

Statistical mapping of tree species over Europe

D. J. Brus · G. M. Hengeveld · D. J. J. Walvoort ·
P. W. Goedhart · A. H. Heidema · G. J. Nabuurs ·
K. Gunia

Received: 25 November 2010 / Revised: 14 February 2011 / Accepted: 29 March 2011
© Springer-Verlag 2011

Abstract In order to map the spatial distribution of twenty tree species groups over Europe at 1 km × 1 km resolution, the ICP-Forest Level-I plot data were extended with the National Forest Inventory (NFI) plot data of eighteen countries. The NFI grids have a much smaller spacing than the ICP grid. In areas with NFI plot data, the proportions of the land area covered by the tree species were mapped by compositional kriging. Outside these areas, these proportions were mapped with a multinomial multiple logistic regression model. A soil map, a biogeographical map and bioindicators derived from temperature and precipitation data were used as predictors. Both methods ensure that the predicted proportions are in the interval [0,1] and sum to 1. The regression predictions were iteratively scaled to the National Forest Inventory statistics

and the Forest map of Europe. The predicted proportions for the twenty tree species were validated by the Bhattacharyya distance between predicted and observed proportions at 230 plot data separated from the calibration data. Besides, the map with the predicted dominant species was validated by computing the error matrix. The median Bhattacharyya distance in the subarea with NFI plot data was 1.712, whereas in the subarea with ICP-Level-I data, this was 2.131. The scaling did not significantly decrease the Bhattacharyya distance. The estimated overall accuracy of this map was 43%. In areas with NFI plot data, overall accuracy was 57%, outside these areas 33%. This gain was mainly attributable to the much denser plot data, less to the prediction method.

This article originates from the context of the EFORWOOD final conference, 23–24 September 2009, Uppsala, Sweden. EFORWOOD—Sustainability Impact Assessment of Forestry-wood Chains. The project was supported by the European Commission.

Communicated by J. Müller.

Electronic supplementary material The online version of this article (doi:10.1007/s10342-011-0513-5) contains supplementary material, which is available to authorized users.

D. J. Brus (✉) · G. M. Hengeveld · D. J. J. Walvoort ·
A. H. Heidema · G. J. Nabuurs
Alterra, Wageningen University and Research Centre,
P.O. Box 47, 6700 AA Wageningen, The Netherlands
e-mail: dick.brus@wur.nl

P. W. Goedhart
Biometris, Wageningen University and Research Centre,
P.O. Box 100, 6700 AC Wageningen, The Netherlands

G. J. Nabuurs · K. Gunia
European Forest Institute, Torikatu 34, 80100 Joensuu, Finland

Keywords Logistic regression · Kriging · Map validation · Bhattacharyya distance · Confusion matrix · Overall accuracy

Introduction

Forests cover over 35% of Europe (MCPFE Liaison Unit Warsaw, UNECE, FAO 2007; Schuck et al. 2003). These forests provide important goods and services to the human society. Next to the economic potential in raw material, e.g. for construction and energy, forests provide many other services to society. These services range from the sequestration of CO₂ to the protection of watersheds from erosion, biodiversity conservation and the provision of recreational area (UN-ECE 2005; EEA 2007; Nabuurs et al. 2007; FAO 2007). For production of, and policies targeting these goods and services at national and EU level, the analyses of the state and future development of these forests used to be done at rather high aggregation levels.

These were often statistics or forest resource models down to the national level or Geographical Information System (GIS)-based tools down to level 3 of the Nomenclature of Territorial Units for Statistics (NUTS) (Schelhaas et al. 2007; Verkerk et al. 2011; EEA 2006; Kuhlmann et al. 2008). Furthermore, a large variation of forest models exists that are suitable for detailed and high-resolution analyses, but that can only be run for small areas and that cannot account for the whole European continent (Hasenauer 2006).

The other analysis area relevant for European forests that operates at the European scale is the climate change field, where biophysical models for the terrestrial biospheric carbon balance of Europe run at the resolution for which soil or climate data are available. This is often at a 10×10 km resolution. A finer resolution in these methods is often hampered by a lack of high-resolution data for the forest vegetation characteristics like species, growing stock, increment and age class.

With modern computation and GIS techniques and the increasing data availability, such as the National Forest Inventory (NFI) plot data, the possibility of high-resolution scenario modelling of areas of large extent is coming within reach. This will not only enable the compatibility of the climate change models and the forest resource models at the European scale. But it will lead to highly improved forest resource modelling, being able to deal with the high-resolution forest diversity in Europe. In this way, analyses of goods and services, and, e.g. policy effectiveness, can be monitored much better. As a first step towards this high-resolution forest resource modelling, detailed and harmonized tree species maps over Europe are needed. These maps are the basis for further high-resolution GIS material on specific variables like growing stock and age class.

The distribution of tree species and forest types is to a large extent constrained by abiotic conditions as soil, precipitation and temperature (Pearman et al. 2008; Bohn et al. 2000). This results in a correlation between abiotic conditions and biotic response that is exploited in the bioclimatic envelope approach (Pearson and Dawson 2003). Within Europe, however, 95% of the forests are managed by humans (MCPFE Liaison Unit Warsaw, UNECE, FAO 2007). Management has a strong influence on the distribution and dominance of species, leading to deviations from the 'natural' forest composition (Pearson and Dawson 2003). This effect of management is reflected in data on occurrence of species. Deviations from the natural situation can be either large scale (e.g. selected species are overrepresented within a certain region) or small scale (e.g. species show more spatial segregation in heavily managed areas). Conversely, management is also to a certain extent limited by the abiotic conditions that confine species distributions.

Recently, pan-European maps for six main tree species groups have been published by Tröltzsch et al. (2009). These maps were constructed by ordinary kriging of the ICP-Forests-Level-I plot data. The maps were then scaled so that they correspond both to the Forest map of Europe (Schuck et al. 2003) on pixel level and to national forest inventory statistics at regional or national levels. Other attempts have been more regional, e.g. Sykes et al. (1996), Thuiller et al. (2003) and Hidalgo et al. (2008).

In the above-mentioned previous studies, the wealth of data in the NFI plots is not used, or the approach is based on bioclimatic envelopes only. Here, we take this a step forward in using NFI plot data, ICP-level-I data and advanced statistical methods with biophysical GIS layers. The aim is to derive 1×1 km resolution tree species map of Europe.

This paper describes in detail the data and information we used in the mapping, the statistical methods used for mapping the tree species and how we validated the results. We present a selection of the maps (all maps are available online) and the validation results. Finally, we discuss the results and draw some conclusions.

