BOREAL ENVIRONMENT RESEARCH 21: 53–70
ISSN 1239-6095 (print) ISSN 1797-2469 (online)

Helsinki 8 April 2016

The role of landscape, topography, and geodiversity in explaining vascular plant species richness in a fragmented landscape

Aleksi Räsänen[1)2)], Markku Kuitunen[1)], Jan Hjort[3)], Asta Vaso[1)], Tuomo Kuitunen and Anssi Lensu[1)]

[1)] *University of Jyvaskyla, Department of Biological and Environmental Science, P.O. Box 35, FI-40014 University of Jyvaskyla, Finland*
[2)] *current address: Department of Environmental Sciences, P.O. Box 65, FI-00014 University of Helsinki, Finland*
[3)] *Department of Geography, P.O. Box 3000, FI-90014 University of Oulu, Finland*

Received 15 Aug. 2014, final version received 4 Aug. 2015, accepted 16 July 2015

Räsänen A., Kuitunen M., Hjort J., Vaso A., Kuitunen T. & Lensu A. 2016: The role of landscape, topography, and geodiversity in explaining vascular plant species richness in a fragmented landscape. *Boreal Env. Res.* 21: 53–70.

We explained vascular plant species richness patterns in a 286 km^2 fragmented landscape with a notable human influence. The objective of this study was two-fold: to test the relative importance of landscape, topography and geodiversity measures, and to compare three different landscape-type variables in species richness modeling. Moreover, we tested if results differ when only native species are considered. We used generalized linear modeling based variation partitioning and generalized additive models with different explanatory variable sets. Landscape and topography explained the majority of the variation but the relative importance of topography and geodiversity was higher in explaining native species richness than in explaining total species richness. Differences between the three landscape type variables were small and they provided complementary information. Finally, topography and geodiversity often direct human action and can be ultimate causes behind both landscape variability and species richness patterns.

Introduction

Biodiversity, which can be defined as the variations in ecosystems, species, and genes, is often measured using vascular plant species richness as a proxy. While plants are just one taxon and partly depend on other organisms, they are the base of the food chain (Whittaker *et al*. 2001). Species richness is explained widely with species–energy relationships (e.g. Evans *et al*. 2005) and species–area relationships (e.g. Connor and McCoy 1979). There are also other factors that affect biodiversity. These factors include climate, which has partly same components as energy, historical factors, stress, stability, disturbance, ecological interactions (Fraser and Currie 1996, Whittaker *et al*. 2001), and environmental heterogeneity, which includes between-habitat and within-habitat variability, together with climatic, soil, and topographical heterogeneity (Stein *et al*. 2014). One part of environmental heterogeneity, which is often analyzed separately, is geodi-

Editor in charge of this article: Albert Porcar-Castell

versity. Geodiversity can be defined as the variability of the Earth's surface materials, forms, and physical processes (Gray 2013), and it has been found to have an effect on biodiversity (Anderson and Ferree 2010, Parks and Mulligan 2010, Hjort *et al*. 2012, Ruddock *et al*. 2013, Lawler *et al*. 2015).

When species richness is modeled on landscape scale, perhaps the most often used predictor variables are topographic variables (Guisan and Zimmerman 2000). Other predictors include, for example, different remotely-sensed variables, such as normalized difference vegetation index as a proxy of productivity (Nagendra 2001, Turner *et al*. 2003, Parviainen *et al*. 2009) and heterogeneity of spectral information as a proxy of habitat heterogeneity (Rocchini *et al*. 2010, 2011). Usually, when remotely-sensed data or spectral information is used in species richness mapping, traditional pixel-based analyses are used. Some of the methodologies that are widely used in landscape mapping, such as object-based image analysis, in which pixels are merged into meaningful objects (Blaschke *et al*. 2014), have not been tested in species richness mapping (Rocchini *et al*. 2010).

Thematic land-cover data that are based on remote sensing are often used in species richness models (Honnay *et al*. 2003, Thuiller *et al*. 2004). There are, however, few studies where thematic land-cover and continuous remote-sensing data, such as spectral values and indices, are compared. One example is research by Cord *et al*. (2014) which compared continuous remote-sensing data with land-cover data in mapping single tree-species occurrences. To the best of our knowledge, there are no studies in which different approaches for assessing landscape heterogeneity are compared. We tested different landscape heterogeneity measures in mapping vascular plant species richness. Two of these measures were object-based and one was based on spectral heterogeneity of remote sensing data.

Topographic variables can be considered to be parts of geodiversity, and they are widely used in explaining species richness. Studies, in which other geodiversity measures such as geomorphology (when not equated to topography), hydrology and soils are included, are fewer (Heikkinen and Neuvonen 1997, Lobo *et al*. 2001, Pausas *et al*. 2003, Titeux *et al*. 2009, Hjort *et al*. 2012). In addition, there are few studies where geodiversity measures are explicit, and where testing is systematic (Hjort *et al*. 2012). It has been observed that geodiversity measures improve species richness models in boreal landscapes in near-natural state (Hjort *et al*. 2012). We analyzed if explicit measures of geodiversity explain vascular plant species richness in a landscape which is fragmented primarily due to human influence.

The goal of this study was to (1) analyze the relative importance of landscape heterogeneity, geodiversity, and topography in modeling vascular plant species richness in a southern-boreal vegetation zone rural landscape in Finland, (2) compare three different approaches to assessing landscape heterogeneity, and (3) investigate if results differ when only native species are taken into account. Geodiversity was measured in terms of geological and hydrological richness. Two of the used landscape-heterogeneity measures were object-based and one of the object-based measures was based on species' habitat preferences (Rossi and Kuitunen 1996) and should have stronger ecological background. We compared how well a habitat type classification system that is based on species identification literature explains species diversity as opposed to spectral information and land-use/land-cover classification. We modeled native species richness and total species richness separately, since non-native species are rich in human habitats while the relative proportion of native species is higher in areas with less human disturbance. We tested if the relative role of different variables and variable groups differ in explaining native species richness versus total species richness. We also discuss the interrelationships between different explanatory variables and the role of human influence in species richness patterns.

Material and methods

Study area

We studied a 286 km^2 rural area (Fig. 1) in southern Finland (Kuitunen 2014) in the southern-boreal vegetation zone (Ahti *et al*. 1968).

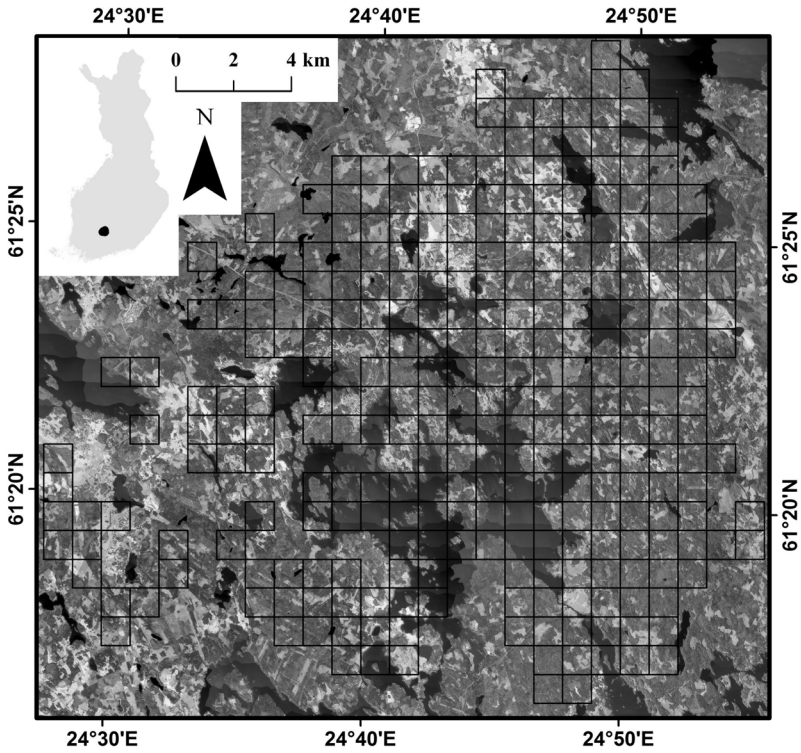

Fig. 1. A black and white version of color-infrared aerial image over the study area. Studied 1-km^2 grid cells are marked with squares. In the upper left corner, a map of Finland is shown with the black dot showing the study area. (Aerial imagery reprinted with permission from ©TerraTec Oy).

