. CRC Press, Boca Raton, FL, USA, pp. 184. - doi: 10.1201/9781439891117

Keppel G (1982). Design and analysis: an experimenter's handbook (4th edn). Pearson, CA, USA, pp. 611.

Kumschick S, Gaertner M, Vilà M, Essl F, Jeschke JM, Pyšek P, Ricciardi A, Bacher S, Blackburn TM, Dick JTA, Evans T, Hulme PE, Kühn I, Mrugala A, Pergl J, Rabitsch W, Richardson DM, Sendek A, Winter M (2015). Ecological impacts of alien species: quantification, scope, caveats and recommendations. BioScience 65: 55-63. - doi: 10.1093/biosci/biu193

Land Type Survey Staff (2006). Land types of South Africa: digital map (1:250 000 scale) and soil inventory databases. ARC - Institute for Soil, Climate and Water, Pretoria, South Africa.

Le Maitre DC, Forsyth GG, Dzikiti S, Gush MB (2016). Estimates of the impacts of invasive alien plants on water flows in South Africa. Water SA 42: 659-672. - doi: 10.4314/wsa.v42i4.17

Lemke D, Brown JA (2012). Habitat modeling of alien plant species at varying levels of occupancy. Forests 3: 799-817. - doi: 10.3390/f3030799

MathWorks (2017). MATLAB and Statistics toolbox release 2017a. The MathWorks, Inc., Natick, MA, USA.

Mittermeier RA, Turner WR, Larsen FW, Brooks TM, Gascon C (2011). Global biodiversity conservation: the critical role of hotspots. In: "Biodiversity Hotspots - Distribution and Protection of Conservation Priority Areas" (Zachos FE, Habel JC eds). Springer, Berlin, Germany, pp. 3-22. - doi: 10.1007/978-3-642-20992-5_1

Naesset E, Gobakken T, Solberg S, Gregoire TG, Nelson R, Ståhl G, Weydahl D (2011). Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: a case study from a boreal forest area. Remote Sensing of Environment 115: 3599-3614. - doi: 10.1016/j.rse.2011.08.021

Nel JL, Richardson DM, Rouget M, Mgidi T, Mdzeke N, Le Maitre DC, Van Wilgen BW, Schonegevel L, Henderson L, Neser S (2004). A proposed classification of invasive plant species in South Africa: towards prioritizing species and areas for management action. South African Journal of Science 100: 53-64. [online] URL: http://journals.co.za/content/sajsci/100/1-2/EJC96213

Peterman WE, Crawford JA, Kuhns AR (2013). Using species distribution and occupancy mod-