Data

The data used in this study consist of forest plot data collected in a European-wide network and National Forest Inventory forest plot data on tree species of eighteen countries. Besides, forest inventory statistics have been used. These statistics are based on forest plot data that are aggregated over regions in the Nomenclature of Territorial Units for Statistics (NUTS). Finally, ancillary information has been used on covariates related to the occurrence of tree species, such as climatic variables and soil type.

Plot data on tree species

We used the ICP-Level-I plot data, covering most of Europe (www.icp-forests.org). The total number of ICP-Level-I plots with tree species counts equals 6,238 (Fig. 1). Apart from this data set, we used the National Forest Inventory (NFI) data of eighteen countries (Table 1; Fig. 1). For some countries, the NFI data cover only part of the country, e.g. France, Spain and Italy. The total number of plots equals 3,35,360. Reported proportions per species represent the share in the forest based on basal area, growing stock or stem number. For Spain, Great Britain and Slovakia, only species presence was reported. In this case, equal proportions were assumed.

More than 200 tree species were reported in the two databases. These species were grouped into eighteen species groups that cover most of the European species, and

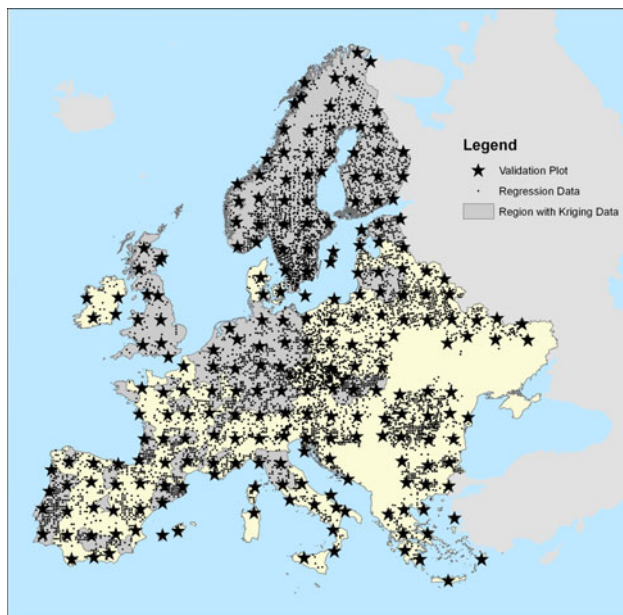


Fig. 1 Location of ICP-Level-I plots and regions with NFI plot data (shaded). The map also shows the location of the validation plots

two miscellaneous groups being ‘other conifers’ and ‘other broadleaved species’ (Table 2).

The forest plot data were split into a set for calibration and a set for validation. This was done by overlaying the combined ICP and NFI data sets with a square grid with a

spacing of 180 km. The plot closest to the centre of each grid cell entered the validation set (Fig. 1). This selection procedure guarantees a fair spatial coverage of Europe and avoids overrepresentation of intensively sampled areas. The validation set consists of 230 plots (216 ICP plots, 14 NFI plots).

Forest inventory statistics

Regional traditions and goals in forest management are not well reflected in the regression of tree species to abiotic covariates. In order to incorporate these effects in the predicted tree species distributions, we used forest inventory statistics on proportions of land areas covered by tree species in forest and woodland area at spatial scales varying from NUTS-0 to NUTS-3. We used the European Forest Information Scenario Model (EFISCEN) database for this (Schelhaas et al. 2003), extended with statistics from:

- Serbia, Montenegro and Ukraine, from the country reports of the forest resource assessment 2005, www.fao.org/forestry/site/32179/en/ (accessed on 31-01-2008))
- Austria (Austrian Forest Inventory 2000–2002)
- Finnmark (Norwegian Forest Inventory)
- Moldova (ECE/FAO, 1997. Forest and Forest products country profile: Republic of Moldova)

Table 1 National Forest Inventory data used in this study; third column indicates the grouping of countries for variogram estimation and kriging (MLR: NFI plot data used in multinomial logistic

regression); columns 4–9: number of plots with one reported species (1), two reported species (2) et cetera; *n*: total number of plots

No	Country	Grouping	1	2	3	4	5	>5	<i>n</i>
31	Netherlands	1	3,156	0	0	0	0	0	3,156
32	Belgium	1	5,788	1,855	711	248	60	12	8,674
33	France	2	208	239	275	261	241	720	1,944
34	Spain	3	10,341	3,267	518	37	2	0	14,165
351	Portugal	4	1,35,327	2,827	560	0	0	0	1,38,714
39	Italy	6	13,605	8	4	1	2	1	13,621
	Tuscany	7	20,987	30,150	15,804	0	0	0	66,941
	Veneto region	MLR							
40	Romania	MLR							
44	United Kingdom	9	6,174	5,689	3,996	2,333	1,173	966	20,331
46	Sweden	10	2,014	1,148	283	40	9	1	3,495
47	Norway	10	8,365	226	40	0	0	0	8,631
49	Germany	11	21,454	16,665	8,554	2,925	715	148	50,461
358	Finland	10	1,464	469	154	18	4	0	2,109
370	Lithuania	5	474	299	24	15	0	3	815
372	Estonia	5	196	277	422	405	257	121	1,678
380	Ukraine	MLR							
385	Croatia	MLR							
386	Slovenia	MLR							
421	Slovakia	8	65	178	143	132	72	36	626

Table 2 Groups of tree species

No	Species group
1	<i>Abies</i> spp.
2	<i>Alnus</i> spp.
3	<i>Betula</i> spp.
4	<i>Carpinus</i> spp.
5	<i>Castanea</i> spp.
6	<i>Eucalyptus</i> spp.
7	<i>Fagus</i> spp.
8	<i>Fraxinus</i> spp.
9	<i>Larix</i> spp.
10	Other broadleaved
11	Other conifers
12	<i>Pinus</i> spp. (not included in groups 15 and 16)
13	<i>Quercus</i> spp. (not included in group 19)
14	<i>Picea</i> spp.
15	<i>Pinus pinaster</i>
16	<i>Pinus sylvestris</i>
17	<i>Populus</i> spp.
18	<i>Pseudotsuga</i> spp.
19	<i>Quercus robur</i> and <i>Quercus petraea</i>
20	<i>Robinia</i> spp.

- Greece (National Inventory of Greece, 1992. General Secretariat of Forests and Natural Environment. Ministry of Agriculture, Independent Edition)

Covariate maps

Nineteen quantitative, full-coverage covariates were used as candidate predictors in regression (Table 3). The first fifteen covariates are bioindicators derived from monthly temperature and precipitation data in the WORLDCLIM database with a spatial resolution of 30 arc seconds (Hijmans et al. 2005). Elevation and slope are terrain attributes stored in GTOPO30, a global digital elevation model of the U.S. Geological Survey with a spatial resolution of 30 arc seconds (approx. 1 km) (http://eros.usgs.gov/#Find_Data/Products_and_Data_Available/GTOPO30).

