DE GRUYTER
OPEN

Selecting site characteristics at different spatial and thematic scales for shrubby cinquefoil (*Potentilla fruticosa* L.) distribution mapping

Kalle Remm

Remm, K. 2016. Selecting site characteristics at different spatial and thematic scales for shrubby cinquefoil (*Potentilla fruticosa* L.) distribution mapping. – Forestry Studies | Metsanduslikud Uurimused 64, 17–38. ISSN 1406-9954. Journal homepage: http://mi.emu.ee/forestry.studies

Abstract. The largest natural population of shrubby cinquefoil (*Potentilla fruticosa*) in the Baltic States was observed in the field to reveal the scale-dependent explanatory value of site characteristics for subsequent spatial distribution modelling of the species. About 700 km was crossed during field observations in 2008–2014. Thinning of the raw field records to ensure a distance of at least 50 metres between each point yielded 1459 presences and 7327 absences. These occurrence data were related to present and historical land cover, soil, elevation, human population density, the proportion of presence sites, and *P. fruticosa* mean coverage in the neighbourhood. Boosted classification tree models were used to compare the value of 60 individual site features at thematically and spatially different levels of generalization as indicators of the species' presence or absence. *P. fruticosa* presence is significantly non-random regarding most of the studied site features but only a few of these are valuable predictors. The proportion of presences in the neighbourhood had the highest indicative value. *P. fruticosa* occurrence also coincides with moist thin calcareous soils according to the soil map, with larger scrubland patches according to the topographical database, and with tussock areas according to a topographical map from the 1930s. The explanatory value of nominal site characteristics primarily drops when the most indicative category is merged with other classes to form a more general category. Site characteristics calculated at the observation point are not always the most effective predictors for *P. fruticosa* occurrence – features of the neighbourhood are related to the occurrence as well. The study area was classified into: confirmed absence area, unclear presence/absence area and probable presence area. Subsequent distribution modelling in the unclear area should be targeted on a species presence/absence, while abundance could be the priority within the probable presence area.

Keywords: *Potentilla fruticosa*, site features, thematic and spatial scale, thinning, proportion of presences.

Author' address: Institute of Ecology and Earth Sciences, University of Tartu, 46 Vanemuise St, 51014 Tartu, Estonia; e-mail: kalle.remm@ut.ee

Introduction

The general approaches for species and habitat distribution mapping can be outlined as 1) direct field observations, 2) remote detection by remote sensing and telemetry techniques, or 3) indirect estimation by experts, using indicators from statistical models or from similarity-based reasoning. Direct observation by an expert can be reliable, but covering any larger area by direct observation is labour-intensive. Therefore, indirect estimations are inevitably required to create a detailed distribution map for a larger area. Site characteristics related to the occurrence of the target

DOI: 10.1515/fsmu-2016-0002

species and its potential habitat, are used to estimate the likely absence, presence or abundance of the species.

For the application of numerical methods, detailed formal elementary features (e.g. site elevation above sea level or land cover/land use type according to a given classification) are extracted from data layers representing environmental conditions in the study area. Among these, neighbourhood effects, i.e. features representing the site neighbourhood (e.g. proportion of suitable habitat or the existence of earlier presence records of the species within a certain radius), have confirmed to be equally useful as the focal feature values (Mack & Harper, 1977; Guisan & Thuiller, 2005; Latimer *et al*., 2006).

The selection of the appropriate spatial and thematic resolution is one of the central issues in land cover and habitat mapping using remote sensing data (Ju *et al*., 2005) and also in landscape ecology (Turner *et al*., 1989; Wiens, 1989) and in distribution mapping. The predictive ability of spatial models depends on scale, as species expectedly have characteristic scales of response to their environment (Thuiller *et al*., 2003; Holland *et al*., 2004; Graf *et al*., 2005; Boscolo & Metzger, 2009). Vale *et al*. (2014) found that fine resolution models are more accurate at selecting marginal habitats, and distribution models with coarse resolution tended to overestimate species distribution at the edge of distribution area. However, the limited availability of high resolution data precludes its frequent use. As a rule, abiotic variables (e.g. climate) tend to be more important in continental or global models, while variables representing habitat and biotic interactions can be more important at finer spatial scales (Boulangeat *et al*., 2012).

Scale has spatial and thematic aspects. Spatial scale is described by grain size and spatial extent (O'Neill *et al*., 1986; Wiens, 1989). Grain is the resolution or minimum mapping unit of the data, while extent is the size of a mapped area. In addition, local extent has been defined as the radius or area of a kernel around each focal point (Thompson & McGarigal, 2002), whereas distance-dependent neighbourhood effects are related to scale by the neighbourhood extent. Thematic resolution is often limited by the available datasets. In the case of a nominal variable, the thematic resolution refers to the level of detail in the categories (categorical scale); in the case of a continuous variable, the thematic scale is expressed by measurement precision. The reliability of spatial predictions affected by thematic resolution has been paid much less attention than the effect of the spatial scale (Liang *et al*., 2013; Zhou *et al*., 2014).

By combining different data layers: their spatial scale and thematic generalization, kernels covering different extent from the surrounding area, and statistics calculated locally from these kernels, the number of possible numerical features for any geographical location approaches infinity. However, it is not reasonable to include excessive number of features to prediction models because since irrelevant explanatory variables add noise to the predictions and increase the risk of model over-fitting (Remm, 2004; Dormann, 2011; Ficetola *et al*., 2014), and also because the pre-processing of each characteristic to a numerical format is a time-consuming task. It is not always easy to identify a priori how many (and which) site features are actually relevant and should be used in predictive models.

In this experiment, comprehensive field observation data of shrubby cinquefoil *Potentilla fruticosa* were used to find out which site characteristics in which spatial and thematic scale are the best predictors of the species' presence and absence in the study region. These spatial features should be involved to distribution models in subsequent studies as explanatory factors, applied for creating detailed full-cover predictive distribution maps, and considered when planning protection measures for the species.

Material and Methods

Study area

The study area covering 819 km² is located in north-western Estonia and is bounded by the Baltic Sea to the north. The capital of Estonia, Tallinn, lies on its eastern boundary (Figure 1).

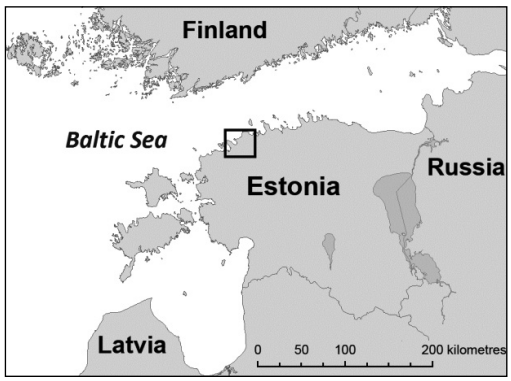


Figure 1. Location of the study area (black rectangle).

Joonis 1. Uurimisala (must ristkülik) paiknemine.

The elevation above sea level (asl) of natural ground in the study area is up to 60 m. The most frequent soil types according to the 1: 10 000 soil map are Sapric Histosol (96 km²), Mollic Gleysol (94 km²), Gleysol (78 km²), Calcaric Regosol (75 km²), Molli-Histic Gleysol (43 km²) and Endogleyic Cambisol (43 km²). According to the Estonian National Topographic Database (ENTD), forest covers 369 km², cultivated land 176 km², natural grassland 103 km², private yards 44 km², unmanaged open land (in this region mainly alvar grassland) 34 km², scrubland 13 km² and inland waters 8 km².

The study area contains the Vääna Landscape Reserve (4.09 km²), created to protect, *inter alia*, the largest natural population of *P. fruticosa* in the Baltic States. Similar alvar sites can be found elsewhere in western and northern Estonia, but the species does not grow there.

Shrubby cinquefoil habitat demands

Shrubby cinquefoil (*Potentilla fruticosa* L. syn. *Dasiphora fruticosa* (L.) Rydb.) (Rosaceae) is a perennial flowering shrub mainly known as a decorative cultivar. Its natural populations are widespread in Asian mountains and in North America. Its distribution in Europe is sporadic (Elkington & Woodell, 1963). Shrubby cinquefoil is a protected plant in Estonia, where the only sustainable population is between Tallinn, Keila and Paldiski. Here, mainly on alvar grasslands situated on Middle and Upper Ordovician limestone, grows the largest natural population in the Baltic States.

P. fruticosa prefers open sites, although resists moderately dense scrub and can survive for decades under a young forest canopy. However, shade is considered the main limiting factor (Gorchakovskiy, 1960; Elkington & Woodell, 1963). Moderate grazing seems to be beneficial for the species, suppressing potential competitors – lush herbs and bushes. The species' attitude to soil characteristics is not clear. According to current knowledge, at least northern European populations grow mainly in moist base-rich soil, although the plant is tolerant of slightly acid soils, drought and temporal flooding (Gorchakovskiy, 1960; Elkington & Woodell, 1963; Roland & Smith, 1969; Reier & Leht, 1999; Lonati *et al*., 2014).