eling to guide survey efforts and assess species status. Journal for Nature Conservation 21: 114-121. - doi: 10.1016/j.jnc.2012.11.005
Ricciardi A, Blackburn TM, Carlton JT, Dick JTA, Hulme PE, Iacarella JC, Jeschke JM, Liebhold AM, Lockwood JL, MacIsaac HJ, Pyšek P, Richardson DM, Ruiz GM, Simberloff D, Sutherland WJ, Wardle DA, Aldridge DC (2017). Invasion science: a horizon scan of emerging challenges and opportunities. Trends in Ecology and Evolution 32: 464-474. - doi: 10.1016/j.tree.2017.03.007
Robinson AP, Walshe T, Burgman MA, Nunn M (2017). Invasive species: risk assessment and management. Cambridge University Press, Cambridge, UK, pp. 426. [online] URL: http://books.google.com/books?id=LYThDgAAQBAJ
Rouget M, Hui C, Renteria J, Richardson DM, Wilson JRU (2015). Plant invasions as a biogeographical assay: vegetation biomes constrain the distribution of invasive alien species assemblages. South African Journal of Botany 101: 24-31. - doi: 10.1016/j.sajb.2015.04.009
Sahragard HP, Ajorloa M (2018). Comparison of logistic regression and maximum entropy for distribution modeling of range plant species (a case study in rangelands of western Taftan, southeastern Iran). Turkish Journal of Botany 42: 28-37. [online] URL: http://journals.tubitak.gov.tr/botany/abstract.htm?id=21843
Särndal C (2010). Models in survey sampling. Statistics in Transition 11 (3): 539-554. [online] URL: http://www.infona.pl/resource/bwmeta1.element.desklight-2413e6e6-1fd4-43ea-b2cd-1c3d20e94ba2
Simberloff D, Genovesi P, Pyšek P, Campbell K (2011). Recognizing conservation success. Science 332: 419. - doi: 10.1126/science.332.6028.419-a
Smith WB (2002). Forest inventory and analysis: a national inventory and monitoring program. Environmental Pollution 116: 233-242. - doi: 10.1016/S0269-7491(01)00255-X
Soti P, Jayachandran K, Koptur S, Volin JC (2015). Effect of soil pH on growth, nutrient uptake, and mycorrhizal colonization in exotic invasive Lygodium microphyllum. Plant Ecology 216 (7): 989-998. - doi: 10.1007/s11258-015-0484-6
Ståhl G, Saarela S, Schnell S, Holm S, Breidenbach J, Healey SP, Patterson PL, Magnussen S, McRoberts RE, Gregoire TG (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems 3: 5. - doi: 10.1186/s40663-016-0064-9
Stockwell DRB (1999). Genetic algorithms II. In: “Machine learning methods for ecological applications” (Fielding AH ed). Kluwer, Dordrecht, Netherlands, pp. 123-144. - doi: 10.1007/978-1-4615-5289-5_5
Symonds MRE, Moussalli A (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike’s information criterion. Behavioral Ecology and Sociobiology 65: 13-21. - doi: 10.1007/s00265-010-1037-6
Van Wilgen BW, Forsyth GG, Le Maitre DC, Wannenburgh A, Kotze JDF, Van Den Berg E, Henderson L (2012). An assessment of the effectiveness of a large, national-scale invasive alien plant control strategy in South Africa. Biological Conservation 148: 28-38. - doi: 10.1016/j.biocon.2011.12.035
Vilà M, Hulme PE (2017). Impact of biological invasions on ecosystem services. Springer-Verlag, Heidelberg, Germany, pp. 354. [online] URL: http://link.springer.com/book/10.1007/978-3-319-45121-3
Vinogradova Y, Pergl J, Essl F, Hejda M, Van Kleunen M, Pyšek P (2018). Invasive alien plants of Russia: insights from regional inventories. Biological Invasions 20: 1931-1943. - doi: 10.1007/s10530-018-1686-3
Volis S (2016). Species-targeted plant conservation: time for conceptual integration. Israel Journal of Plant Sciences 63: 232-249. - doi: 10.1080/07929978.2015.1085203
Wang O, Zachmann LJ, Sesnie SE, Olsson AD, Dickson BG (2014). An iterative and targeted sampling design informed by habitat suitability models for detecting focal plant species over extensive areas. PLoS One 9 (7): e101196. - doi: 10.1371/journal.pone.0101196
Webster R, Lark RM (2013). Field sampling for environmental science and management. Routledge, London, UK, pp. 192.
Williams KJ, Belbin L, Austin MP, Stein JL, Ferrier S (2012). Which environmental variables should I use in my biodiversity model? International Journal of Geographical Information Science 26: 2009-2047. - doi: 10.1080/13658816.2012.698015
Wilson JRU, Faulkner KT, Rahlao SJ, Richardson DM, Zengeya TA, Van Wilgen BW (2018). Indicators for monitoring biological invasions at a national level. Journal of Applied Ecology 55: 2612-2620. - doi: 10.1111/1365-2664.13251
Xu H, Qiang S, Genovesi P, Ding H, Wu J, Meng L, Han Z, Miao J, Hu B, Guo J, Sun H, Huang C, Lei J, Le Z, Zhang X, He S, Wu Y, Zheng Z, Chen L, Jarošík V, Pyšek P (2012). An inventory of invasive alien species in China. NeoBiota 15: 1-26. - doi: 10.3897/neobiota.15.3575

Supplementary Material

Appendix 1 - Description on the approach used to filter IAP species within the SAPIA database.

Appendix 2 - The stratification procedures followed of environmental variables.

Fig. S1 - Approach followed to filter the SAPIA database.

Link: Kotze_2767@suppl001.pdf