Besides, two categorical variables were used as candidate predictors, biogeographical region and soil class. For biogeographical region, we used the biogeographical regions data set of the European Environment Agency (Roekaerts 2002). The map polygons were aggregated into four major regions (see online supplementary material). For soil class, we used the FAO/Unesco Soil map of the World 1:50,000,000 (FAO 1992). The soil map units in the study area were grouped into 13 soil groups on the basis of soil properties relevant for tree growth (see online supplementary material).

Mapping methods

Two entirely different statistical methods were used for mapping the tree species (Fig. 2). In areas with NFI plot data, the maps were constructed by spatial interpolation (kriging) of the plot data, whereas outside these areas, a regression model calibrated on the ICP-Level-I plot data was used for mapping. This means that in areas with NFI plot data, we build on the spatial autocorrelation of the proportion of a tree species, whereas outside these areas, the relation between the proportion of a tree species and covariates mapped at the same plot is exploited. In areas with NFI plot data, the spatial density of plot data is very high, and no gain in precision is expected by taking covariates into account. On the other hand, in areas with ICP-Level-I plot data only, the density is that low that we expect no profit from spatial autocorrelation, and we entirely rely on statistical relations between the proportion of a tree species and the covariates. With both methods, the proportions of the twenty tree species are predicted for forest plots at the nodes of a 1 km × 1 km grid. Hereafter, we will elaborate on the implementation of both statistical methods.

Both the regression and the kriging predictions represent (expected) proportions of *the forested area*. In order to scale the results of the regression model to statistics of national and regional forest inventories and to the proportions of coniferous and broadleaved forest as depicted on the Forest map of Europe, the predicted proportions were multiplied by the proportion of land area covered by forest as depicted in the Forest map of Europe (Schuck et al. 2003). This results in predicted species proportions of *land area*. For the scaling, we used the procedure of Tröltzsch et al. (2009), see also Päivinen et al. (2009). We slightly adapted this procedure to allow for species hierarchy. This was necessary to account for differences in tree species resolution between the reported statistics and the predicted distributions (legend of the tree species maps). E.g. if within a reporting unit, all pinus species are grouped as *Pinus* spp., then the predicted land areas for *Pinus pinaster* and *Pinus sylvestris* were summed in scaling. And if within a reporting unit no account is given for *Abies*, then the predicted land area for this tree species was summed with the land areas of all other broadleaved in scaling. The iteration of scaling to the regional statistics and the Forest map of Europe is repeated until either all scaling factors were within 0.0001 from unity, or the largest change in scaling factors between iterations was less than 0.001, or 25 iterations were performed. This scaling was not applied to the compositional kriging results within areas with NFI plot data because the NFI statistics are based on the NFI plot data.

Table 3 Quantitative covariates used as candidate predictors in multinomial multiple logistic regression; last two columns show the mean and standard deviation (SD) in the data set used for calibration of the multinomial multiple logistic regression model

Covariate	Acronym	Mean	SD
Annual mean temperature	AnnTemp	73.1	35.4
Mean diurnal range (mean of monthly Tmax _ Tmin)	MeanDiur	85.2	12.6
Isothermality (MeanDiur/TempRange _100)	Isotherm	29.6	4.94
Temperature seasonality (Standard Deviation *100)	TempSeas	7,120	1,254
Max. temperature of warmest month	MaxTWarm	226	34.0
Min. temperature of coldest month	MinTCold	-60.8	50.2
Mean temperature of wettest quarter	TempWet	131	43.3
Mean temperature of driest quarter	TempDry	24.6	82.4
Annual precipitation	AnnPrec	751	224
Precipitation of wettest month	PrecWetM	94.4	26.9
Precipitation of driest month	PrecDryM	37.3	15.5
Precipitation seasonality (coe_cient of variation)	PrecSeas	29.9	10.1
Precipitation of wettest quarter	PrecWetQ	262	76.8
Precipitation of warmest quarter	PrecWarmQ	226	78.2
Precipitation of coldest quarter	PrecColdQ	162	79.0
Elevation of plot	Elevation	457	423
Slope of plot	Slope	0.737	0.853
Easting	Xetrs	46,03,644	7,07,837
Northing	Yetrs	31,47,230	7,96,711

Besides tree species maps with the predicted tree species proportions, we also constructed a map with the predicted dominant tree species, i.e. the tree species with the largest predicted proportion.

Multinomial multiple logistic regression

Multinomial logistic regression is a direct extension of ordinary logistic regression, which itself is a special case of a generalized linear model. When there are k species, there are also k counts y_1, \dots, y_k with corresponding probabilities of occurrence π_1, \dots, π_k . The counts then follow a multinomial distribution, i.e. $(y_1, \dots, y_k) \sim \text{Multinomial}(N; (\pi_1, \dots, \pi_k))$, with N the sum of the counts, i.e. the total number of trees in a plot. Analogous to ordinary logistic regression, the natural logs of the quotients $\pi_1/\pi_k, \pi_2/\pi_k, \dots, \pi_{k-1}/\pi_k$ are modelled as a linear combination of predictors (covariates). Note that this implies that there are $(k - 1)$ linear predictors.

The model ensures that all probabilities are in the interval $[0,1]$ and that the probabilities add up to 1. Further information about multinomial logistic regression can be found in McCullagh and Nelder (1989) and in Hosmer and Lemeshow (2000).

For calibrating the regression model, the NFI plots from Ukraine, Romania, Slovenia, Croatia and the Veneto Region in Italy (800 plots in total) were added to the ICP-Level-I plot data. The density of NFI plots in these countries and region is low, resulting in unsatisfactory results from compositional kriging. Addition of these NFI plots