Data and methods

The methodological framework of this study consists of the following stages. The relations between data pre-processing, data processing stages and intermediate results are given in the data flow diagram (Figure 2):

1. Field observation of the *P. fruticosa* coverage at a possibly large number of locations.
2. Evening out the observation density by thinning out closely neighbouring locations.
3. Calculating site features for the retained locations.

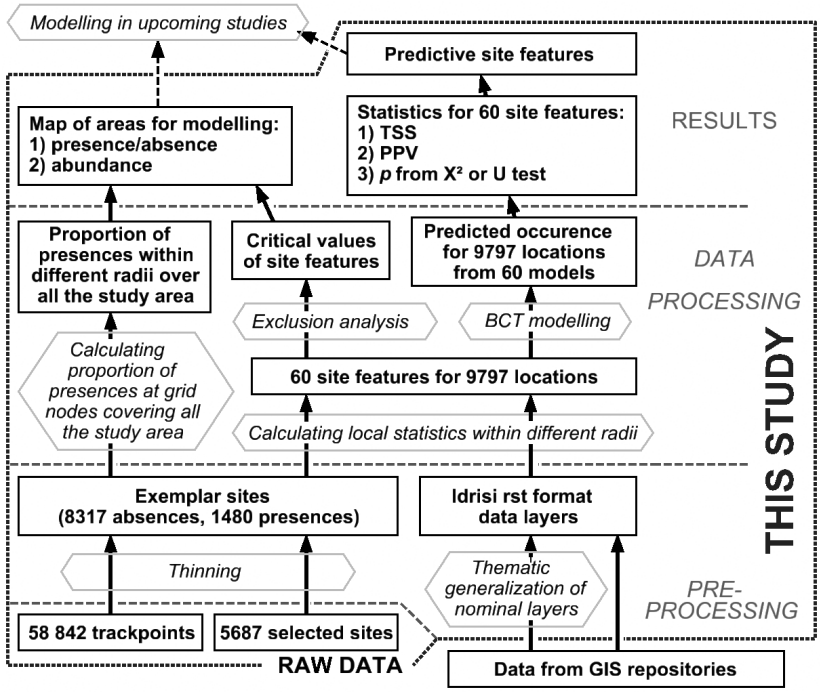
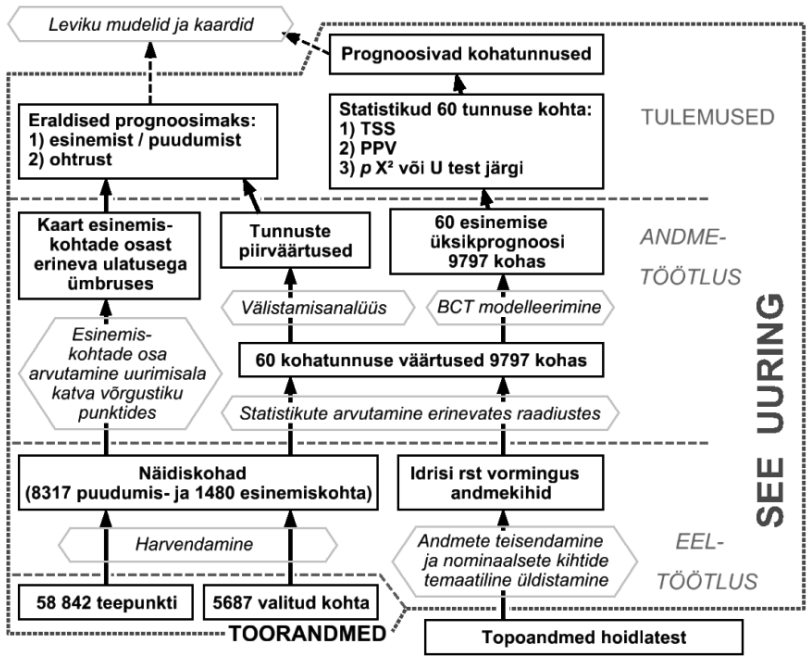


Figure 2. Connections between calculated items (rectangles) and methods used in this investigation. See text for explanation.



Joonis 2. Seosed arvutustulemuste (ristkülikud) ja toimingute vahel.

4. Relating single site features to the species presence/absence to find the best predictors.
5. Estimating intervals of feature values that exclude the species occurrence for certain.
6. Estimating intervals of feature values that, in most cases, indicate the species occurrence.
7. The study area was classified into three categories according to the critical values of the most indicative site features: 1) area where site features clearly exclude the species presence; 2) area, where the species occurrence is probable; 3) area where the species occurrence is unclear.

Field observations

Field observations were made when *P. fruticosa* was in bloom in the summers of 2008–2014 by conducting walking tours across the terrain. The sites for field observations were dynamically planned according to the accumulating experience on *P. fruticosa* distribution and its habitat preferences. The survey was focused on sites that should be relatively suitable for the species, but where it had not been recorded. Locations characterized by site features excluding the species (e.g. currently cultivated fields and the south-west part of the study area) were paid less attention. Dynamic planning of observation tracks enabled to increase the species detection to the sampling effort ratio and to prevent collecting superfluous amount of absence data, which is inevitable in case of regular and random sampling.

The coordinates of *P. fruticosa* presence locations, the movement track, and selected typical absence locations were recorded using a Garmin Vista HC + GPS receiver. The total length of the observation tracks during 82 observation days was 700 km. The coordinates of a presence location were measured when standing on a shrubby cinquefoil bush; the species was not visible within about 20 m in actively selected typical absence locations. The relative area covered by the species at each presence site was recorded but not used in this study as a dependent variable.

The density of observations was not equal throughout the study area, since the areal proportion of habitat types, ground visibility, terrain roughness, restricted areas and probability of species occurrence were also considered when planning field tours and moving across terrain. Private and industrial yards (43 km^2), a gunnery-practice ground (7 km^2), and the territory of a military air field (9 km^2) were the largest restricted parts of the study area. Private lands were generally not an obstacle for the field survey, as it is permitted to access unrestricted private property in Estonia, unless the owner forbids it. Field movement was not restricted to tracks, since following roads and paths causes unequal likelihood of locations to be observed (sampling bias) (Kadmon *et al*., 2004; Albert *et al*., 2010). Moreover – intentional dodges were frequent when moving along a vehicle track.

Thinning

The raw field data contained 58 842 track points automatically recorded by the GPS receiver and 5687 intentionally recorded observations (2854 presence and 2833 absence sites) in locations considered typical by the observer. A method for reducing the effect of uneven sampling, removing redundant points and evening the observation density is spatial thinning (Remm *et al*., 2009; Boria *et al*., 2014). The raw records were thinned using an online spatial data calculator (SDC) (Remm & Kelviste, 2014) to ensure at least 50 m distance between each accepted location in order to even out the density of observations, reduce the share of automatically recorded absences, and avoid spatially close records. Aiello-Lammens *et al*. (2015) published a thinning function spThin in R. Thinning in the SDC differs from it by: 1) enabling different source and target points, 2) enabling different types of input (including 1D, 2D and

3D), 3) by starting from the first point in a list, not from the location with the greatest number of neighbouring occurrences.

The thinning in the SDC was a step-by-step process implemented in the following order:

1. Removal of intentionally recorded absence sites less than 50 m from presence sites.
2. Thinning of retained absence sites to ensure intervals of at least 50 m.
3. Removal of track points less than 50 m from the retained absence sites.
4. Removal of track points less than 50 m from the presence sites.
5. Thinning of retained absence points to ensure intervals of at least 50 m.
6. Thinning of presence sites to ensure intervals of at least 50 m.

The retained intentionally observed and automatically recorded absence sites gave a total of 8317 absence exemplars, which, together with the 1480 retained presence sites (total = 9797), were used in the calculations (Figure 3). The species cover estimations in the thinned locations are freely available as an archived dataset (Remm, 2016). The dataset advantages are the recorded absences; all observations made during the same season, during a relatively short period of years and predominantly by one person; data thinning reduced pseudoreplication and disproportion between the number of recorded presence and absence locations.

The mean density of thinned records per terrestrial study area is 0.12 points per hectare (0.09 in forests and 0.08 on fields), while being higher in the typical *P. fruticosa* habitats: scrubland 0.44, natural grassland 0.27 and in open swampy ground 0.24 records per hectare.

Data layers

Areal categories from the Estonian National Topographic Database (ENTD) were used to describe present land cover and land use. The spatial resolution of this database corresponds to a 1: 10 000 map; data were from the year 2013. The areal categories used at different levels of thematic generalization are listed in the Supplement (Remm, 2015). The most detailed level includes 48 categories, the medium level, 15 and the most generalized level, 6 categories.