The geographic coordinates (WGS84) of the site are 61°16´–61°30´N and 24°26´–24°55´E. The main land-cover and land-use types in the area are coniferous and deciduous forests followed by (in descending order) lakes, agricultural areas and peatlands. Most of the forest area is used for timber production with rotation-based forestry and most of the peatlands are drained for forestry purposes. Agricultural areas (mostly fields), roads, and settlements are found in most parts of the area, and the landscape is fragmented. There are 742 (80–415 in each 1-km^2 grid cell) vascular plant species identified in the area (Kuitunen 2014). A total of 387 (68–242 in 1-km^2 grid cell) of these plants are native species.

Data sets

We used the following data: a Corine Land Cover (CLC) 2006 land-use/land-cover data from the Finnish Environment Institute at 25-meter resolution (©Finnish Environment Institute 2010, partly ©Finnish Forest Research Institute, Ministry of Agriculture and Forestry, National Land Survey, Population Register Centre), a habitat type classification (HTC, *see* Appendix 1), three-band (green, red, near infra-red) aerial imagery at 40-cm resolution taken in summer 2011 (TerraTec Oy, Helsinki, Finland; ©Finnish Forest Centre Pirkanmaa), airborne laser scanning data from years 2008 and 2012 together with a 1:10 000 resolution topographic database from the year 2010 (©National Land Survey of Finland), 1:20 000 digital Quaternary deposit (hereafter soil, ©Geological Survey of Finland 2007) and 1:200 000 digital bedrock maps (©Geological Survey of Finland 2009), as well as the vascular plant species inventory data (Kuitunen 2014).

In the used vascular plant species inventory data (Kuitunen 2014), the presence of vascular plant species inside 1-km^2 grid cells were surveyed between 1983 and 2011. Overall, 286 grid cells were inventoried. For each grid cell, the presence of different species was recorded but other factors such as species abundances were not collected. We used the overall number of species observed in each grid cell. Each of the grid cells was initially surveyed once by

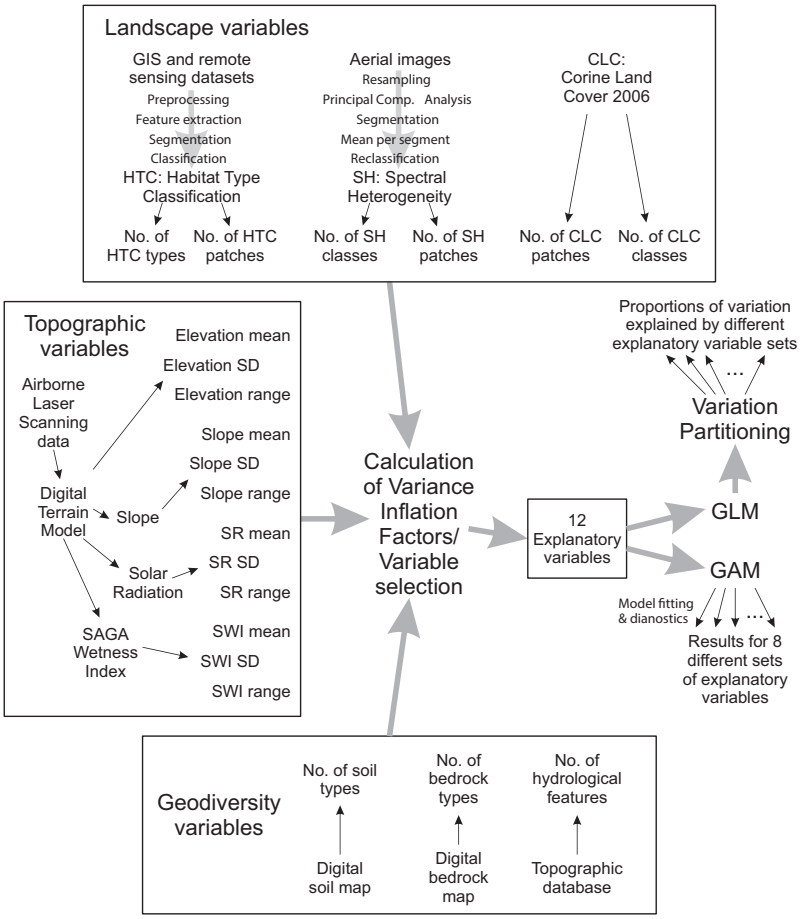

Fig. 2. Workflow of the used methodology. GLM refers to generalized linear model and GAM to generalized additive model.

one person. In some grid cells, additional species were later detected and added to the data after the initial inventory (Kuitunen 2014). The sampling effort of different grid cells varied a bit, but it should not have an effect on the species richness results. One of the grid cells was cross-checked by two persons, and there was a 5% difference in the number of species. A total of 20 grid cells were only partly inventoried. In those grid cells, the biotopes in the non-inventoried parts were similar to the biotopes in the inventoried parts; and thus, we did not expect to find significant number of new species from the unsurveyed parts of the grid cells. We also tested some of the analyses without these 20 partly-inventoried grid cells and the differences were small. For instance, the amount of deviance explained was 2 percentage-points greater in a generalized additive model with all variables when partly inventoried cells were omitted. With the help of Hämet-Ahti *et al*. (1998), we divided plants in the data set into two groups: native plants and other plants.

Overview of the methodology

The workflow of the methodology is illustrated in Fig. 2. First, we used already existing or constructed GIS layers that represented different landscape, topography or geodiversity features. Second, from each layer, we calculated one to three variables. We calculated these variables inside 1-km^2 grid cells that were the same grid cells as in the vascular plant species inventory data. Third, we removed multicollinear variables using variation inflation factor calculation. Fourth, we used calculated variables as explana-

tory variables with which we explained species richness of the grid cells using generalized linear and generalized additive models.

Explanatory variables

We divided the explanatory variables into three groups: landscape, topography, and geodiversity (Table 1). We calculated all explanatory variables inside 1-km^2 grid cells. The data sets had higher resolutions than 1 km^2 but we generalized finer resolution information to grid cells using variables that are elaborated below.

The landscape variables consisted of variables calculated from the HTC, spectral heterogeneity (SH), and CLC 2006 land-use/land-cover classifications. The HTC included 22 different habitat types (Table A1), and it is based on species identification literature (Hämet-Ahti *et al*. 1998). Based on the habitat preferences of different species, suitable habitat types were defined (Rossi and Kuitunen 1996). Our HTC does not directly tell that some habitat type has specific species composition. Instead, it should be interpreted to indicate that some species can exist in a specific habitat. We wanted to test if the theoretical HTC can explain observed species richness patterns. From the HTC, we calculated the number of habitat types and habitat patches per 1 km^2. SH refers to spectral information available in remote sensing data. SH can be seen as a proxy of environmental heterogeneity. Similar heterogeneity measures were used also previously in species richness mapping (Rocchini *et al*. 2010, 2011). We calculated SH based on the aerial imagery set (collected by Terratec) from the year 2011. We resampled all three color bands of the aerial images to 10-meter resolution using band and pixel-wise mean values. From the resampled combined image of the whole area, we calculated the first principal component (PC) to reduce the dimensionality of data. According to eigenvalues, PC1 could represent 85% of the total variation. We generalized the PC1 values using a segmentation technique

Table 1. Explanatory variables used in generalized linear modeling and generalize additive modeling. Variables that were used in the modeling after collinearity screening are set in boldface. VIF refers to variation inflation factors that are shown only for variables that were left after collinearity screening.