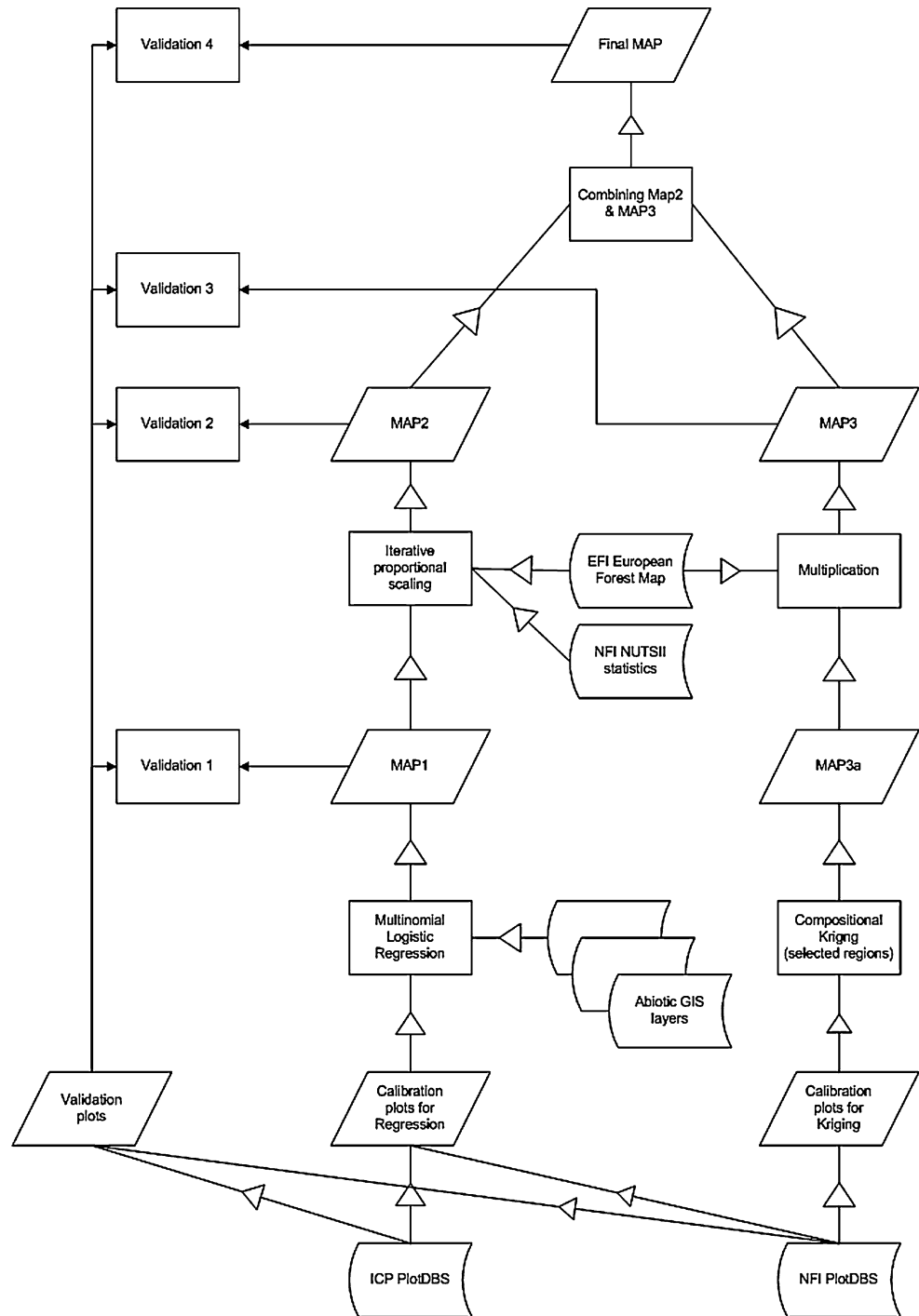
also improved European spatial coverage, especially for trees that are very rare in the ICP data set. This resulted into a calibration data set of 6,822 plots (6022 ICP plus 800 NFI) with the proportions of 20 different tree species. For the 800 NFI plots, with proportions of species only, the total number of trees is not available. The somewhat arbitrary choice was made to analyse the percentages rather than the counts, with a binomial denominator of 100 for all plots. In this way, all plots are weighted equally in the statistical analysis. Since 100 is not the real binomial denominator, an overdispersion factor was added to the multinomial variance and then quasi-likelihood was used, rather than maximum likelihood (McCullagh and Nelder 1989). We used GenStat (Payne et al. 2007) to fit the regression model.

Table 4 shows the number of plots in which a given species does not occur (0%) and constitutes 0–10% of the trees, 10–20%, *et cetera*. Note that percentage 100 denotes the number of plots in which that species is the only species present. Summing the category 100% over the species gives 2,840 plots with a single species present. Some species, notably *Eucalyptus* spp., ‘other conifers’ and *Pseudotsuga* spp., are hardly present in the data. At the other end of the spectrum, species *Picea* spp. and *Pinus sylvestris* spread most across the plots.

Model selection

At first sight, the data set at hand seems rather large with 6,822 plots. However, most species are quite rare.

Fig. 2 Flow chart of the mapping method



Consequently, using many covariates in a regression model may yield unstable estimators of regression coefficients and thus possibly poor predictions. This is especially the case when covariates are correlated. Some form of selection of covariates is therefore necessary. The aim is then to find a parsimonious model with as few covariates as necessary that is almost 'as good' as a model with many covariates. In general, such a model will produce better predictions. The following line of attack was

chosen. First, the qualitative covariates biogeographical region and soil class are considered to be the most important covariates. This is confirmed by preliminary logistic regressions, separately for each species, which showed that biogeographical region and soil class are significant for almost all species. These covariates are therefore included in every linear predictor. The remaining covariates are classified into the following four groups:

Table 4 Number of calibration plots where a given tree species equals 0%, 0–5%... 100% of the total number of trees

Coverage	<i>Abies</i>	<i>Alnus</i>	<i>Betula</i>	<i>Carpinus</i>	<i>Castanea</i>	<i>Eucalyptus</i>	<i>Fagus</i>	<i>Fraxinus</i>	<i>Larix</i>	Other broadleaved
0	6,263	6,626	5,552	6,450	6,661	6,756	5,610	6,511	6,324	5,857
0–5	117	33	191	63	23	3	185	76	221	226
5–25	162	81	531	169	58	13	324	138	176	375
25–50	139	34	242	77	27	7	209	59	62	182
50–75	65	16	127	45	21	5	131	23	18	125
75–95	27	14	71	15	16	6	135	11	11	33
95–100	4	2	10	1	3	2	70	1	3	4
100	45	16	98	2	13	30	158	3	7	20
Coverage	Other conifer	<i>Pinus</i>	<i>Quercus</i>	<i>Picea</i>	<i>Pinus pinaster</i>	<i>Pinus sylvestris</i>	<i>Populus</i>	<i>Pseudotsuga</i>	<i>Quercus rob.,petr.</i>	<i>Robinia</i>
0	6,725	6,477	6,099	4,010	6,623	4,102	6,435	6,776	5,830	6,682
0–5	31	31	56	144	12	206	90	4	119	12
5–25	34	61	120	408	20	406	157	15	301	50
25–50	12	37	87	423	17	301	70	6	193	25
50–75	7	31	56	496	15	309	25	5	128	15
75–95	3	39	74	527	28	331	10	3	89	12
95–100	–	18	40	167	12	84	1	1	34	5
100	10	128	290	647	95	1,083	34	12	128	21

1. Elevation and slope
2. Temperature related variables
3. Precipitation related variables
4. Coordinates of the plots

A maximum of two covariates from a group are allowed to enter the regression model, and only when a covariate is significant at the 1% level. This largely prevents heavily correlated covariates from being selected, and it ensures that covariates from different coherent groups can be selected. The covariate groups are handled in the order given above such that covariates added in a previous step are not subject to selection again. This considerably limits the number of possible models. Adding a covariate to the model means that the covariate is added to every linear predictor. The slope covariate was log-transformed after adding 1, i.e. $\ln(\text{slope} + 1)$, because it is expected that larger slopes will differ only marginally in their effect on the presence of species. For the other skew distributed covariates, there was no apparent rationale for transformation. To increase numerical precision, the quantitative covariates were standardized so that all covariates had mean zero and variance one.