In addition to the present land cover, the main areal categories from a 1: 50 000 topographical map surveyed in the second half of the 1930s were included, since the present *P. fruticosa* distribution is presumably related to previous land use. This historical map is more generalized than the ENTD data; land cover and land use is flexibly depicted by different combinations of topographical symbols, which were classified into 18 categories at the most detailed level. These categories were digitized from scanned map sheets as vector polygons. The meaning of categories included from the historical topographical map and their number does not correspond to the classification used in the ENTD, since the scale and mapping principles of these data sources do not match. E.g. the historical map includes a special symbol for tussocks, used either as a separate areal category or in combination with grassland, marshland or shrub symbols. Unfortunately, tussock areas are not represented by any single ENTD land cover category nor by a combination of categories.

Soil data were obtained from the Estonian 1: 10 000 soil map and land elevation from digital elevation models derived from detailed LiDAR measured raw data. The soil types with their names according the original map and the closest World Reference Base for Soil Resources (WRB) taxonomic unit are listed in the Supplement (Remm, 2015), since categories used in the Estonian soil map are not directly transferable to the WRB soil system. The above-mentioned land cover, soil and elevation data were obtained from the Estonian Land Board, and human population data from Statistics Estonia (Ministry of Finance). The human population data layer

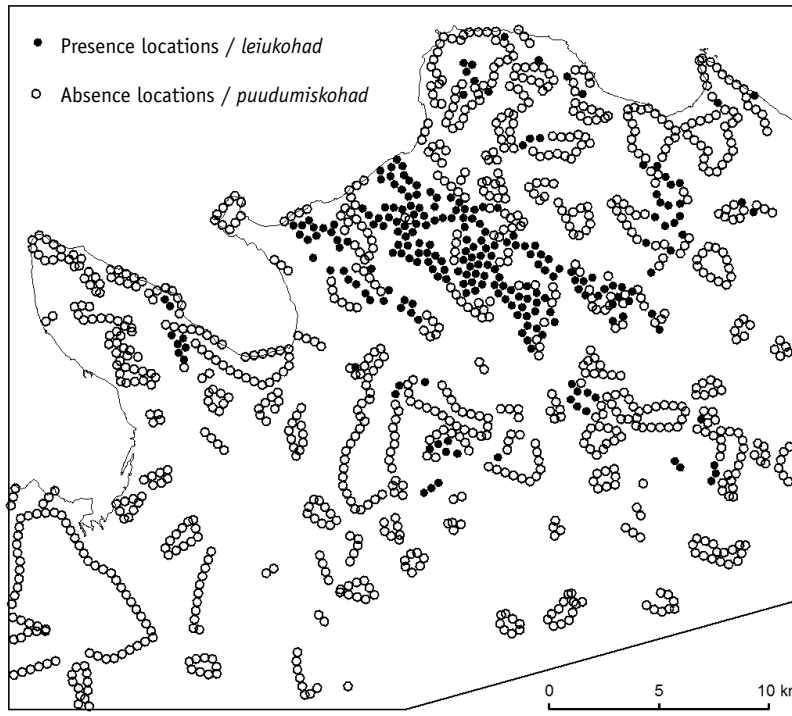


Figure 3. *P. fruticosa* generalized records from this investigation (presences filled, absences empty circles). One circle represents about 6 thinned locations.

Joonis 3. Põõsasmarana vaatluskohad üldistatud kujul. Täidetud ringide kohas on leiukohad, tühjade kohas puudumiskohad. Üks ring joonisel vastab keskmiselt kuuele harvendatud vaatluskohale.

had a 1000 m grid interval, the other layers were rasterized to 10 m grids.

Site features

The recorded sites were described using 60 site characteristics (features) calculated from data layers representing land cover, soil and terrain properties, human population density and also *P. fruticosa* observation results from the vicinity (Table 1). A site feature was defined by: 1) the data layer and its thematic detail (resolution), 2) spatial scale, as the radius limiting the inclusion extent of the values, 3) a statistic calculated from the values within the radius in the layer.

The values of the site features at the thinned locations were calculated using the SDC. Radii of 0 (local value), 100, 200, 500 and 1000 m were used to derive features corresponding to different spatial scales, except for the human density layer, which has an original grid interval of 1 km. The statistics calculated within a given radius were the mean of values in a numerical layer and the most frequent (modal) category in the case of a nominal data layer. For local features, the closest value was read from the data layer. Land elevation was included as the mean elevation and as the relative elevation compared to the mean in a given radius. Different thematic resolutions were represented by merging initial detailed categories into a smaller number of more generalized units, as given in Remm (2015).

The proportion of presences among records from the neighbourhood represents spatial continuity of the distribution (autocovariance). The proportion of presences

Table 1. Data layers and site features calculated at different thematic and spatial resolutions. PLC – present land cover, HLC – historical land cover, PP – proportion of presence locations among observed locations, HPD – human population density per km², LV – local value, RV – relative value (compared to the mean), N – number of categories, R – radius in metres.

Tabel 1. Andmekihid ja nendest arvutatud kohatunnused erinevas temaatilises ja ruumilises mõõtkavas. PLC – kaasaegne maakate, HLC – ajalooline maakate, soil – muld, elevation – maapinna kõrgus, PP – leiukohtade osa vaatluskohtadest, HPD – alaliste elanike arv ruutkilomeetril, LV – lokaalväärtus, RV – suhteline väärtus võrreldes keskmisega, N – kategooriate arv, mode – mood, mean – keskmine, R – raadius meetrites.

Layer *Kiht*	N	Statistic *Statistik*	R	Layer *Kiht*	N	Statistic *Statistik*	R	Layer *Kiht*	N	Statistic *Statistik*	R	Layer *Kiht*	Statistic *Statistik*	R
PLC	48	LV		HLC	18	LV		Soil	59	LV		Elevation	LV	
		Mode	100			Mode	100			Mode	100		Mean	100
			200				200				200			200
			500				500				500			500
			1000				1000				1000			1000
	15	LV			6	LV			25	LV			RV	100
		Mode	100			Mode	100			Mode	100			200
			200				200				200			500
			500				500				500			1000
			1000				1000				1000	PP		100
	6	LV			2	LV			10	LV				200
		Mode	100			Mode	100			Mode	100			500
			200				200				200			1000
			500				500				500	HPD	LV	
			1000				1000				1000		Mean	1000

per number of records applied within different radii, instead of the raw number of presences, was used to reduce the effect of variable survey intensity. The focal record was not included when calculating the proportion of presences in the neighbourhood of an observation site. The use of proportions instead of the number of presences minimizes the effect of different observation density

Estimating predictive values of features

The explanatory value of site features was compared by: 1) looking for feature values that directly indicate *P. fruticosa* presence or absence, 2) by the correctness of predictions using this single variable (Figure 2).

Feature values excluding species presence or excluding absence were tested for both categorical and numerical features. For categorical features, this is simple examination of presence and absence frequency in the categories. In the case of numerical features, value intervals including at least 1% of observations (100 cases) and containing no recorded presences were considered as excluding species presence. The opposite is an absence excluding interval, which should include ≥ 1% of cases and no absences. For finding the excluding intervals, a gradual search algorithm is available in the SDC (*Regions of frequency → Code*).

The algorithm works as follows:

1. Sort cases according to values.
2. Mark the first case in each group of equal values.
3. Check correspondence to pre-determined proportion and frequency crite-

ria while extending the value interval starting from a first case. Groups of equal values are always included together.

When looking for a simple but effective and universal model applicable for a bivariate dependent variable, where the frequency of categories (presence/absence) is unequal, using single numerical and categorical predictors, the boosted classification tree (BCT) method was preferred since it is highly resistant to over fitting (since subsampling is included to model calibration) and has confirmed to be among the most effective prediction methods (Elith *et al*., 2006; Elith & Graham, 2009; Zurell *et al*., 2009). The predicted presence or absence was calculated from a BCT model per each feature at each spatial and thematic scale. Interactions of site characteristics were not studied, as the aim of this investigation was to compare the explanatory value of features one by one.

The modelling results were compared in three aspects: 1) goodness-of-fit between observed and model-predicted presences and absences, 2) the proportion of true positive predictions among model-predicted presences, and 3) statistical significance of the difference in feature values between *P. fruticosa* presence and absence sites. The True Skill Statistic (TSS) – also called Hanssen-Kuipers Skill Score (Hanssen & Kuipers, 1965) – was used as the objective function to compare the predictive ability of the BCT models. The TSS value is calculated as the proportion of true positive cases plus the proportion of true negative cases minus one. The TSS statistic is preferable as it does not depend on the proportion of categories, and integrates correct prediction of both presences and absences (McPherson *et al*., 2004).