Group/layer	Variable	Min	Max	Median	VIF
Landscape					
habitat type classification (HTC)	**no. of types**	5	16	11	2.567
	no. of patches	13	385	148.5	1.876
Corine Land Cover (CLC)	**no. of types**	3	26	14	4.110
	no. of patches	6	328	171.5	2.408
spectral heterogeneity (SH)	**no. of types**	11	47	30	2.420
	no. of patches	27	196	108.5	–
Topography (topo)					
elevation (elev) (m a.s.l.)	**mean**	87.45	156.75	108.63	1.954
	SD	0.06	19.79	5.42	–
	range	1.08	65.59	26.91	–
slope (°)	mean	0.02	7.34	3.17	–
	SD	0.11	6.75	2.89	2.604
	range	2.71	45.71	21.85	–
solar radiation (SR) (WH m^{-2})	**mean**	653318	689543	673908	1.539
	SD	1366.52	51415	20695.1	–
	range	59647.3	513024	224682	–
SAGA wetness index (SWI)	mean	10.56	21.09	14.38	–
	SD	0.09	5.13	2.70	–
	range	1.88	16.40	12.70	2.026
Geodiversity (GD)					
bedrock	**no. of rock types**	1	5	2	1.208
Soil	**no. of soil types**	1	7	4	1.705
hydrology	**no. of hydrological features**	1	5	2	1.711

in which pixels are merged into homogenous objects called segments (*see* Blaschke *et al*. 2014). We used a segmentation technique that was identical to the segmentation used in habitat type classification (*see* Appendix 1). We calculated a mean value of PC1 per segment. Finally, we quantized these segment-related values to 64 classes using equal intervals. We calculated the continuous values for all segments and transformed them to classes to get meaningful objects and to reduce noise that is present in remote sensing data (for other benefits, *see* Blaschke *et al*. 2014). In initial evaluations, we tested different quantization options and 64 classes had the highest correlation with the response variable. From each 1-km^2 grid cell, we calculated the number of patches and the number of different principal component class values. From the CLC classification level four (*see* Appendix 2), we calculated the number of patches and variety of land-use/land-cover classes per 1 km^2. The level four is a Finnish modification and specification of the European CLC level three made to match the data set better with the Finnish context (*see* Appendix 2).

Topographic variables are used widely in species richness mapping and they are among the most important predictor variables (e.g. Guisan and Zimmermann 2000). In our case, topographic variables give information about, e.g., moisture conditions and microclimatic variations inside the area. We calculated the topographic variables based on the airborne laser scanner data. We constructed a digital terrain model at 10-meter resolution by triangulating points classified as ground. From the digital terrain model, we filled empty spots and pits. From the filled digital terrain model, we calculated the following layers: slope, solar radiation, and a SAGA wetness index. Solar radiation indicates how much irradiation each point in a landscape receives from the sun. We estimated total amount of solar irradiation per year from one day in each week in 30 minutes intervals using ArcGIS 10.2 (Esri, Redlands, CA, USA) and its Area Solar Radiation tool. The SAGA wetness index models moisture conditions using local and neighborhood slope and upslope contributing area (Böhner and Selige 2006). It is a modification of the standard topographic wetness index in which only local slope is taken into account. For these four layers, i.e. elevation, slope, solar radiation, and SAGA wetness index, we calculated mean values, standard deviations and ranges per 1-km^2 grid cells.

We compiled geodiversity variables from the digital soil and bedrock maps, and from the National Land Survey of Finland topographic database. More precisely, we computed measures of geodiversity by simply summing the total number of different soil types (e.g. clay, sand, till), rock types (e.g. gabbro, granodiorite, mica schist) and hydrological features (i.e. springs, streams, rivers, ponds and lakes) in the 1-km^2 study grid cells (Hjort and Luoto 2010). For example, the soil richness is the sum of different soil types regardless of the number and cover of the specific features in the study grid cells. The approach to compile geodiversity information was highly simplified but, in previous studies, it has been shown to describe variability in the Earth's surface materials, forms, and physical processes well at the landscape scale (Hjort and Luoto 2010, 2012).

Statistical analyses

We performed two sets of statistical analyses: one to explain and predict the total species richness per grid cell, and the other to the native species richness per grid cell. We acknowledged that native species and total species richness were highly correlated (Pearson's $r = 0.92$) but we wanted to analyze if their distributions had different explanations.

We used two approaches in the analyses. Using generalized linear models (GLM, Nelder and Wedderburn 1972), we examined the relative importance of different variable groups with variation partitioning (Real *et al*. 2003, Heikkinen *et al*. 2004). We used generalized additive models (GAM, Hastie and Tibshirani 1986) to test the relative strengths of different variables and variable groups. We tested, if they explain a larger proportion of the variation in species richness than GLMs and if there are non-linear dependencies between explanatory variables and dependent variables. In GAMs, the dependence between a dependent variable and explan-

atory variables is modeled semi-parametrically with smoother functions. GLMs and GAMs are among the most widely used statistical techniques in the species distribution models and they have been used successfully many times (e.g. Guisan *et al*. 2002, Guisan and Thuiller 2005). Some of the main advantages of these two methods are that different residual error distribution models can be assumed and that temporal and spatial dependencies can be taken into account. In addition, the probably non-linear effect of each explanatory variable to the (often transformed) response variable can be visualized for GAM models as we show below in results. Due to the overdispersion of the dependent variables, we used a quasipoisson distribution with a log-link function (Zuur *et al*. 2009) for the response. As an example, dispersion parameter was estimated to be 3.65 in our GLM run with all species and lasso penalty variable selection (Tibshirani 1996).

To deal with multicollinearity, we calculated variance inflation factors (Zuur *et al*. 2009) of potential explanatory variables. We used a threshold value of five for variance inflation factors (Zuur *et al*. 2009). First, from each topographic layer, we left only the topographic variable which had the highest correlation with the total plant species richness in the model. Second, we removed the variable which indicated the number of SH patches from the model. We used a total of 12 explanatory variables in the model (Table 1).

Variation partitioning refers to techniques in which variation of the response can be divided into several components. In variation partitioning, partial GLMs and a full GLM are calculated. In each partial GLM some variable group is left out of the model so that the importance of that group in explanation can be measured. We divided variation into seven components: into the pure effects of landscape, topography, and geodiversity as well as into the combined effects of each combination of two or three pure components. For detailed instructions how to do the calculations, *see* e.g. Real *et al*. (2003) and Heikkinen *et al*. (2004). We performed variable selection using lasso penalty (Tibshirani 1996) available the R package *lqa* (ver. 1.0-3., Ulbricht 2012) with R (ver. 2.15.2, R Core Team 2013). We sought an optimal regularization parameter with intervals of 0.1 using five-fold cross-validation and by minimizing deviance loss. In variation partitioning, we included also quadratic terms of explanatory variables.

We fitted GAM models with the R package *mgcv* (Wood 2006, 2011). In the smoothing parameters, we let the degrees of freedom vary between zero and four and we used restricted maximum likelihood estimation. We selected the variables by implementing an extra penalty term as suggested by Marra and Wood (2011). To analyze the effect of each landscape variable and geodiversity variables, we compared eight GAM models: (1) full model including all variables, three models that included topographic and geodiversity variables but only one landscape variable, i.e. (2) HTC, (3) CLC, or (4) SH type richness, (5) model omitting patch diversity variables, and three models including (6) only topography variables, (7) topography and geodiversity variables and (8) topography and landscape variables. We performed a six-fold cross-validation for model comparison. Finally, spatial autocorrelation of the full GAM residuals was minor (Moran's I was at maximum 0.15, $p < 0.001$). Therefore, we decided not to take spatial autocorrelation into account in the models. Overall, all model assumptions were met in different GLM and GAM runs and there were no clear structure in residual plots.