Compositional kriging

In areas with NFI plot data, tree species maps were constructed by kriging. Whereas in the previous section, the

relation between the probability of occurrence of a species and covariates observed at the same plot is modelled, here the spatial autocorrelation of the forest proportion of a tree species with itself, but measured at different plots, is modelled. No covariates are involved in the modelling.

In ordinary kriging, the forest proportion of a species at an unobserved plot is predicted as a weighted average of the observed proportions in the neighbourhood (Webster and Oliver 2007). The model of spatial autocorrelation is the key in finding the optimal weights, i.e. the weights that lead to predicted proportions that are unbiased and have minimum prediction error variance. Broadly speaking, the smaller the distance between an observed plot and the prediction location, the stronger the autocorrelation of the proportions at these two plots will be, and consequently, the larger the weight attached to this observed plot. For more details, we refer to Webster and Oliver (2007).

We can interpolate the proportions of the twenty species separately by ordinary kriging as described above. However, doing so, there is no guarantee that the predicted proportions sum to 1. Besides, it is not guaranteed that the predicted proportions are in the interval [0,1]. Walvoort and de Gruijter (2001) showed how these two additional constraints can be accounted for in computing the optimal kriging weights.

To account for non-stationarity, we estimated experimental variograms for countries separately. To obtain more reliable variograms, some countries with small numbers of

plots were grouped with a neighbouring country with similar tree species (Table 1). Predictions at the nodes of the 1 km × 1 km grid are based on the nearest observed plots, within a search radius of 25 km from the prediction plot. The minimum number of plots used in interpolation was 1, and the maximum was 12.

Validation

The predicted forest proportions for the twenty tree species were validated by computing the Bhattacharyya distance between the vectors with predicted and observed proportions at the 230 validation plots (Bhattacharyya 1943):

$$D = -\ln\left(\sum_{c=1}^{20} \sqrt{p_c \hat{p}_c}\right)$$

with p_c and \hat{p}_c , the observed and predicted proportion for tree species c , respectively. The argument of the ln function, referred to as the Bhattacharyya coefficient (BC), has a minimum of 0 when there is no match between the two vectors, and maximum 1 with a perfect match. The Bhattacharyya distance has a lower bound of 0 (perfect match) and goes to infinity when BC goes to 0. The map of predicted dominant tree species was validated by computing the error matrix (confusion matrix). From this error matrix, we estimated the overall accuracy (overall purity) and the user's and producer's accuracies (Stehman 1997). The user's accuracy for a given species c is the proportion of land with *predicted* dominant species c that is correctly classified. The producer's accuracy for a given species c is the proportion of land with *observed* dominant species c that is correctly classified.

We expected differences in the quality of predicted proportions and predicted dominant tree species between areas with NFI plot data and areas with ICP-level-I plot data only, mainly due to differences in the plot density. To check this, we also computed Bhattacharyya distances and error matrices for these two subareas separately (see Fig. 2). The total number of validation plots equals 99 for the subarea with NFI data and 131 for the subarea with ICP data only. Finally, in order to quantify the effect of the scaling step on the quality of the regression predictions, we computed quality measures before and after scaling (see Fig. 2).

Results

Multinomial multiple logistic regression

The selected model contains the following 10 covariates: biogeographical region, soil class, elevation, slope, annual

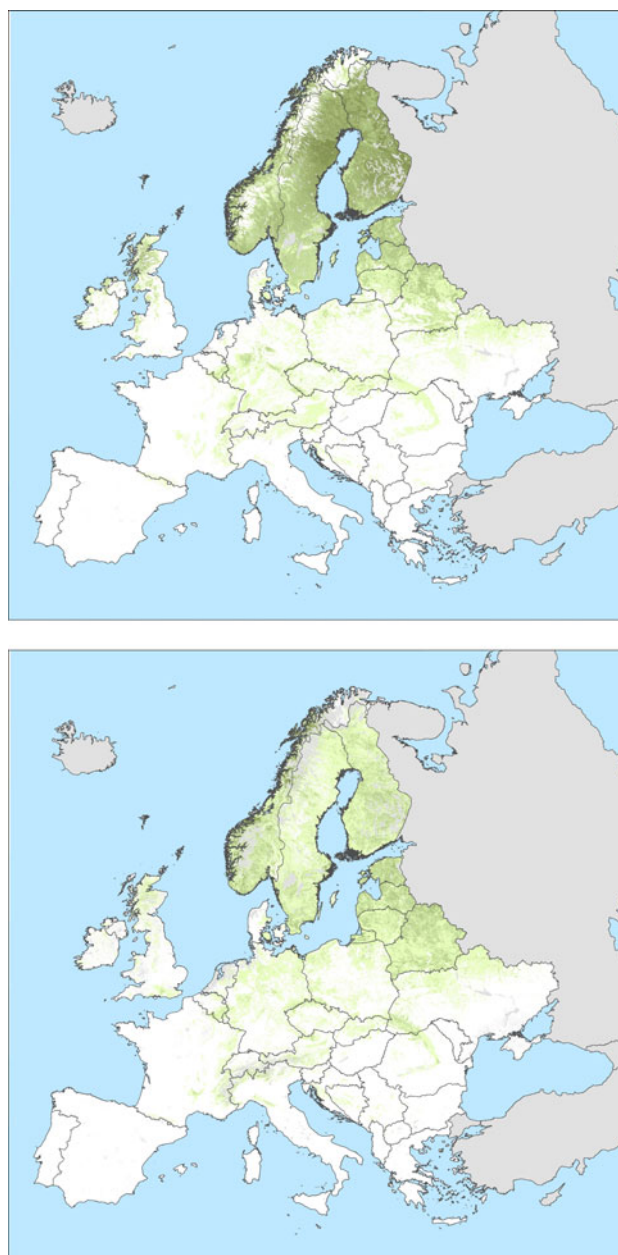


Fig. 3 Predicted proportion of land area covered by *Betula* spp. as obtained by multinomial logistic regression model, before (*upper*) and after (*lower*) scaling to forest inventory statistics; the *darker*, the larger proportions

mean temperature, temperature seasonality, annual precipitation, precipitation of warmest quarter, easting and northing (for the table with the estimated parameters for the quantitative covariates of the selected model, see online supplementary material). None of the estimated parameter values were extremely positive or negative, indicating that there are no major problems with multicollinearity.