Another statistic for the comparison of site features was indicator precision or positive predictive value (PPV), which is the proportion of true positives among all positive predictions. The PPV enables both feature-level estimates and the comparison of single categories of a nominal feature. In the case of a single category, PPV is the proportion of the presence sites among observed sites characterized by this particular presence-indicating category.

The statistical significance of the predictors was estimated using the SDC, by comparing the frequency distribution of explanatory categories in presence sites and absence sites using the χ^2 test, and by comparing mean values of numerical variables using the Mann-Whitney U test.

Results

Predictive values of site features

Although all the 60 explanatory variables were statistically significantly ($p < 0.001$) related to the occurrence of *P. fruticosa*, as the number of observations is large and species occurrence does not correspond randomly to different feature values, most of them were found to be weak predictors (Table 2). *P. fruticosa* presence is characterized first of all by a higher density of presences in the vicinity, especially within the nearest 100 m (TSS = 0.85, PPV = 0.70). The best single categories of land cover and soil, which predict the species presence sites correctly in most cases (PPV ≥ 0.5) if they are the modal category within the given radius, were: 1) land cover type scrubland (indicative in all radii, the highest PPV = 0.88 if radius = 200 m), 2) tussock surface according to the historical map (all radii, PPV = 0.85 if combined with grassland) and 3) relatively thin gleyic soil laying on limestone bedrock in the study area: (Gleyic Rendzic Leptosol *Gh*, PPV = 0.53 locally and Endogleyic Leptosol *Khg*, PPV = 0.52 if radius = 1000 m) (Table 3).

The area where at least one of these four predictors is favourable covers 38.4 km². However, a single favourable predictor may be occasional, so, for this species, it is preferable to delineate the favourable area by the presence of at least two favourable conditions. The area where at least two of

Table 2. Predictive ability of the BCT models according to the true skill statistic (TSS) and positive predictive value (PPV). Thematic resolution: D – detailed, M – medium, G – generalized; other abbreviations as in the Table 1. Values > 0.5 are in bold.

Tabel 2. BCT mudelite prognoosiv võime Hanssen-Kuipersi skoori (TSS) ja positiivse prognoosiväärtuse (PPV) järgi. Temaatiline üldistustase: D – üksikasjalik, M – keskmine, G – üldine; teised lühendid nagu tabelis 1. Väärtused > 0,5 on rasvases kirjas.

Layer *Kiht*	Detail *Üldistus*	Statistic *Statistik*	R	TSS	PPV	Layer *Kiht*	Detail *Üldistus*	Statistic *Statistik*	R	TSS	PPV
PLC	D	LV		0.396	0.297	Soil	D	LV		**0.582**	0.353
		Mode	100	0.359	0.309			Mode	100	**0.569**	0.347
			200	0.334	0.332				200	**0.551**	0.351
			500	0.263	0.358				500	**0.515**	0.338
			1000	0.164	0.221				1000	0.460	0.301
	M	LV		0.396	0.297		M	LV		**0.583**	0.367
		Mode	100	0.361	0.311			Mode	100	**0.570**	0.363
			200	0.361	0.311				200	**0.554**	0.358
			500	0.264	0.361				500	**0.504**	0.353
			1000	0.163	0.220				1000	0.438	0.331
	G	LV		0.154	0.202		G	LV		0.214	0.185
		Mode	100	0.186	0.202			Mode	100	0.192	0.181
			200	0.182	0.197				200	0.180	0.179
			500	0.135	0.183				500	0.144	0.172
			1000	0.090	0.164				1000	0.132	0.185
HLC	D	LV		0.441	0.291	Elevation		LV		0.419	0.263
		Mode	100	0.438	0.278			Mean	100	0.423	0.261
			200	0.416	0.268				200	0.400	0.249
			500	0.330	0.240				500	0.417	0.255
			1000	0.209	0.189				1000	0.427	0.265
	M	LV		0.438	0.289			RV	100	0.244	0.207
		Mode	100	0.422	0.270				200	0.203	0.201
			200	0.393	0.257				500	0.085	0.166
			500	0.310	0.441				1000	0.123	0.189
			1000	0.204	0.430	PP			100	**0.851**	**0.703**
	G	LV		0.251	**0.735**				200	**0.845**	**0.605**
		Mode	100	0.239	**0.724**				500	**0.801**	**0.559**
			200	0.223	**0.726**				1000	**0.773**	**0.568**
			500	0.134	**0.718**	HPD		LV		0.317	0.235
			1000	0.016	**0.833**			Mean	1000	0.304	0.253

Table 3. Single indicative categories of nominal data layers that indicate *P. fruticosa* presence correctly in most cases (PPV > 0.5) if they are the most frequent spatial unit. Categories are listed in their best predictive radius (R). LV – local value.

Tabel 3. Nominaalsete andmekihtide üksikkategooriad, mis näitavad põõsasmarana esinemist õigesti enamikus kohtades (PPV > 0,5), kus see kategooria on sagedaseim raadiuses R. LV – lokaalväärtus (R = 0). Present land cover – kaasaegne maakate, soil map – mullakaart, historical map – ajalooline kaart, scrubland – põõsastik, Gleyic Rendzic Leptosol – paepealne gleimuld, Endogleyic Leptosol – gleistunud paepealne muld, all tussock areas – mätlik ala (ka koos teiste märkidega), tussocks with grassland – mätlik ala koos rohumaa märkidega, pure tussock area – vaid mätliku ala märgid, tussocks with scrubland – mätlik ala koos võsa märkidega.

Data layer *Andmekiht*	R [m]	Category *Kategooria*	PPV
Present land cover	200	Scrubland	0.88
Soil map	LV	Gleyic Rendzic Leptosol	0.53
	1000	Endogleyic Leptosol	0.52
Historical map	1000	All tussock areas	0.83
	LV	Tussocks with grassland	0.85
	200	Pure tussock area	0.78
	1000	Tussocks with scrubland	0.65

these four predictors are favourable, covers 1.4% (11.4 km²) of the terrestrial study area and has a density of find sites of 92.4% of all observed sites.

Alternative categories of the same features cover larger parts of the study area but are weak predictors. E.g. modal land cover category "scrubland" coincides with *P. fruticosa* presences to a great extent but the predictive value of land cover data layer, if all categories are included, was weak at all spatial and thematic scales (TSS ≤ 0.40) (Table 2). The binary variable *Dominating historical land cover within 1000 m is tussock area* is even more specific, having a high PPV = 0.83, but has a low TSS = 0.016. That means *P. fruticosa* can likely be found in large alvar patches mapped as tussock areas, but the feature is practically useless as an explanatory variable in all other regions of the study area.

There are also some less confident numerical predictors. *P. fruticosa* presence was recorded most often at altitudes of 19–35 m asl and in places where other presence sites occur within 1 km (Figure 4). More than a half (54%) of all recorded presences are located in square kilometres with no permanent habitants.

Excluding values

No landscape feature defines *P. fruticosa* presence for certain (excludes absence), that is, the species is not constantly present at any values of site features. The absence excluding feature values are possible only in case of an extremely abundant species since whatever current or previous limiting factor can exclude the species despite optimal conditions concerning other site features.

Regardless of the large number of observations, species absence is also rarely confirmed by map categories. Only a few of them meet the >1% of observations and no presences condition. *P. fruticosa* was never found where water is the most common land cover category within 100 m – it avoids sea coasts and does not occur at lake shores. More frequent sandy soil types were also found to exclude *P. fruticosa*. As site features, these are Umbric Gleysol (LkG) and Endogleyic Umbric Podzol (Lkg) in all radii and Haplic Podzol (L) in