Results

According to the variation partitioning results, landscape and topographic variables explained most of the variation of the vascular plant species richness (Fig. 3). Independently, geodiversity explained little (0.2% to 1.8%) of the variation but the combined effect of geodiversity and landscape (6.8% to 16.1%) as well as all variable groups (14.7% to 15.4%) was considerable. Results were slightly different for the total species richness and native species richness. Most notably, the fraction explained by topographic diversity was greater for native species richness whereas the fractions explained by landscape, landscape and geodiversity together, and landscape and topography together were smaller for

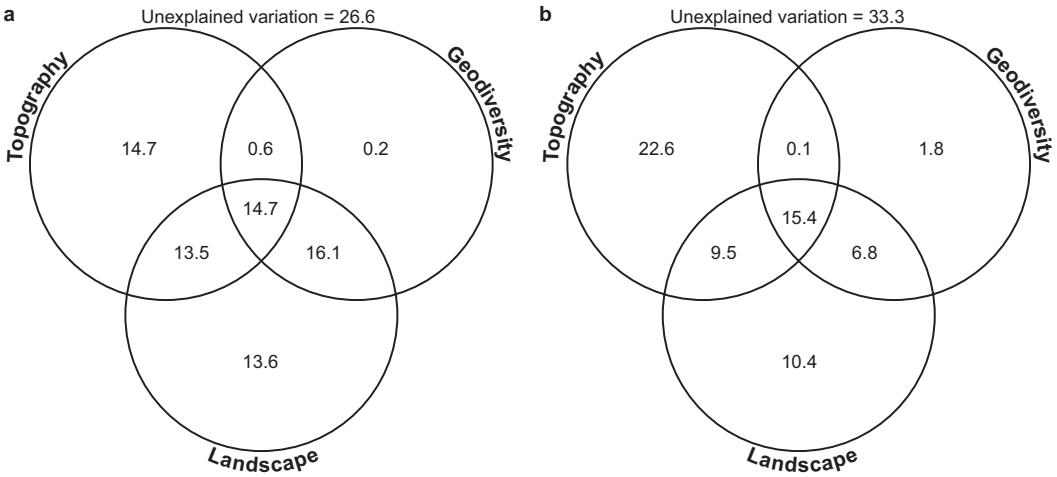

Fig. 3. Percentages of variance explained divided into different fractions based on partial and full generalized linear models. Fractions are independent contributions of three variable groups and shared contributions of different variable groups. Results are shown for (**a**) all species and (**b**) native species.

native species richness. The fraction of unexplained variation was larger for native species richness.

Many of the variables did not have great effects on the total or native species richness (Fig. 4). For instance, bedrock type richness and range of SAGA wetness index smoothers were both penalized to zero in both models. Some variables had clear positive or negative effects (Fig. 4). Different landscape-type diversity variables had rising or bell-shaped curves whereas the mean elevation showed a clear negative trend.

GLM with all variables explained a slightly greater proportion of the total species richness (75.6%) than GAM with all variables (74.6%; Table 2) because some explanatory variables had been penalized to zero in GAM. In explaining native species richness, GLM and GAM explained similar proportion of the total deviance. In cross-validation, GAM had greater Spearman's rank correlation between observed and predicted species richness (r_S = 0.828) than GLM (r_S = 0.823). Of the GAMs which included only a subset of variables, the models excluding patch variables (74.3%) or geodiversity variables (73.8%) explained little less of the species richness than the full model (74.6%). Actually, the model without geodiversity (r_S = 0.838) had higher prediction capability in cross-validation than the full model (r_S = 0.828). Of the GAMs that included only one landscape variable, GAM with CLC types had slightly higher prediction capability than GAMs with HTC types or SH types in explaining the total species richness and predicting total and native species richness. GAM with SH types outperformed by narrow margin GAM with HTC in all comparisons. GAM with geodiversity and topography variables had higher explanation and prediction capability than the GAM with topography variables only. The explanation and prediction capability of these topography-only models was lower than of all the other models (Table 2).

Discussion

The role of landscape, topography and geodiversity

A large part of the variance in vascular plant species richness could be explained using topographic and landscape variables (Fig. 3). Of different explanatory variables, mean elevation and different landscape type diversity variables had the most pronounced relationships with species richness (Fig. 4). Generally, the number of species was higher in areas with larger landscape variability (Stein *et al*. 2014; Fig. 4). Usually the areas with larger landscape variability have also stronger human influence, i.e., include agricul-

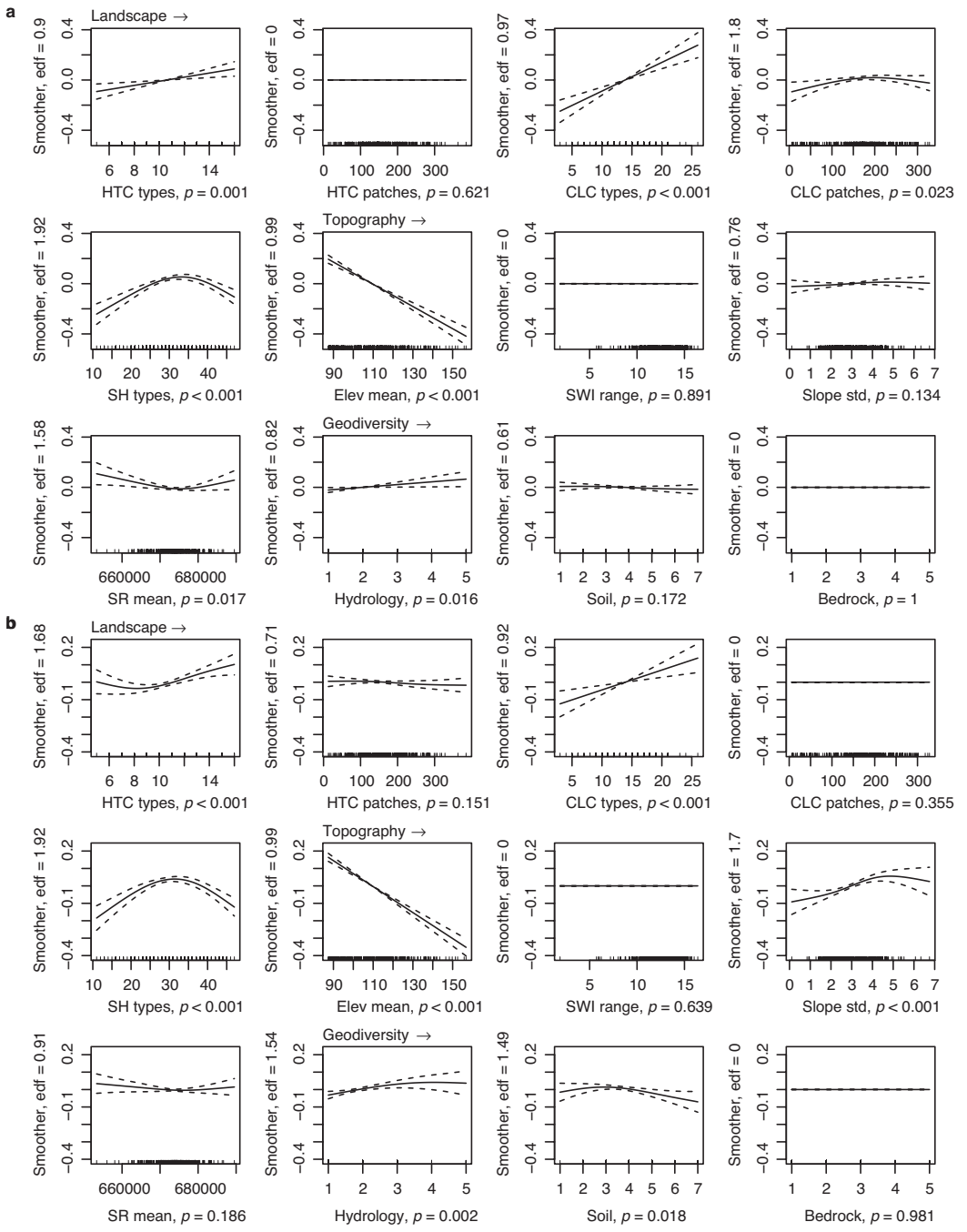

Fig. 4. Generalized additive model smooth estimates of the models with all variables for (**a**) total species richness and (**b**) native species richness. Two standard errors above and below the estimates are also plotted with dashed lines. The statistical significance of each variable (*p* value) in the model is given next to the name of the variable. *x*-axes represent explanatory variable values and *y*-axes are smoother-function values used in predicting the log-transformed response variable. When the drawn curve is above zero, then at those values the explanatory variable increases the total or native species richness, and when the curve is below zero, at those values the explanatory variable decreases the species richness as compared with the baseline set by the intercept term of the model. In the *y*-axes labels, edf = estimated degrees of freedom of the smoother function. Abbreviations of explanatory variables are given in column LAYER of Table 1.

tural or artificial areas, and the areas with stronger human influence incorporate more species (Honnay *et al*. 2003, Wania *et al*. 2006, Pautasso 2007). It is argued that in areas affected by humans, the degree of disturbance and nutrient availability varies; hence, there may exist more niches for different plant species (Honnay *et al*. 2003). Nevertheless, the GAM smoothers of spectral heterogeneity (SH) types were bell-shaped (Fig. 4). There were, thus, weak indications that species richness could decrease in very fragmented areas with many different habitat types. Other explanation is that the thematic resolution of SH is too fine. Inside one meaningful landscape or habitat type, different spectral types can be found. These spectral types can look different in the remote sensing data but can be essentially identical habitats and have similar species composition.