Parameter estimates in ordinary regression models have a straightforward interpretation. However, estimates of a multinomial logistic regression model are hard to interpret

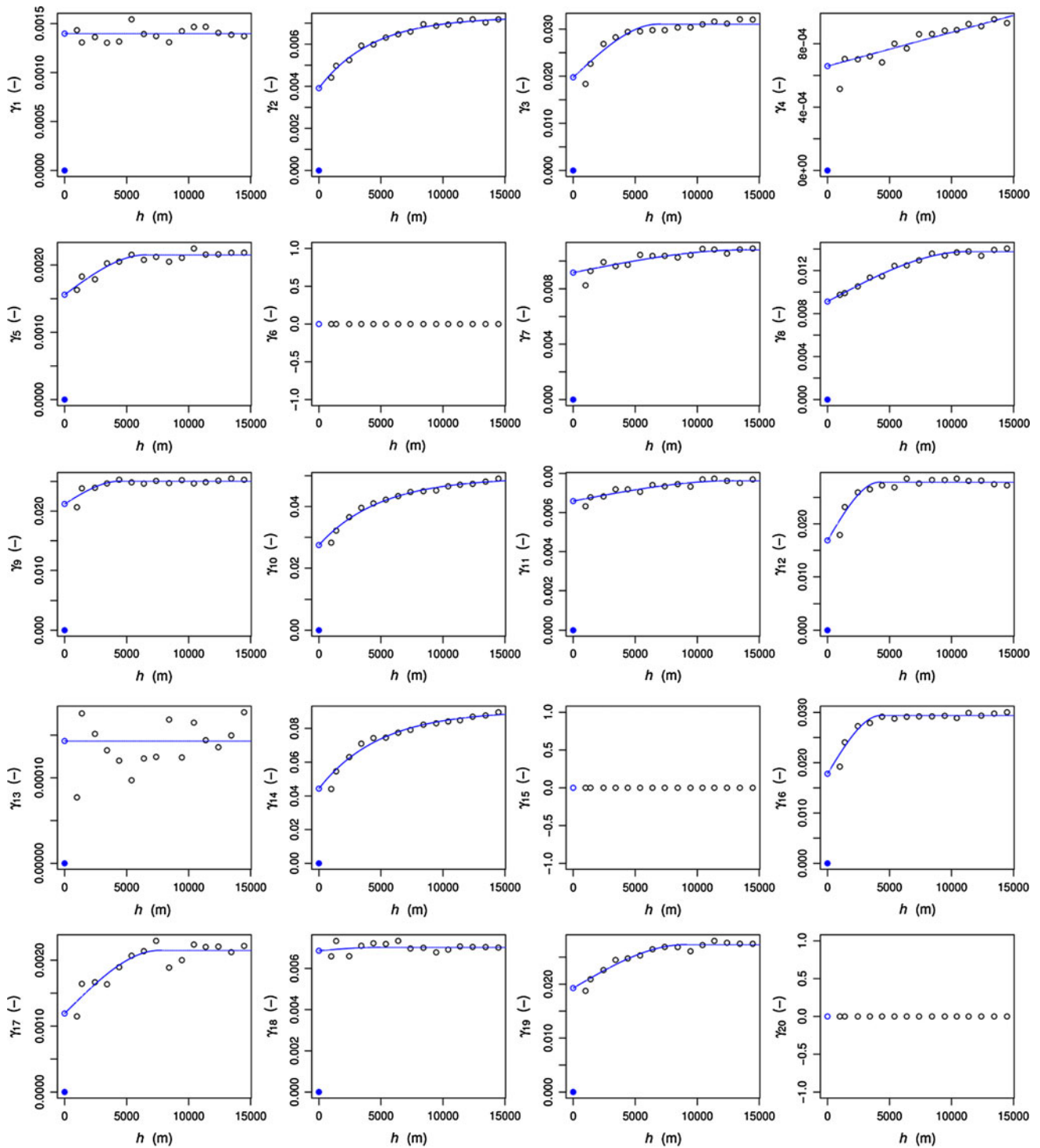


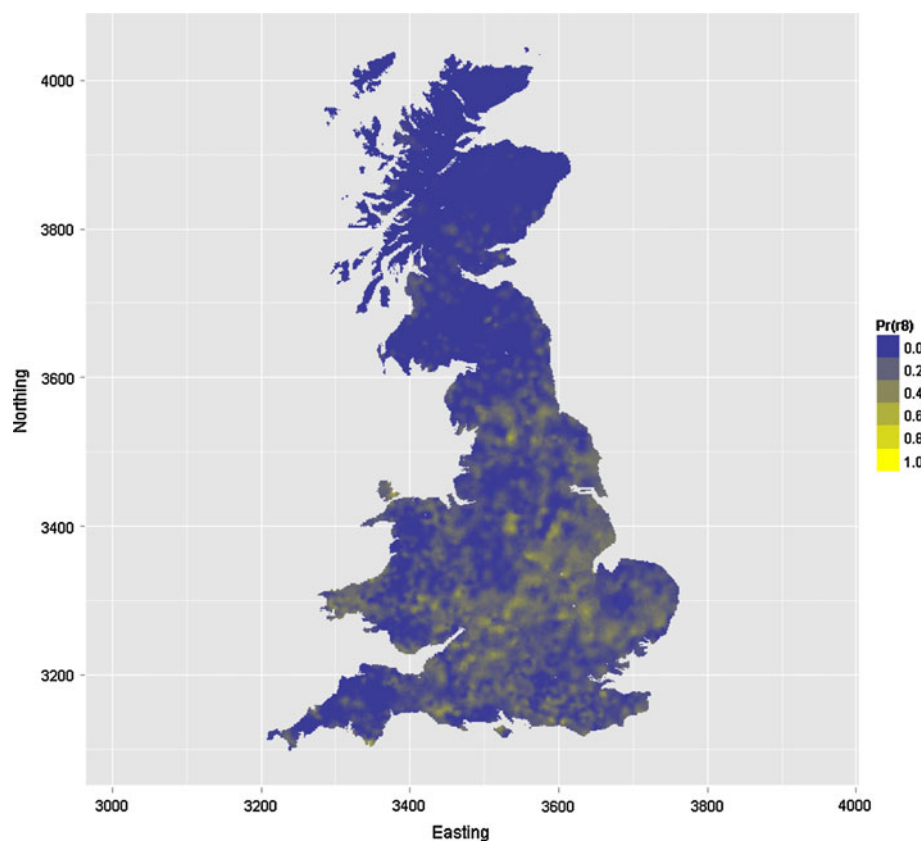
Fig. 4 Experimental and fitted variogram for twenty tree species in United Kingdom; Subscript of γ refers to species number (see Table 2). For *Eucalyptus* spp. (species 6), *Pinus pinaster* (species 15) and *Robinia* spp. (species 20), there were no observations

since the effect of a covariate on the predicted probabilities depends on multiple estimates. Fitted values of the model can be used instead to get a grip on the model.