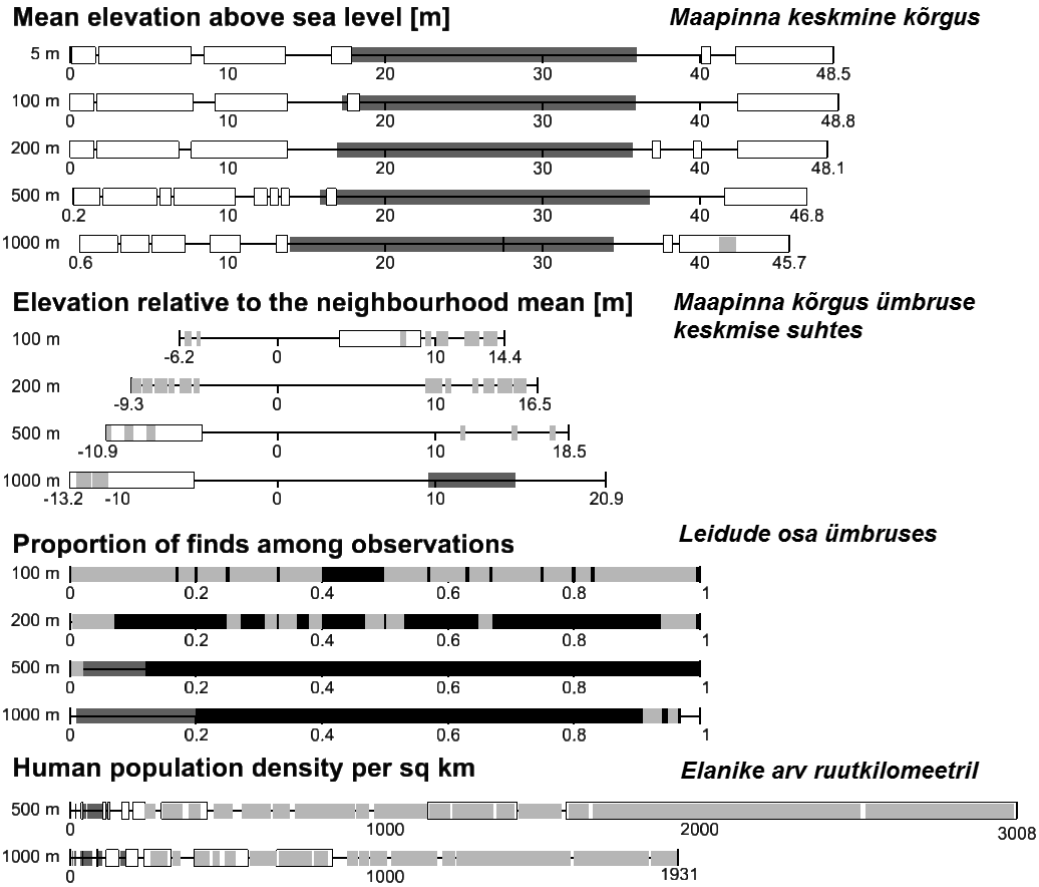


Figure 4. Range of values of numerical site features indicative of *P. fruticosa* presence or absence. Bars mark value intervals that contain at least 1% of all observations and match the following proportion of presences (PP): PP ≥ 0.5 (black), PP = 0.25–0.49 (dark grey), PP = 0.1–0.24 or < 1% of records (thin line), PP = 0 (empty box), light grey intervals contain no records.

Joonis 4. Põõsasmarana esinemist või puudumist näitavate numbriliste kohatunnuste väärtusvahemikud. Eristatud vahemikes on vähemalt 1% kõigist vaatluskohtadest ja need vastavad leiukohtade osale (PP) järgmiselt: PP ≥ 0.5 (must), PP = 0,25–0,49 (tumehall), PP = 0,1–0,24 või < 1% vaatlustest (peen joon), PP = 0 (seest valge ristkülik). Helehallides vahemikes andmed puuduvad.

radii up to 200 m. These excluding categories cover 5.1% (41.3 km²) of the terrestrial study area (Figure 5).

There are other map categories that never coincided with *P. fruticosa* sites, but the number of observations matching these areal categories did not meet the 1% frequency criterion. E.g. a few observations are recorded as being located in the sea according to the maps, but this is because of coast line drift due to sand movement and sediment deposition. The species was predominantly not found in places where no presences were recorded in the neighbourhood but, unexpectedly, only absence records in the vicinity are not a firm absence indicator, since there are exclusions – five solitary presence locations have only absence records within 1 km.

Among values of the numerical variables, 53 excluding intervals, often disrupted by occasional presences, were de-

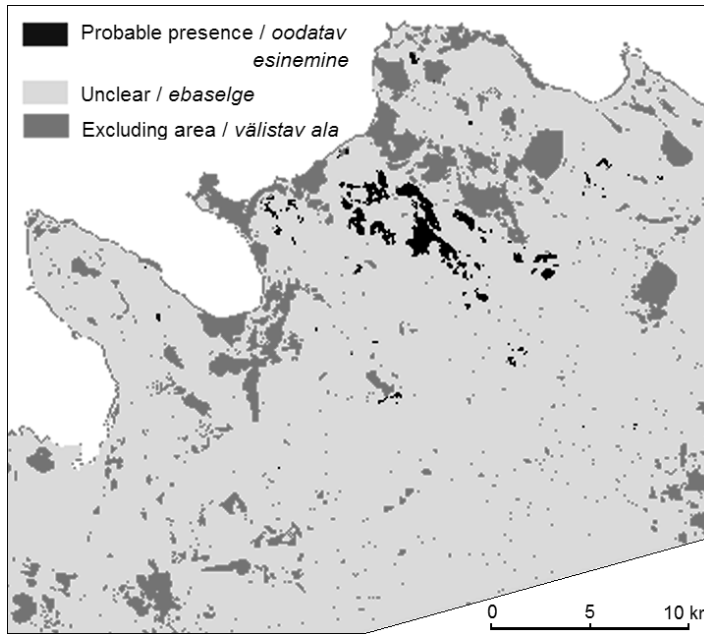


Figure 5. Area covered by feature values excluding *P. fruticosa* (dark grey), area where two of the indicative site features predict *P. fruticosa* presence (black), and undetermined area (light grey).

Joonis 5. Põõsasmarana esinemist välistavate kohatunnuste ala (tumehall), põõsamarana tõenäoline esinemisala vähemalt kahe kõrge indikaatorväärtusega tunnuse järgi (must) ja määratlemata ala (helehall).

limited (Figure 4). In general, *P. fruticosa* is absent in the lowest parts the study area (coastal plains) and in the highest altitude regions at the southern boundary of the area. Extreme values of relative elevation also exclude species presence. *P. fruticosa* was not found growing naturally in square kilometres where the human population exceeded 1575 inhabitants.

The occurrence of *P. fruticosa* remains undetermined in most (93.5%) of the study area (Figure 5) due to: 1) spatially close presence and absence records, 2) missing direct observations nearby or 3) low predictive value of the site features.

Spatial and thematic resolution

According to the TSS, thematically and spatially more detailed features are by and large more useful in predicting *P. fruticosa* occurrence than more generalized site characteristics (Table 4). Although this is not always the case, e.g. scrubland as the modal category has the highest predictive value when applied at a radius of 200 m, but not focally and not within 100 m (Figure 6). That means that species presence is statistically related to larger scrubland patches dominating in land cover within some hundreds of metres, rather than to single small groves.

Spatial generalization of the historical map increases the proportion of false negative predictions, as the PPV values tend to be higher at more generalized spatial and thematic scales (Table 4). This trend in mean values is mainly based on the historical map – more general versions of the map highlight the largest tussock areas, which coincide with the *P. fruticosa* populations, while most of the *P. fruticosa* sites remain unrecognized. Consequently,

Table 4. Mean TSS and PPV fit of single explanatory variable BCT models predicting the presence/absence of *P. fruticosa* depending on spatial scale and detail of the categorical explanatory variables. PLC – present land cover, HLC – historical land cover. The highest values in each subdivision are in bold.

Tabel 4. Kohatunnuseid ükshaaval sisaldavate BCT mudelite keskmine vastavus TSS ja PPV statistikute järgi põõsasmarana esinemise või puudumise prognoosimisel. Rasvases kirjas on kõrgeim väärtus igas alajaotuses. Tabeli ülemises osas on kolm temaatilise üldistuse taset kategooriate arvu järgi (üksikasjalik, keskmine ja üldine), alumises osas ruumiline detailsus tunnuse arvutamise ulatuse järgi meetrites. Muu tähistus nagu tabelis 1.

		Mean		PLC		HLC		Soil	
		TSS	PPV	TSS	PPV	TSS	PPV	TSS	PPV
Thematic resolution as the number of categories	Detailed	**0.402**	0.289	**0.303**	**0.303**	**0.367**	0.253	**0.535**	0.338
	Medium	0.397	0.331	0.309	0.300	0.353	0.337	0.530	**0.354**
	General	0.165	**0.372**	0.149	0.190	0.173	**0.747**	0.172	0.180
Spatial scale as kernel radius [m]	LV	**0.384**	0.335	**0.315**	0.265	**0.377**	0.438	**0.460**	**0.302**
	100	0.371	0.328	0.302	0.274	0.366	0.424	0.444	0.297
	200	0.355	0.326	0.292	0.280	0.344	0.417	0.428	0.296
	500	0.289	**0.354**	0.221	**0.301**	0.258	0.466	0.388	0.288
	1000	0.208	0.310	0.139	0.202	0.143	**0.484**	0.343	0.272

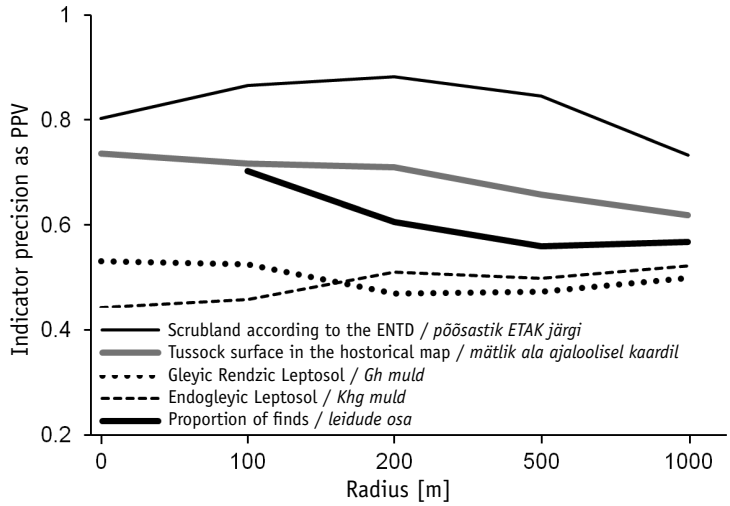


Figure 6. The predictive value of site features in recognizing *P. fruticosa* occurrence locations depending on the radius. Variables are compared at their medium level of thematic resolution.