Topographic variables were among the variables with the highest explanatory power and topography explained the greatest amount of variance in variation partitioning (Figs. 3 and 4, Table 2). In previous studies, topography variables are frequently used and they are noticed to be working well (e.g. Guisan and Zimmermann 2000, Stein *et al*. 2014). In this study, mean elevation was the most important topographic variable (Fig. 4). Also in previous studies in boreal landscapes, mean elevation was among the predictors with the highest explanatory capability (e.g. Heikkinen and Neuvonen 1997, Parviainen *et al*. 2008). Those study areas were located further north and the elevational differences were greater. In a relatively flat landscape, such as our study area, elevation does not have a large climatic effect since the difference between highest and lowest mean elevation was not more than 70 m. Nonetheless, mean elevation is probably a proxy for other factors such as soil productivity and moisture as well as proximity of waterbodies. In addition, human influence and land-use history concentrate in the areas close to waterbodies and on productive lands. Elevation may also be a proxy for human influence and has an effect on landscape type patterns.

Geodiversity variables were not among the most important explanatory variables and geodiversity explained independently little of the variation in variation partitioning analysis (Figs. 3 and 4). Its explanatory power was slightly greater in explaining the native species richness than the total species richness (Fig. 3). In a previous study, where geodiversity measures were included (Hjort *et al*. 2012), geodiversity explained more of the variation in vascular plant species richness than climatic and topographic variables. In their study, study areas were pre-

Table 2. Results of full generalized linear model (GLM) and different generalized additive model (GAM). We calculated explained deviance, r^2 (adjusted) values and Spearman's rank correlation coefficients (r_s) between observed and predicted species richness for the different model. A six-fold cross-validation was used and presented values are mean values of six calibration or test sets. Topo refers to topography, GD to geodiversity, HTC to habitat type classification, CLC to Corine Land Cover and SH to spectral heterogeneity.

		All species				Native species			
		Calibration			Test	Calibration			Test
		explained deviance (%)	r^2 adjusted	r_s	r_s	explained deviance (%)	r^2 adjusted	r_s	r_s
GLM	all variables	75.6		0.863	0.823	68.3		0.823	0.777
GAM	all variables	74.6	0.743	0.860	0.828	68.3	0.673	0.826	0.794
	topo + GD + HTC	64.6	0.643	0.809	0.780	62.2	0.611	0.787	0.759
	topo + GD + CLC	71.5	0.706	0.845	0.818	62.8	0.613	0.817	0.791
	topo + GD + SH	68.8	0.677	0.830	0.793	63.4	0.625	0.789	0.763
	patches omitted	74.3	0.742	0.858	0.827	67.9	0.670	0.826	0.797
	topography only	46.4	0.448	0.690	0.663	49.0	0.476	0.660	0.617
	topo + GD	59.4	0.583	0.778	0.745	57.8	0.561	0.747	0.718
	topo + landscape	73.8	0.736	0.855	0.838	66.0	0.652	0.818	0.808

dominantly natural, and landscape measures were not included. In our study, the explanatory power of models including topography and geodiversity was about the same magnitude as the explanatory power of geodiversity, topography and climate by Hjort *et al*. (2012). Moreover, in Hjort *et al*. (2012), geodiversity included a measure of geomorphological richness. Inclusion of geomorphology could potentially increase the explanatory power of geodiversity measures. Nevertheless, geomorphological mapping requires a skilled interpreter and is time-consuming whereas other geodiversity measures are rather quick to quantify over larger areas using existing GIS data sets.

It has been found that geodiversity is positively linked with habitat diversity (Jačková and Romportl 2008). The greater the geodiversity values are the more diverse landscapes are in regards of habitat or land-use/land-cover types. In our study, where geodiversity explained independently little of the total deviance, the combined effect of geodiversity and landscape was greater than the effect of geodiversity alone (Fig. 3). Some geodiversity features such as mafic and intermediate bedrock types or esker deposits were also used in constructing the habitat type classification (Appendix 1). Additionally, agricultural areas, residential areas or human influence are not distributed evenly or randomly. Human influence is usually strongest on, e.g., high productive soils and near waterbodies, which both can be regarded as measures of geodiversity. Geodiversity may, together with topography, direct landscape variability and human influence. It can be possible that geodiversity and topography are the ultimate factors in explaining biological diversity but especially in areas with notable human influence, more proximate factors such as landscape features may have stronger explanatory power. More research is needed to check if this argument is correct. Furthermore, in the study area, some portions of geodiversity, such as rock types have a stronger effect on the richness of bryophytes than vascular plant species (Kuitunen 2014). Finally, the role of geodiversity could be greater if, other types of geodiversity variables than simple richness, i.e. number of types in grid cells, would be used (e.g. Ruban 2010, Beier *et al*. 2015).

Landscape type variables

Corine Land Cover (CLC) and spectral heterogeneity (SH) had as high or higher explanatory capabilities than habitat type classification (HTC) (Fig. 4 and Table 2). This was somewhat surprising since CLC and SH do not have an ecological background as opposed to the used HTC. According to the GAM results, CLC had higher explanatory capability than HTC in explaining total species richness, whereas in explaining native species they were on par (Table 2). One reason behind this difference might be that in CLC there are more agricultural and artificial area classes (number of classes is 12) than in HTC (five classes) (Appendices 1 and 2). Most of the non-native species in the area inhabit these human habitats whereas native species are more diverse in natural habitats. Among natural areas, CLC had only a little more classes (19 classes) than HTC (17 classes) and the classes were different. SH had a higher correlation with CLC than with HTC. The reason is probably that, based on visual interpretation, SH had a higher variation in artificial and agricultural areas than HTC.

Based on the GAM results (Table 2), all three landscape heterogeneity measures gave complementary information, probably because classes were divided differently in all three measures. We, therefore, advise to use different habitat type or land-use/land-cover type classifications in species richness modeling, if the intention is to explain as much as possible of the variation. Yet, different landscape type variables can be collinear. In our case, HTC and CLC type variables were rather highly correlated and they had the highest variance inflation factor values (Table 1) but the values were under the suggested thresholds (Zuur *et al*. 2009).

According to our results, the variable derived directly from remotely-sensed data, SH, explained slightly less of the variation in species richness than CLC but slightly more than HTC (Fig. 4 and Table 2). In the study by Cord *et al*. (2014), continuous remote-sensing variables outperformed thematic land-cover data. Their remotely sensed variables included numerous temporal, net primary productivity and seasonality metrics derived from multi-temporal MODIS

satellite data and they studied a single tree-species in the whole area of Mexico. Even simple remotely-sensed measures, such as SH, can be on par with thematic data but more specific measures based on remotely-sensed data can reveal unseen features in thematic data. At our scale and study area, more specific measures may not work as well as in the study by Cord *et al*. (2014) due to rather high local landscape variability in a rather fine-scale data. In our case, metrics — such as normalized difference vegetation index, — would merely be proxies of landscape heterogeneity than productivity which they ideally should measure (*see* e.g. Pettorelli *et al*. 2005). These indices are probably more applicable on natural areas than on natural-agricultural mosaic (*see* e.g. Parviainen *et al*. 2009).

Two of our used landscape measures, HTC and SH, were object-based instead of pixel-based. While we did not compare pixel-based measures with object-based measures *per se*, object-based measures worked well since HTC and SH variables explained a significant amount of the variation of species richness. In initial analyses, object-based SH had stronger correlations with species richness than pixel-based SH, but pixel-based and object-based measures were highly collinear. In future studies, pixel-based measures should be compared with object-based measures to test if their predictive capabilities differ.

Differences between native species richness and total species richness

The correlation between native species richness and total species richness was considerable (Pearson's $r = 0.92$) and native species were a subset of total species. In other words, the pattern in which the relative strength of different explanatory variables was approximately similar in explaining both native and total species richness (Fig. 3), may be an artifact, i.e. trivial but real (*see* Palmer *et al*. 2008).