We compared, for each species separately, the number of plots for which the *fitted* percentage of that species is between 0 and 5, and so on, with the number of plots with

corresponding *observed* percentages. In general, as can be expected from a regression model, the fitted percentages were much smoother than the observed percentages. The fitted values were much more confined to smaller percentages, and for most species, there were hardly any plots where the fitted percentage was above 50%. Notable

Fig. 5 Map of predicted forest proportion for *Fraxinus* spp. (species 8) in United Kingdom (map not yet multiplied by proportion of land covered by forest)



exceptions were species *Picea* spp. (species 14) and *Pinus sylvestris* (species 16), although for these species, the number of plots with a fitted forest coverage of 75% or larger was much smaller than the number of plots with an observed coverage of 75% or larger (for the table, see online supplementary material).

Figure 3 shows the regression predictions of the proportion of land area covered by *Betula*, before and after scaling (for other species, see online supplementary material). As can be seen in Sweden and Finland, the probabilities were lowered considerably by the scaling.

Compositional kriging

As an illustration, Fig. 4 shows the experimental and fitted variogram for the twenty species in the United Kingdom (for variograms of other regions, see online supplementary material). The variograms with $\gamma(h) = 0$ for all h , for species *Eucalyptus* spp. (6), *Pinus pinaster* (15) and *Robinia* spp. (20), indicate that these species are not observed in any forest plot in the UK.

Figure 5 shows the predicted forest proportion for *Fraxinus* spp. in the United Kingdom (for other species and other regions, see online supplementary material). The predicted proportions in Scotland are small compared to England and Wales.

Map of dominant tree species

Figure 6 shows the predicted dominant tree species for Europe as obtained by multinomial logistic regression and compositional kriging. As can be seen in the map, the dominant conifers shift from *Picea* spp. in the North, through *Pinus sylvestris* in Poland and northern Germany and *Picea* spp. again in the alpine region to ‘other pines’ in the Mediterranean. The distribution of broadleaved species has a West-to-East trend with domination of *Quercus robur/petraea* in the West and more *Fagus* spp. and *Alnus* spp. in the East, and *Betula* spp. in Scandinavia and the Baltic states.

Validation

The median Bhattacharyya distance between the vectors with predicted and observed proportions was 2.060 (Table 5). For the subarea with NFI plot data (where predictions were obtained by compositional kriging), this median distance was considerably smaller compared to the subarea with ICP-Level-I data only (with multinomial logistic regression predictions): 1.712 versus 2.131, respectively. Testing the mean over the 99 validation plots in the areas with NFI plot data of pairwise differences of D obtained by kriging and scaled regression showed that

Fig. 6 Map of predicted dominant tree species

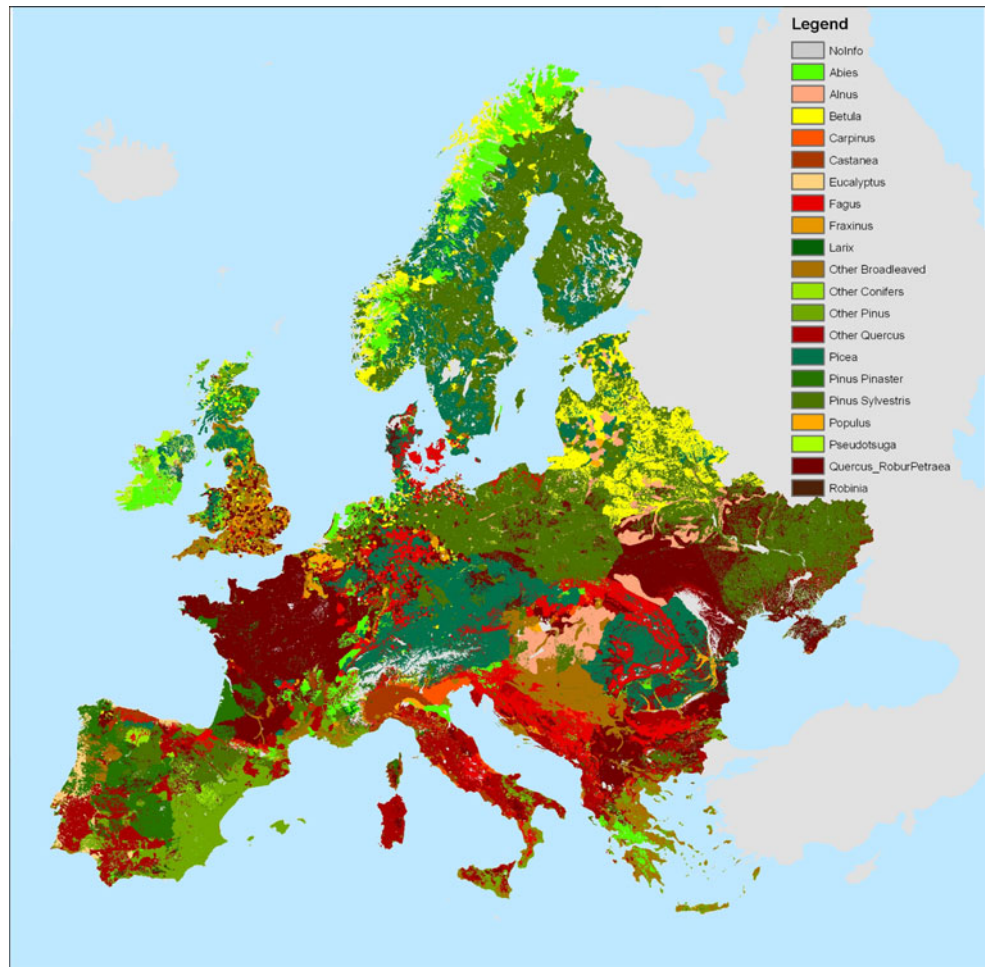


Table 5 Bhattacharyya coefficient (BC), Bhattacharyya distance (*D*) and purity for entire map, for the part of the map with regression predictions from ICP plot data, and for the part of the map with kriging predictions from NFI plot data

Area/map	Prediction method	BC		<i>D</i>		Purity (%)
		Mean	Median	Mean	Median	
ICP	Regression	0.127	0.121	2.253	2.113	31
ICP	Scaled regression	0.129	0.117	2.358 (3)	2.131	33
NFI	Kriging	0.206	0.178	1.810 (1)	1.712	57
NFI	Scaled regression	0.125	0.128	2.312	2.054	52
Entire map	Scaled regression/Kriging	0.136	0.127	2.265 (4)	2.060	43

For the latter part also the quality measures are shown for the scaled regression predictions. Number in parentheses: number of locations with $D = +\infty$, not used in estimating the mean Bhattacharyya distance

the quality of compositional kriging predictions of the proportions was significantly higher than those obtained by multinomial logistic regression (mean difference -0.502 with standard error of 0.060). Testing the mean over all 230 validation plots of pairwise differences of *D* obtained by scaled regression and regression showed that the scaling did not significantly improve the quality of the predicted proportions.