Joonis 6. Kohatunnuse indikaatorväärtus positiivse prognoosiväärtuse (PPV) järgi sõltuvalt tunnuse arvutamise raadiusest. Nominaalseid tunnuseid on võrreldud keskmise temaatilise üldistuse tasemel, antud kategooria on selles raadiuses kõige sagedasem üksus selles andmekihis.

spatial generalization of the historical map increases the proportion of false negative predictions.

By virtue of a more generalized scale and lower planar precision, the characteristics derived from the historical map are less sensitive to spatial generalization compared to category merging (thematic generalization) (Table 4). *P. fruticosa* distribution largely matches the historical map tussock areas, a land cover category that was kept separate from other categories at all generalization levels. The explanatory value of the soil map drops abruptly when Gleyic Rendzic Leptosol is merged with other gleyic soils, and the value of the land cover decreases when merging scrubland with forest categories. A rule can be summarized, that the change in indicative value of a categorical layer, when altering the thematic detail, depends on how the most indicative single categories have been merged during generalization.

Discussion

P. fruticosa preference for Endogleyic Leptosol and Gleyic Rendzic Leptosol in our data confirms the notion that the natural populations of the species in northern Europe prefer thin calcareous soils (alvars), although, according to previous authors and our observations of cultivars, the species is tolerant of different soil conditions and the main limiting factor is shade (Elkington & Woodell, 1963; Marosz, 2004).

The limiting role of shade is confirmed by this study, as two of the best predictive site features (tussock signs in the historical map and scrubland as the present land cover) represent non-forest areas, since the tussock signs are on some occasions combined with grassland and bushes but never with signs of a forest area. The scrubland and forest categories are mutually exclusive categories also in the current land cover data. In nature, the undergrowth in forests is shaded by canopy but in alvar scrublands there is sufficient light for *P. fruticosa* to grow, as the shade from junipers in most places is not as dense as from the forest canopy.

P. fruticosa's preference for alvar soils in natural and semi-natural habitats can be explained by a weaker competition for light, since soil at these sites has been unsuitable for the establishment of dense forest in the period after the last glaciation. According to this, the main natural threat to *P. fruticosa* in the study area is afforestation of alvar grasslands, and to a lesser extent, competition with junipers, deciduous bushes and herbaceous species. Even a reduction in browsing and the closing down of a cattle farm near the population core area will probably have a long-term negative effect due to the gradual afforestation of the pastures.

Several methodological problems are related to distribution monitoring, including the selection and critical assessment of data sources. For example, the tussock areas on the historical map largely match the current distribution of *P. fruticosa,* but not everywhere. The first reason for this – a 1: 50 000 topographical map is spatially more generalized than the view of a field mapper; secondly – according to the experience from field observations, a mapped tussock area does not necessarily indicate *P. fruticosa* dwarf bushes, but can also denote *Vaccinium uliginosum*, *Betula nana*, *Calluna vulgaris* bushes or *Molinia caerulea Carex cespitosa* tussocs. There are tussock areas according to the historical map, where it is currently hard to assess whether *P. fruticosa* grew there 80 years ago or not (e.g built-up areas).

A number of studies indicate that a major source of sampling bias in the field mapping of animals and plants is imperfect detection, which leads to under estimated occurrence of the study object (Kéry *et al*., 2005, 2010; Bornand *et al*., 2014; Lahoz-Monfort *et al*., 2014), especially if its abundance is low (McCarthy *et al*., 2012). The detectability of even common, persistent vegetation

species is far below 100% (Kéry *et al*., 2006; Clarke *et al*., 2012). The recorded absence of a focal species is likely a false absence where the environment is conducive and the species was observed in the neighbourhood (Manceur & Kühn, 2014); however, data gathering using prior expert knowledge and relationships between existing occurrence data and environmental variables improves detection rate of rare objects (Albert *et al*., 2010). Although recognition of *P. fruticosa* in summer is mostly easy, as the plants reach above the dense grass layer and are blossoming, a few false absence records presumably occur in the original field records, but their share was presumably minimized by higher observation density at major distribution patches, combined with thinning out the excess absences and track points within 50 m of presence sites.

If we consider the width of the on-foot observed track to be approximately 10 m and ignore the partial overlap of the observed belt, the directly observed area still comprises only about 1% of the study area, showing that field surveys alone cannot completely cover any larger study area in detail. Consequently, the distribution pattern of a species has to be modelled, not only in the case of global and regional studies, but also if landscape complexity and mapping resolution makes it unrealistic to cover the total study area with direct observations.

Sites that were relatively suitable for the species but where it had not been registered were preferably observed in order to increase detection relative to the time spent. Preferential sampling based on a priori knowledge is a frequent decision to maximize discovery and cover the target-specific site variability (Luoto *et al*., 2001; Roleček *et al*., 2007; Giljohann *et al*., 2011; Steege *et al*., 2011). If we had applied a spatially random or regular sampling schema, most of our time in the field would have been spent on sites unsuitable for the species, the proportion of absence records would have been higher, and many occurrence sites would have remained undiscovered. The results of preferential sampling can better represent the species occurrence than most data in natural history museums and species occurrence registers commonly used in ecological and methodological studies. In this case, as: 1) the species absences were recorded; 2) all observations are made during the same season and during a relatively short period; 3) observations were made predominantly by one person; 4) data thinning reduced pseudo-replication and disproportion between the number of recorded presence and absence locations; 5) the use of proportions instead of the number of presences minimizes the effect of different observation density.

A special effort was made in this study to extract absence sites from recorded moving tracks. A species may absent in a location due to: 1) currently unsuitable environment; 2) dispersal limitations, historical factors, local extinctions or 3) due to incomplete and biased samples (Lobo *et al*., 2010). Some authors (Jiménez-Valverde *et al*., 2008; Gogol-Prokurat, 2011) even suggest discarding records on species' absence because of the high risk of false negative records. According to other authors, direct recording of absence sites reduces biases in predicted presence/absence (Phillips *et al*., 2009; Václavík & Meentemeyer, 2009; Phillips & Elith, 2013). If absence locations are not recorded in addition to presence sites, these should be generated as pseudo-absences, extracted at random from the sites in the study area where the species has not been recorded (Elith & Leathwick, 2007).

Registering absences during surveying can result in so-called zero-inflated data dominated by absence records. A high overall prediction fit is easy to obtain if a large part of the study area is clearly unsuitable for the species. To avoid the obscuring effect of surplus absences, eliminating surely expected absence areas has been suggested as the first step in distribution modelling (Mullahy, 1986; Heilbron, 1994; Welsh *et al*., 1996; Barry & Welsh, 2002), although it

is still not a common practice (Titeux *et al*., 2007). Another means to reduce the excess of absence records is spatial thinning as proposed in this study.

Site features for the recognition of *P. fruticosa* sites derived from topographical data layers at the most detailed spatial resolution were the most efficient according to TSS, but not following PPV (Table 2 and 4). Relatively low values of the PPV statistic compared to the TSS point to prevailing false positive predictions – which means, the spatially detailed site features tend to over-estimate species occurrence. Spatially smoothed features (except elevation asl) are not able to distinguish small patches, but are more reliable in species occurrence predictions. The spatially most detailed remote sensing data layer has also not been the best indicator for identification of vegetation units according to some other authors (Marceau *et al*., 1994; Ghosh *et al*., 2014). Supposedly, the optimal spatial scale has to be estimated empirically according to the dependent variable and the properties of explanatory data layers.

The predictive power of a categorical explanatory feature may depend remarkably on the order of merging the detailed categories. Formal merging of highly indicative categories of a categorical feature, while reducing thematic resolution, may increase or decrease the predictive value. The best option would be to reduce the only predictor to categories statistically matching the dependent variable. Unfortunately, such a perfect match cannot usually be found in real data.

In this study, categories of explanatory variables were merged according to prior knowledge about their meaning, which might not yield the best classification for the particular prediction task. A statistical algorithm for merging categories of a predictive variable is Exhaustive CHAID (Chi-squared Automatic Interaction Detector), which tests all pairs of categories and merges the categories according to the statistical significance to the classification of the dependent variable until a single splitting pair remains (Kass, 1980). Exhaustive CHAID may require significant computing time if the number of categories is large. Categories merged purely according to statistical significance for a particular target may not be easy to interpret meaningfully and applicable for other data.