Despite the strong correlations between native and total species richness there were some differences in the results (Fig. 3 and Table 2). Topographic variables and geodiversity variables were more important in explaining native species richness than in explaining total species richness. Vice versa, the relative effect of landscape was smaller on native species richness. One reason behind this difference might be that human influence does not increase native species richness as much as it increases total species richness. It can also point out the fact that geophysical factors have an effect on species richness, but due to human intervention this effect is not always visible and straightforward. On the other hand, when native species richness was the dependent variable, a smaller amount of variance could be explained. This might highlight the importance of other factors and random processes in explaining native species richness. More research is needed in evaluating the relationship between native species richness and total species richness and if their distributions have different explanations.

Comparison of GLM and GAM

GLM and GAM had approximately similar explanatory and predictive capabilities (Table 2). Most of the relationships modeled with GAM were linear (Fig. 4) and it can be thought that there was little need to model nonlinear relationships. Although GAM did not have a higher explanatory and predictive capability than GLM, GAM revealed that some of the variables had nonlinear relationships with species richness (Fig. 4). Those were not revealed with GLM. We argue that modeling with GAMs is not needed in many cases unless the relationships between the dependent and the explanatory variables are expected to be nonlinear, or if the shape of the relationship is not known at all.

Conclusions

We explained vascular plant species richness in a fragmented landscape with notable human influence. We examined the relative role of landscape, topography, and geodiversity, compared three different landscape heterogeneity measures, and modeled total and native species richness separately using GLMs and GAMs. Based on the results, we draw three main conclusions. Firstly,

majority of species richness was explained with landscape and topography variables whereas geodiversity explained little of the variation. Topography and geodiversity often direct human action and can thus be ultimate causes behind both landscape heterogeneity and species richness. Secondly, three landscape heterogeneity measures (HTC, SH, CLC) gave complementary information. Differences between the measures were small with CLC having the highest explanatory capability. Although HTC had a stronger ecological background, it was outperformed by the other two measures. Thirdly, in explaining native species richness, the relative role of topography and geodiversity, and the amount of unexplained variation was larger than in explaining total species richness. It may be that human influence is smaller and random processes more important on native species richness than on total species richness, but this must be addressed in future research.

Acknowledgements: This research was funded by Maj and Tor Nessling Foundation. We thank Raino Lampinen from Finnish Museum of Natural History for providing us the species inventory data. We are grateful to Antti Peltonen from Finnish Forest Centre Pirkanmaa for providing us aerial imagery from the area. MK received from Jenny and Antti Wihuri foundation a sabbatical scholarship for the year 2013 that was partly funded also by the EU IMPERIA-project (LIFE11 ENV/FI/905) and the University of Jyväskylä. JH acknowledges the Academy of Finland (grant numbers 267995 and 285040).

References

Ahti T., Hämet-Ahti L. & Jalas J. 1968. Vegetation zones and their sections in northwestern Europe. *Ann. Bot. Fennici* 5: 169–211.

Anderson M.G. & Ferree C.E. 2010. Conserving the stage: Climate change and the geophysical underpinnings of species diversity. *PLoS ONE* 5(7): e11554, doi:10.1371/journal.pone.0011554.

Beier P., Sutcliffe P., Hjort J., Faith D.P., Pressey R.L. & Albuquerque F. 2015. A review of selection-based tests of abiotic surrogates for species representation. *Conserv. Biol.* 29: 668–679.

Blaschke T., Hay G.J., Kelly M., Lang, S., Hofmann P., Addink E., Queiroz Feitosa R., van der Meer F., van der Werff H., van Coillie F. & Tiede D. 2014. Geographic object-based image analysis — towards a new paradigm. *ISPRS J. Photogramm.* 87: 180–191.

Böhner J. & Selige T. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In: Böhner J., McCloy K.R. & Strobl J. (eds.), *SAGA — Analysis and modelling applications*, Göttinger Geographische Abhandlungen, vol. 115, pp. 13–28.

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

Breiman L. & Cutler A. 2007. *Random forest: classification description*. Available at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Connor E.F. & McCoy E.D. 1979. The statistics and biology of the species area relationship. *Am. Nat.* 113: 791–833.

Cord A.F., Klein D., Mora F. & Dech S. 2014. Comparing the suitability of classified land cover data and remote sensing variables for modeling distribution patterns of plants. *Ecol. Model.* 272: 129–140.

Evans K.L., Warren P.H. & Gaston K.J. 2005. Species–energy relationships at the macroecological scale: a review of the mechanisms. *Biol. Rev.* 80: 1–25.

Fraser R.H. & Currie D.J. 1996. The species richness-energy hypothesis in a system where historical factors are thought to prevail: coral reefs. *Am. Nat.* 148: 138–159.

Gallant J.C. & Dowling T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resour. Res.* 39: 1347–1360.

Grabs T., Seibert J., Bishop K. & Laudon H. 2009. Modeling spatial patterns of saturated areas: a comparison of the topographic wetness index and a dynamic distributed model. *J. Hydrol.* 373: 15–23.

Gray M. 2013. *Geodiversity: valuing and conserving abiotic nature*, 2nd ed., John Wiley & Sons, Chichester, UK.

Guisan A. & Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8: 993–1009.

Guisan A. & Zimmermann N.E. 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135: 147–186.

Guisan A., Weiss S.B. & Weiss A.D. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* 143: 107–122.

Guisan A., Edwards T.C.Jr. & Hastie T. 2002. Generalized linear and generalized additive models in studies of species distributions. *Ecol. Model.* 157: 89–100.

Hämet-Ahti L., Suominen J., Ulvinen T. & Uotila P. (eds.), 1998. *Retkeilykasvio*, 4th ed. Luonnontieteellisen keskusmuseon kasvimuseo, Helsinki.

Haralick R.M., Shanmugam K. & Dinstein I. 1973. Textural features for image classification. *IEEE T. Syst. Man. Cyb.* 3: 610–621.

Hastie T. & Tibshirani R. 1986. Generalized additive models. *Stat. Sci.* 1: 297–318.

Heikkinen R.K. & Neuvonen S. 1997. Species richness of vascular plants in the subarctic landscape of northern Finland: modelling relationships to the environment. *Biodivers. Conserv.* 6: 1181–1201.

Heikkinen R.K., Luoto M., Virkkala R. & Rainio K. 2004. Effects of habitat cover, landscape structure and spatial variables on the abundance of birds in an agricultural—forest mosaic. *J. Appl. Ecol.* 41: 824–835.

Hjort J. & Luoto M. 2010. Geodiversity of high-latitude landscapes in northern Finland. *Geomorphology* 115: 109–116.

Hjort J. & Luoto M. 2012. Can geodiversity be predicted from space? *Geomorphology* 153–154: 74–80.

Hjort J., Heikkinen R.K. & Luoto M. 2012 Inclusion of explicit measures of geodiversity improve biodiversity models in a boreal landscape. *Biodivers. Conserv.* 21: 3487–3506.

Honnay O., Piessens K., Van Landuyt W., Hermy M. & Gulinck H. 2003. Satellite based land use and landscape complexity indices as predictors for regional plant diversity. *Landscape Urban Plan.* 63: 241–250.

Jačková K. & Romportl D. 2008. The relationship between geodiversity and habitat richness in Šumava National Park and Křivoklátsko PLA (Czech Republic): a quantitative analysis approach. *J. Landscape Ecol.* 1: 23–38.

Kalliola R. 1973. *Suomen kasvimaantiede*. WSOY, Porvoo.

Kuitunen T. 2014. *Luopioisten kasvisto*. Available at http://www.luopioistenkasvisto.fi/.

Lawler J.J., Ackerly D.D., Albano C.M., Anderson M.G., Dobrowski S.Z., Gill J.L., Heller N.E., Pressey R.L., Sanderson E.W. & Weiss S.B. 2015. The theory behind, and the challenges of, conserving nature's stage in a time of rapid change. *Conserv. Biol.* 29: 618–629.

Liaw A. & Wiener M. 2002. Classification and regression by randomForest. *R News* 2: 18–22.

Lobo J.M., Castro I. & Moreno J.C. 2001. Spatial and environmental determinants of vascular plant species richness distribution in the Iberian Peninsula and Balearic Islands. *Biol. J. Linn. Soc.* 73: 233–253.

Marra G. & Wood S. 2011. Practical variable selection for generalized additive models. *Comput. Stat. Data An.* 55: 2372–2387.