In this study, several excluding values of categorical site features were found. Setting value intervals that exclude the species occurrence on the axis of a continuous feature assumes splitting the axis and checking if any of the intervals meet the given frequency and data density premises. The task would be a trivial braking data to quantiles if splitting only according to the number of records is needed, or production of a histogram if the class boundaries were given. The combination of the two conditions complicates the issue as the number of possible (split) points on a continuous axis is infinite. The proposed regions of frequency algorithm starts from data ordering and checks split points only at existing values.

Overlaying the spatial distribution of excluding and presence-predicting site features does not result in a detailed predictive distribution map, but outlines regions for further field survey and special modelling effort. Delimiting the species-excluding area enables attention to be focussed on the unclear presence/absence area and on the likely presence area when creating and calibrating predictive distribution models in subsequent studies on detailed distribution of the species. Presence/absence modelling or more field observations are needed for the undetermined area, while species abundance modelling could be a priority in the probable presence area.

Conclusions

1. All site features were significantly related to the occurrence of *P. fruticosa* but most of them are weak predictors.

2. The best predictors of *P. fruticosa* occurrence are 1) the proportion of presences in the neighbourhood, 2) moist thin calcareous soils according to the soil map, 3) larger scrubland patches according to the topographical database, and 4) tussock areas according to the topographical map from the 1930s.
3. Thematically and spatially the most detailed site features are not always the best indicators.
4. The change in indicative value of a categorical layer, when altering the thematic detail, depends on how the most indicative single categories have been merged during generalization.
5. Restricting a species distribution modelling by excluding the area where the species surely cannot be found could support the modelling effort, although, in this case, the species occurrence remains undetermined in most (93.5%) of the study area. Combination of different predictors at different scales is needed for detailed distribution modelling.
6. Innovative algorithms and tools for preprocessing distribution modelling data were applied for: 1) delimiting the species confirmed absence, probable and unclear occurrence area; 2) thinning of observed locations; 3) finding intervals in values of a numerical variable which match given ratio between states of a Boolean variable.

Acknowledgements. The author thanks the Estonian Land Board and Statistics Estonia for supplying cartographical and statistical data, Mare Remm for participating during field observations, Arno Kanal for advice on soil terminology, Kiira Mõisja for support in historical map digitizing, Malle Leht, Liina and Mare Remm for remarks on the manuscript, and Ilmar Part for proof-reading the manuscript. The investigation was financially supported by the Estonian Ministry of Education and Research (Project IUT 2-16).

References

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, 541–545.

Albert, C.H., Yoccoz, N.G., Edwards, T.C. Jr, Graham, C.H., Zimmermann, N.E., Thuiller, W. 2010. Sampling in ecology and evolution – bridging the gap between theory and practice. – Ecography, 33, 1028–1037.

Barry, S.C., Welsh, A.H. 2002. Generalized additive modelling and zero inflated count data. – Ecological Modelling, 157, 179–188.

Boria, R.A., Olson, L.O., 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.

Bornand, C.N., Kéry, M., Bueche, L., Fischer, M. 2014. Hide-and-seek in vegetation: time-to-detection is an efficient design for estimating detectability and occurrence. – Methods in Ecology and Evolution, 5, 433–442.

Boscolo, D., Metzger, J.P. 2009. Is bird incidence in Atlantic forest fragments influenced by landscape patterns at multiple scales? – Landscape Ecology, 24, 907–918.

Boulangeat, I., Gravel, D., Thuiller, W. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. – Ecology Letters, 15, 584–593.

Clarke, K.D., Lewis, M., Brandle, R., Ostendorf, B. 2012. Non-detection errors in a survey of persistent, highly-detectable vegetation species. – Environmental Monitoring and Assessment, 184, 625–635.

Dormann, C.F. 2011. Modelling species' distributions. – Jopp, F. *et al*. (eds.). Modelling complex ecological dynamics: an introduction into ecological modelling for students, teachers & scientists. Springer, 179–196.

Elith, J., Graham, C.H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography, 32, 66–77.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., *et al*. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography, 29, 129–151.

Elith, J., Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with

multivariate adaptive regression splines. – Diversity and Distributions, 13, 265–275.

Elkington, T.T., Woodell, S.R.J. 1963. *Potentilla fruticosa* L. (*Dasiphora fruticosa* (L.) Rydb.). – Journal of Ecology, 51, 769–781.

Ficetola, G.F., Bonardi, A., Mucher, C.A., Gilissen, N.L.M., Padoa-Schioppa, E. 2014. How many predictors in species distribution models at the landscape scale? Land use versus LiDAR-derived canopy height. – International Journal of Geographical Information Science, 28, 1723–1739.

Ghosh, A., Fassnacht, F.E., Joshi, P.K., Koch, B. 2014. A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. – International Journal of Applied Earth Observation and Geoinformation, 26, 49–63.

Giljohann, K.M., Hauser, C.E., Williams, N.S.G., Moore, J.L. 2011. Optimizing invasive species control across space: willow invasion management in the Australian Alps. – Journal of Applied Ecology, 48, 1286–1294.

Gogol-Prokurat, M. 2011. Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. – Ecological Applications, 21, 33–47.

Gorchakovskiy, P.L. 1960. On the distribution and habitat conditions of *Dasiphora fruticosa* (L.) Rydb. in connection with relict character of its localities in the Ural Mountains. – Zapiski Sverdlovskogo Otdeleniya Vsesoyuzhnogo Botanicheskogo Obshestva, 1, 3–22. (In Russian).

Graf, R.F., Bollman, K., Suter, W., Bugmann, H. 2005. The importance of spatial scale in habitat models: capercaillie in the Swiss Alps. – Landscape Ecology, 20, 703–717.

Guisan, A., Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecology Letters, 8, 993–1009.

Hanssen, A.W., Kuipers, W.J.A. 1965. On the relationship between frequency of rain and various meteorological parameters. – Mededelingen van de Verhandlungen, 81, 2–15.

Heilbron, D. 1994. Zero-altered and other regression models for count data with added zeros. – Biometrical Journal, 36, 531–547.

Holland, J.D., Bert, D.G., Fahrig, L. 2004. Determining the spatial scale of species' response to habitat. – BioScience, 54, 227–233.

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, 885–890.

Ju, J., Gopal, S., Kolaczyk, E.D. 2005. On the choice of spatial and categorical scale in remote sensing land cover classification. – Remote Sensing of Environment, 96, 62–77.

Kadmon, R., Farber, O., Danin, A. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecological Applications, 14, 401–413.

Kass, G.V. 1980. An exploratory technique for investigating large quantities of categorical data. – Applied Statistics, 29, 119–127.

Kéry, M., Gardner, B., Monnerat, C. 2010. Predicting species distributions from checklist data using site-occupancy models. – Journal of Biogeography, 37, 1851–1862.

Kéry, M., Royle, J.A., Schmid, H. 2005. Modeling avian abundance from replicated counts using binomial mixture models. – Ecological Applications, 15, 1450–1461.

Kéry, M., Spillmann, J.H., Troung, C., Holderegger, R. 2006. How biased are estimates of extinction probability in revisitation studies? – Journal of Ecology, 94, 980–986.

Lahoz-Monfort, J.J., Guillera-Arroita, G., Wintle, B.A. 2014. Imperfect detection impacts the performance of species distribution models. – Global Ecology and Biogeography, 23, 504–515.

Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J.A. 2006. Building statistical models to analyze species distributions. – Ecological Applications, 16, 33–50.

Liang, Y., He, H.S., Fraser, J.S., Wu, Z. 2013. Thematic and Spatial Resolutions Affect Model-Based Predictions of Tree Species Distribution. – PLoS ONE 8(7).

Lobo, J.M., Jiménez-Valverde, A., Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography, 33, 103–114.

Lonati, M., Pascale, M., Operti, B., Lombardi, G. 2014 Synecology, conservation status and IUCN assessment of *Potentilla fruticosa* L. in the Italian Alps. – Acta Botanica Gallica, 161, 159–173.

Luoto, M., Kuussaari, M., Rita, H., Salminen, J., von Bondsdorff, T. 2001. Determinants of distribution and abundance in the clouded apollo butterfly: a landscape ecological approach. – Ecography, 24, 601–617.

Mack, R.N., Harper, J.L. 1977. Interference in dune annuals: spatial pattern and neighbourhood effects. – Journal of Ecology, 65, 345–363.

Manceur, A.M., Kühn, I. 2014. Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. – Methods in Ecology and Evolution, 5, 739–750.

Marceau, D.J., Gratton, D.J., Fournier, R.A., Fortin, J.-P. 1994. Remote sensing and the measurement of geographical entities in a forested environment. 2. The optimal spatial resolution. – Remote Sensing of Environment, 49, 105–117.

Marosz, A. 2004. Effect of soil salinity on nutrient uptake, growth, and decorative value of four ground cover shrubs. – Journal of Plant Nutrition, 27, 977–989.

McCarthy, M.A., Moore, J.L., Morris, W.K., Parris, K.M., Garrard, G.E., Vesk, P.A., Rumpff, L., Giljohann, K., Camac, J., Bau, S.S., Friend, T., Harrison, B., Yue, B. 2012. The influence of abundance on detectability. – Oikos, 122, 717–726.

McPherson, J.M., Jetz, W., Rogers, D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – Journal of Applied Ecology, 41, 811–823.

Mullahy, J. 1986. Specification and testing of some modified count data models. – Journal of Econometrics, 33, 341–365.

O'Neill, R.V., DeAngelis, D.L., Waide, J.B., Allen, T.F.H. 1986. A Hierarchical Concept of Ecosystems. Princeton University Press, Princeton, NJ.

Phillips, S.J., Dudik, 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, 181–197.

Phillips, S.J., Elith, J. 2013. On estimating probability of presence from use–availability or presence–background data. – Ecology, 94, 1409–1419.

Reier, Ü., Leht, M. 1999. *Potentilla fruticosa* L. in Estonia and Latvia: origin, present situation and taxonomic position. – Sander, H. (ed.). Dendroloogilised uurimused Eestis 1, 23–36.

Remm, K. 2004. Case-based predictions for species and habitat mapping. – Ecological Modelling, 177, 259–281.

Remm, K. 2015. Classifications of areal categories in the Estonian Topographical Database, in the 1: 50 000 historical map from 1930ies, in the Estonian 1: 10 000 soil map, at three different thematic scales as used in the shrubby cinquefoil (*Dasiphora fruticosa* (L.) Rydb.) distribution studies in north-western Estonia. [WWW document]. – URL http://doi.org/10.13140/RG.2.1.3283.6645 [Accessed 3 May 2016].

Remm, K. 2016. Shrubby cinquefoil (*Dasiphora fruticosa*) cover records from a study site in the NW Estonia. [WWW document]. – URL http://doi.org/10.13140/RG.2.1.4987.6724 [Accessed 3 May 2016].

Remm, K., Kelviste, T. 2014. An online calculator for spatial data and its applications. – Computational Ecology and Software, 4, 22–34.

Remm, K., Linder, M., Remm, L. 2009. Relative density of finds for assessing similarity-based maps of orchid occurrence. – Ecological Modelling, 220, 294−309.

Roland, A.E., Smith, E.C. 1969. The flora of Nova Scotia, part II: The dicotyledons. – Proceedings of the Nova Scotian Institute of Science, 26, 277–743.

Roleček, J., Chytrý, M., Hájek, M., Lvončík, S., Tichy, L. 2007. Sampling design in large-scale vegetation studies: Do not sacrifice ecological thinking to statistical purism! – Folia Geobotanica, 42, 199–208.

Steege, H. ter, Haripersaud, P.P., Banki, O.S., Schieving, F. 2011. A model of botanical collectors' behavior in the field: never the same species twice. – American Journal of Botany, 98, 31–37.

Thompson, C.M., McGarigal, K. 2002. The influence of research scale on bald eagle habitat selection along the lower Hudson River, New York (USA). – Landscape Ecology, 17, 569–586.

Thuiller, W., Araújo, M.B., Lavorel, S. 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. – Journal of Vegetation Science, 14, 669–680.

Titeux, N., Dufrene, M., Radoux, J., Hirzel, A.H., Defourny, P. 2007. Fitness-related parameters improve presence-only distribution modelling for conservation practice: the case of the red-backed shrike. – Biological Conservation, 138, 207–223.

Turner, M.G., Constanza, R., Sklar, F.H. 1989. Methods to evaluate the perfomance of spatial simulation models. – Ecological Modelling, 48, 1–18.

Václavík, T., Meentemeyer, R.K. 2009. Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? – Ecological Modelling, 220, 3248–3258.

Vale, C.G., Tarroso, P., Brito, J.C. 2014. Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara–Sahel transition zone. – Diversity and Distributions, 20, 20–33.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. – Ecological Modelling, 88, 297–308.

Wiens, J.A. 1989. Spatial scaling in ecology. – Functional Ecology, 3, 385–397.

Zhou, W., Qian, Y., Li, X., Li, W., Han, L. 2014. Relationships between land cover and the surface urban heat island: seasonal variability and effects of spatial and thematic resolution of land cover data on predicting land surface temperatures. – Landscape Ecology, 29, 153–167.

Zurell, D., Jeltsch, F., Dormann, C.F., Schröder, B. 2009. Static species distribution models in dynamically changing systems: how good can predictions really be? – Ecography, 32, 733–744.

Erinevas ruumilises ja temaatilises mõõtkavas kohatunnuste valimine põõsasmarana (*Potentilla fruticosa* L.) leviku kaardistamiseks

Kalle Remm

Kokkuvõte

Uuriti kohatunnuste indikaatorväärtust ja selle sõltuvust üldistustasemest põõsasmarana (*Potentilla fruticosa* L. sün. *Dasiphora fruticosa* (L.) Rydb.) leviku detailseks kaardistamiseks selle liigi Baltimaade suurima loodusliku asurkonna piirkonnas Loode-Eestis. Välivaatlustel aastatel 2008–2014 läbiti jalgsi umbes 700 km. Vaatluspunktid harvendati vahemaale vähemalt 50 m, et ühtlustada vaatluste tihedust ruumilises ning esinemise ja puudumise vahekorra mõttes. Harvendamise järel jäi kasutusse 1459 leiukohta and 7327 puudumiskohta. Liigi esinemist/puudumist nendes kohtades seostati kohta kirjeldavate tunnustega, mis arvutati maakatte üksustest Eesti topograafilises andmekogus ja 1930ndate aastate 1: 50 000 kaardil ning mullakaardi, maapinna kõrguse, alaliste elanike tiheduse, liigi leidude tiheduse ja keskmise katvuse andmetest erinevas raadiuses. Selgitamaks leviku kaardistamisel võimalike kohatunnuste indikaatorväärtust, võrreldi seose tugevust ja olulisust liigi esinemise puudumise ja 60 üksiku erineval temaatilisel ja ruumilisel üldistustasemel kohatunnuse vahel BCT (*boosted classification tree*) mudeli abil. Lisaks tunnuste indikaatorväärtuste mõõtmisele on hinnangulise levikukaardistamise eeltöö jaoks uudsete metoodiliste võtetena esitatud vaatluste harvendamise ja binaarse muutuja klassi-de etteantud sageduskriteeriumitele vastavate numbrilise tunnuse väärtusvahemike leidmise algoritmid.

Uuringus selgus, et paljudest võimalikest kohatunnustest on liigi leviku kaardistamisel märkimisväärse indikaatorväärtusega vaid üksikud. Põõsasmarana leviku detailseks kaardistamiseks selle uuringu alal on need tunnused: leiukohtade osakaal ümbruses olevates vaatlustes, paepealne gleimuld või gleistunud paepealne muld mullakaardi, suuremad põõsastikualad topograafilise andmekogu järgi ja mättalise ala märgid ajaloolisel kaardil. Kohatunnuste andmekihtide ühte ja ainsat kõige tõhusamat üldistuse taset ei ole. Enamasti seostuvad detailsemad kohatunnused liigi esinemise või puudumisega paremini, kuid mitte alati. Liigi leviku modelleerimisel tasuks kombineerida erineva üldistustasemega andmeid. Nominaalse andmekihi indikaatorväärtuse muutus temaatilisel üldistamisel sõltub eelkõige kõrgema indikaatorväärtusega üksikkategooria teiste kategooriatega liitmisest.

Mingi nähtuse leviku hinnangulist kaardistamist toetab muuhulgas uurimisala osadeks jagamine. Kohatunnuseid ükshaaval kasutades õnnestus eristada ala, kus liigi esinemine on mõne tunnuse ja olemasolevate andmete järgi välistatud (leviku modelleerimisel võib selle osa uuris-

alast välja jätta) ja ala, kus liigi esinemine on tõenäoline. Selles piirkonnas on liigi esinemiskohti rohkem, mis võimaldab detailsemalt kirjeldada liigi paiknemismustrit ja ohtrust. Paraku enamikul uurimisalast (93,5%) ei ole koha tunnuseid ükshaaval kasutades liigi esinemine või puudumine selge. Seega on leviku modelleerimisel esmaseks ülesandeks leida liigi esinemist või puudumist kõige tõhusamalt prognoosivad kohatunnuste komplektid.

Received May 5, 2016, revised August 3, 2016, accepted August 30, 2016