Murphy P.N.C., Ogilvie J. & Arp P.A. 2009. Topographic modeling of soil moisture conditions: a comparison and verification of two models. *Eur. J. Soil Sci.* 60: 94–109.

Murphy P.N.C., Ogilvie J., Connor K. & Arp P.A. 2007. Mapping wetlands: a comparison of different approaches for New Brunswick, Canada. *Wetlands* 27: 846–854.

Nagendra H. 2001. Using remote sensing to assess biodiversity. *Int. J. Remote Sens.* 22: 2377–2400.

Nelder J.A. & Wedderburn R.W.M. 1972. Generalized linear models. *J. Roy. Stat. Soc. A* 135: 370–384.

Palmer M.W., McGlinn D.J. & Fridley J.D. 2008. Artifacts and artifictions in biodiversity research. *Folia Geobot.* 43: 245–257.

Parks K.E. & Mulligan M. 2010. On the relationship between a resource based measure of geodiversity and broad scale biodiversity patterns. *Biodivers. Conserv.* 19: 2751–2766.

Parviainen M., Luoto M. & Heikkinen R.K. 2009. The role of local and landscape level measures of greenness in modelling boreal plant species richness. *Ecol. Model.* 220: 2690–2701.

Parviainen M., Luoto M., Ryttäri T. & Heikkinen R.K. 2008. Modelling the occurrence of threatened plant species in taiga landscapes: methodological and ecological perspectives. *J. Biogeogr.* 35: 1888–1905.

Pausas J.G., Carreras J., Ferré A. & Font X. 2003. Coarse-scale plant species richness in relation to environmental heterogeneity. *J. Veg. Sci.* 14: 661–668.

Pautasso M. 2007. Scale dependence of the correlation between human population presence and vertebrate and plant species richness. *Ecol. Lett.* 10: 16–24.

Pettorelli N., Vik J.O., Mysterud A., Gaillard J.-M., Tucker C.J. & Stenseth N.C. 2005. Using the satellite derived NDVI to assess ecological responses to environmental change. *Trends Ecol. Evol.* 20: 503–510.

R Core Team 2013. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Räsänen A., Kuitunen M., Tomppo E. & Lensu A. 2014. Coupling high-resolution satellite imagery with ALS-based canopy height model and digital elevation model in object-based boreal forest habitat type classification. *ISPRS J. Photogramm.* 94: 169–182.

Real R.A., Barbosa M., Porras D., Kin M.S., Márquez A.L., Guerrero J.C., Palomo L.J., Justo E.R. & Vargas J.M. 2003. Relative importance of environment, human activity and spatial situation in determining the distribution of terrestrial mammal diversity in Argentina. *J. Biogeogr.* 30: 939–947.

Riley S.J., DeGloria S.D. & Elliot R. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5: 23–27.

Rocchini D., McGlinn D., Ricotta C., Neteler M. & Wohlgemuth T. 2011. Landscape complexity and spatial scale influence the relationship between remotely sensed spectral diversity and survey-based plant species richness. *J. Veg. Sci.* 22: 688–698.

Rocchini D., Balkenhol N., Carter G.A., Foody G.M., Gillespie T.W., He K.S., Kark S., Levin N., Lucas K., Luoto M., Nagendra H., Oldeland J., Ricotta C., Southwort J. & Neteler M. 2010. Remotely sensed spectral homogeneity as a proxy of species diversity: recent advances and open challenges. *Ecol. Inform.* 5: 318–329.

Rossi E. & Kuitunen M. 1996. Ranking of habitats for the assessment of ecological impact in land use planning. *Biol. Conserv.* 77: 227–234.

Ruban D.A. 2010. Quantification of geodiversity and its loss. *P. Geologist. Assoc.* 121: 326–333.

Ruddock K., August P.V., Damon C., LaBash C., Rubinoff P. & Robadue D. 2013. Conservation in the context of climate change: practical guidelines for land protection at local scales. *PLoS ONE* 8(11): e80874, doi:10.1371/journal.pone.0080874.

Stein A., Gerstner K. & Kreft H. 2014. Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecol. Lett.* 17: 866–880.

Thuiller W., Araujo M.B. & Lavorel S. 2004. Do we need land-cover data to model species distributions in Europe? *J. Biogeogr.* 31: 353–361.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* 58: 267–288.

Titeux N., Maes D., Marmion M., Luoto M. & Heikkinen R.K. 2009. Inclusion of soil data improves the performance of bioclimatic envelope models for insect species distributions in temperate Europe. *J. Biogeogr.* 36: 1459–1473.

Turner W., Spector S., Gardiner N., Fladeland M., Sterling E. & Steiniger M. 2003. Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* 18: 306–314.

Ulbricht J. 2012. *lqa: penalized likelihood inference for GLMs*. R package version 1.0-3, available at http://CRAN.R-project.org/package=lqa.

Wania A., Kühn I. & Klotz S. 2006. Plant richness patterns in agricultural and urban landscapes in Central Germany — spatial gradients of species richness. *Landscape Urban Plan*. 75: 97–110.

Whittaker R.J., Willis K.J. & Field R. 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. *J. Biogeogr*. 28: 453–470.

Wood S.N. 2006. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, Boca Raton, FL.

Wood S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. Roy. Stat. Soc. B* 73: 3–36.

Zuur A.F., Ieno E.N., Walker N.J., Saveliev A.A. & Smith G.M. 2009. *Mixed effects models and extensions in ecology with R*. Springer, New York.

Appendix 1

Overview of the used habitat type classification (HTC) approach

Data sets and calculated layers and features

In the habitat type classification (HTC), we used the following GIS and remote sensing data: two sets of aerial imagery; an airborne laser scanner data from the National Land Survey of Finland; 1:20 000 resolution digital Quaternary deposit (hereafter soil) and 1:200 000 resolution digital bedrock maps from the Geological Survey of Finland; forestry planning polygons and polygons of Forest Act habitats from the Finnish Forest Centre Pirkanmaa from years 2000–2010; as well as a 1:10 000 resolution topographic database and a 1:50 000 resolution SLICES land-use database from the year 2010 from the National Land Survey of Finland.

The first set of aerial imagery was taken in the summer of 2011 by TerraTec Oy for the Finnish Forest Centre Pirkanmaa. It consisted of three bands: green, red and infra-red in 40 cm spatial resolution. The other aerial image set was taken by the National Land Survey of Finland during springs 2010, 2011 and 2012 using an Intergraph DMC camera. The spatial resolution of the data was 50 cm and it consisted of four bands: blue 400–580 nm, green 500–650 nm, red 590–675 nm, and near infrared 675–850 nm.

The airborne laser scanner data were collected in the springs of 2008 and 2012 by the National Land Survey of Finland. Data contained at least 0.5 points per 1 m^2 and the flying altitude was on average 2000 meters. The used scan angle was ±20° and the laser pulse footprint in terrain approximately 50 cm. The mean error in the elevation information is at maximum 15 centimeters and in the planar information at maximum 60 cm. The data was delivered as point clouds with automatic classification to ground hits, low vegetation hits, low error hits and unclassified hits.

From the airborne laser scanner data, we constructed two primary layers. A digital terrain model was constructed by triangulating the points classified as ground. A digital surface model was constructed by triangulating the first hits only. Moreover, we excluded the hits classified as unclassified or low points from the analysis. Before the triangulation, we thinned the surface to 1-m^2 resolution. We constructed a canopy height model by subtracting the digital terrain model from the digital surface model. To eliminate unrealistic values, we further manipulated the canopy height model to include values only between zero and 40 m. We processed the airborne laser scanner data using LAStools (rapidlasso, Gilching, Germany).

We did not use the digital terrain model in the analysis as such, but we derived five layers from it. SAGA wetness index models moisture conditions using local and neighborhood slope and upslope contributing area (Böhner and Selige 2006). It is a modification of standard topographic wetness index in which only local slope is taken into account. We quantified SAGA wetness index, because the topographic wetness index is known to underestimate the extent and the contiguity of wetlands (Grabs *et al*. 2009, Murphy *et al*. 2009). Terrain ruggedness index quantifies the amount of elevation

difference locally (Riley *et al*. 1999), topographic position index measures the relative altitudinal position of a pixel (Guisan *et al*. 1999) and multiresolution index for valley bottom flatness identifies the areas that are relatively low or flat (Gallant and Dowling 2003). In a distance to water layer, a slope raster was used as a cost surface and, from each pixel, a cost distance to a stream or a water body was calculated (Murphy *et al*. 2007, 2009). We calculated the terrain ruggedness index using 3 × 3-pixel window size, the topographic position index with the radius of 100 m, and the multiresolution index for valley bottom flatness using the value of 28 for initial threshold for slope as suggested by Gallant and Dowling (2003). In calculating the distance to water layer, we modeled streams using a $D\infty$ flow direction and a 40 000 m^2 threshold value using TauDEM tools (ver. 5.0, http://hydrology.usu.edu/taudem/taudem5/index.html). We calculated the distance to water layer using ArcGIS (version 10.1, Esri, Redlands, CA, USA) and the other topographic layers using SAGA-GIS (ver. 2.0.8, http://www.saga-gis.org/).

Classification of habitat types

Our habitat type classification workflow is a simplified version of the analysis by Räsänen *et al*. (2014). In this study, we modified the classification workflow to match with the available local data sets and to classify all habitat types used in Rossi and Kuitunen (1996). We classified 22 habitat types using object-based image analysis (OBIA) and ancillary data (Table A1).

In the first phase, we used OBIA methodology in mapping three forest habitat types. We performed Fractal Net Evolution Approach segmentation in eCognition Developer 8.8 software (Trimble, Sunnyvale, CA, USA) using a scale parameter value 10 together with a parameter value 0.5 both to color and to compactness. In segmentation, we used all aerial image bands together with the airborne laser scanner-based canopy height model and SAGA wetness index layers in 10-meter reso-

Table A1. Different habitat types mapped and different approaches or data sets used in mapping them.

Habitat type	Approach/data set
Herb-rich and other deciduous forests	OBIA
Esker forests	OBIA, soil map
Dry upland forest sites	OBIA
Moist upland forest sites	OBIA
Rich fen	Forest Act habitats polygons
Open mires	NLS topographic database
Pine mires	NLS topographic database
Spruce mires	NLS topographic database
Oligotrophic lakes	NLS topographic database
Eutrophic lakes	NLS topographic database
Streams and rivers	NLS topographic database
Springs	NLS topographic database, Forest Act habitats polygons
Riparian habitats	NLS topographic database, Forest Act habitats polygons
Flooded areas	NLS topographic database, Forest Act habitats polygons
Beaches	NLS topographic database, Soil map
Non-calcareous rocky areas	NLS topographic database
Calcareous rocks and quarries	NLS topographic database, Bedrock map
Dry meadows	NLS topographic database, Forest Act habitats polygons, Forestry planning polygons
Wet meadows	NLS topographic database
Cultivated areas	NLS topographic database
Parks and gardens	NLS topographic database, SLICES land use database
Industrial and urban areas	NLS topographic database, SLICES land use database

lution. We resampled the aerial image bands to ten meter resolution by calculating mean values. We gave all layers an equal weight.

For all segments, we calculated 122 feature values based on the aerial imagery and the airborne laser scanner data. For all 13 layers, we calculated mean values and standard deviations per segment. For all aerial imagery bands and the canopy height model, we calculated 12 Gray-Level Co-occurrence Matrix (GLCM) and Grey-Level Difference Vector (GLDV) texture features proposed by Haralick *et al*. (1973) using eCognition Developer 8.8. We calculated the features to all directions using 8 bit quantization and they were the following: GLCM homogeneity, contrast, dissimilarity, entropy, angular 2nd moment, mean, standard deviation, and correlation as well as GLDV angular 2nd moment, entropy, mean, and contrast.

Of the 98 196 segments, whose size range was 100–19 000 m^2, we used 3790 as training data in a random forest classifier (Breiman 2001) which was trained using the package randomForest (Liaw and Wiener 2002) in R (ver. 2.15.2, http://www.R-project.org/.). In random forest, a majority vote over several bootstrapped classification trees is taken. When a tree is built, approximately 1/3 of the data is left out of the bootstrap sample and is called out of bag (OOB) data. The OOB data are used for error rate estimation, which is averaged over all trees. Because of the OOB, independent test data or cross-validation is not needed when random forest is used (Breiman 2001, Breiman and Cutler 2007).

For the training data segments, we obtained habitat types from the corresponding forestry planning polygons. The forestry planning data consisted of 4227 polygons, a total of 57.8 km^2, and had information, for instance, about the habitat type and tree stand. We classified the data into three habitat types based on habitat type and tree species in the data set. Initially, we divided all habitat types into four successional stages based on stand development class information. In this study, however, we considered all successional stages of each habitat type as one class. Because of the recent open regeneration areas, we manually modified the habitat type of some forestry planning polygons and deleted some of the polygons altogether to match the aerial images. In total, we used 49.3 km^2 of the data. We used as training data all those segments that had at least a 60% share of area inside one habitat type based on the reference polygons.

After the OBIA classification, we classified forested peatlands of the National Land Survey topographic database into spruce and pine mires based on predicted forest habitat type. We classified all segments, whose majority soil type was esker deposit, as esker habitats. We updated the rest of the habitat types to the HTC straight according to ancillary layers (Table A1). Within these habitat types, we made the following adjustments. We classified a lake as eutrophic if inside a 100-m buffer around the lake over 50% of land-use was cultivated areas or meadows. We mapped riparian areas using 15-m buffers for lakes and streams as well as a 5-meter buffer for small streams, brooks, creeks, and ditches. We did not map 5-meter buffers to peatland areas. For 15-m lakeside buffers, we classified all areas that were on a mineral soil as beaches. For springs, we used a 5-meter buffer. We classified a rocky area as calcareous, if the bedrock type was calcareous or mafic or intermediate based on the classification by Kalliola (1973). We classified a meadow as dry meadow, if it was on mineral soil (excluding clay) and if its mean SAGA wetness index value was smaller than 15. When all habitat types had been classified, we converted the vector data set into a 10-meter resolution raster data set.

Classification accuracy analysis

The classification accuracy of the HTC was calculated using forestry planning polygons as a reference with a simple pixel-based cross-tabulation matrix. All area that was mapped as forests in the classification as well as in the reference was used in the classification accuracy calculation. Additionally, an OOB error rate of the random forest classifier was calculated on a segment level.

The classification accuracy of the HTC was 47% in the areas classified as forest habitat types if different successional stages were considered separate classes. A rather similar result was given by

random forest OOB error rate which was 49%. When only three different forest habitat types were classified, classification accuracy was 60%.

The major reason for the low classification accuracy is probably noisy training data. The information about habitat type and boundaries might not always be accurate in forestry planning data sets, since they are not principally intended to be used in this kind of task. Another reason is that we measured only the classification accuracy in the classification of forest habitat types. Forest habitat types can be difficult to interpret also by skilled professionals in the forest and their differences in remotely sensed data are often small.

Appendix 2

List of different land-use/land-cover types in Corine Land Cover classification

Code	CLC type
1110	Continuous urban fabric
1120	Discontinuous urban fabric
1210	Industrial or commercial units
1220	Road and rail networks and associated land
1310	Mineral extraction sites
1320	Dump sites
1421	Summer cottages
1422	Other sport and leisure facilities
2111	Non-irrigated arable land, in use
2112	Non-irrigated arable land, abandoned
2220	Fruit trees and berry plantations
2310	Pastures
3111	Broad-leaved forest on mineral soil
3112	Broad-leaved forest on peatland
3121	Coniferous forest on mineral soil
3122	Coniferous forest on peatland
3123	Coniferous forest on rock exposure
3131	Mixed forest on mineral soil
3132	Mixed forest on peatland
3133	Mixed forest on rock exposure
3241	Transitional woodland/shrub, canopy cover < 10%
3242	Transitional woodland/shrub, canopy cover 10%–30%, on mineral soil
3243	Transitional woodland/shrub, canopy cover 10%–30%, on peatland
3244	Transitional woodland/shrub, canopy cover 10%–30%, on rock exposure
3247	Transitional woodland/shrub, abandoned agricultural land
3320	Bare rock
4111	Inland marshes, on land
4112	Inland marshes, on water
4121	Peatbogs
5110	Water courses
5120	Waterbodies