The estimated overall accuracy of the map with dominant tree species equals 43%. For the subarea with NFI plot data, this was considerably larger compared to the subarea with ICP-Level-I data only: 57 and 33%, respectively. For the latter subarea, the scaling slightly improved the result: the overall accuracy before scaling was 31%. The large difference in purity between the part of the map obtained by multinomial logistic regression and the part obtained by

Table 6 User's and producer's accuracies for twenty tree species; a user's accuracy of 1/6 for species 1 means that one plot out of the six plots with predicted dominant species 1 is correctly classified

	Spc 1	Spc 2	Spc 3	Spc 4	Spc 5	Spc 6	Spc 7	Spc 8	Spc 9	Spc 10
User	1/6	0/3	8/19	0/0	0/3	2/2	9/21	0/1	1/3	1/9
Prod	1/2	0/5	8/15	0/0	0/2	2/2	9/22	0/5	1/2	1/11
	Spc 11	Spc 12	Spc 13	Spc 14	Spc 15	Spc 16	Spc 17	Spc 18	Spc 19	Spc 20
User	0/1	10/19	11/20	18/37	4/8	30/56	0/0	0/1	4/20	0/0
Prod	0/0	10/22	11/23	8/37	4/6	30/51	0/3	0/0	4/21	0/1

A producer's accuracy of 1/2 for this species means that one plot out of the two plots with observed dominant species 1 is correctly classified

compositional kriging suggests that kriging outperformed regression substantially. However, in the area with NFI plot data, the difference in purity obtained with these two methods was much smaller (5% only), showing that the gain in purity must be largely contributed to the higher density of plot data rather than the prediction method.

The user's and producer's accuracies of the map varied strongly between the tree species (Table 6). However, for many species, there are only a few plots where this species is observed or predicted as the dominant species, so that the estimates of the producer's and user's accuracies are very imprecise for these species. Focusing on tree species with more than 10 plots in total, the user's accuracy varied from 9% ('other broadleaved') to 59% (*Pinus sylvestris*). The producer's accuracy varied from 20% (*Quercus robur* and *Quercus petraea*) to 54% (*Pinus sylvestris*).

Discussion and conclusions

We have presented a method for constructing a map of tree species cover from various data sources. The main data sources were the ICP plot data and a selection of NFI inventory plot data. These plot data were extended with forest inventory statistics at the country to NUTS-II level, the forest cover map of Europe (Schuck et al. 2003) and a collection of abiotic covariate maps covering the whole of Europe. The strong spatial clustering of the plot data prompted the use of a dual mapping strategy. In areas with a low density of (ICP) plot data, these data were used for calibrating a multinomial multiple logistic regression on abiotic covariates, whereas in areas with a high density of (NFI) plot data, these plot data were spatially interpolated using compositional kriging. The final map was constructed by using the kriged map where it was available and the regression map in other areas.

The dual mapping strategy enables simple and quick updating of the map once NFI plot data of more regions become available. The NFI plot data of a new region can be used for mapping the tree species proportions in that region. This kriged map then replaces the regression map

of that region. Outside this new region, the regression map remains unchanged because the new data are not used in recalibrating the multinomial logistic regression model.

In multinomial logistic regression, there were a lot of competing models with different sets of covariates. The selected model is therefore somewhat arbitrary. The applied model selection strategy, in which covariates were grouped into clusters and selecting two covariates at most from the same cluster, reflects the limited information available. This strategy is fruitful in diminishing the risk of selecting a model with collinear covariates.

In calibrating the multinomial logistic regression model, all plots were equally weighted, as for some plots, the total number of trees in a plot was not available. An alternative would have been to use the forest coverage in percentages as the multinomial denominator, which is equivalent to weighing the plots with the coverage percentage. However, with coverage percentages as low as 10%, this seems hardly justifiable.

In compositional kriging, no covariates are used. It would be interesting to see whether spatial interpolations can be improved by exploiting the abiotic covariates. This can be done by extending the compositional kriging matrix equation with the covariates, leading to universal compositional kriging. An interesting alternative would be prediction by a generalized linear geostatistical model (GLGM), the spatial analogue of a generalized linear mixed model (Diggle and Ribeiro Jr. 2007). In the multinomial logistic regression model, for the tree species described above, the ratios of the probabilities of occurrences are modelled as a linear combination of covariates. This is the fixed effect. In this generalized linear model (GLM), it is assumed that the deviance residuals of the model are spatially uncorrelated. In a GLGM, this GLM is extended with a submodel describing the spatial correlation of the deviance residuals (the random effect).

In order to incorporate regional differences in forest management and species choice that is not reflected in the calibration of the regression model, the results of the regression model were scaled to fit the regional forest inventory statistics. This is done using the iterative

proportional scaling method proposed by Tröltzsch et al. (2009). In areas where regional statistics were not available for the twenty species separately, but at a higher hierarchical level only, scaling factors were necessarily computed at that higher level. This results in a potential misfit, e.g. *Eucalypt* will be hierarchically grouped under ‘other broadleaved’ in Poland because it is not reported separately in the regional inventory statistics of Poland. Potentially, this could lead to a high scaling factor for *Eucalypt* if the ‘other broadleaved’ class is underrepresented in the regression predictions and consequently scaled with a factor >1 . Apparently, the regression model is robust enough to counter this effect, as no major anomalous shifts of species were observed during the scaling process.

Acknowledgments This project was carried out within the framework of European FP5 projects CARBO-Europe IP and EFORWOOD IP. We are greatly indebted to the Forest Focus programme and the National Forest Inventory institute’s correspondents. NFI plot data were received from Jacques Rondeux and Martine Waterinckx, Belgium; Juro Cavlovic, Croatia; Veiko Aderman, Estonia; Kari Korhonen, Finland; Thierry B elouard, France; Heino Polley, Germany; Marino Vignoli, Remo Bertani, Giorgio Dalmasso and Maurizio Teobaldelli, Italy; Andrius Kuliesis, Lithuania; Wim Daamen and Henny Schoonderwoerd, Netherlands; Stein Tomter, Norway; Susanna Barreiro and Margarida Tom e, Portugal; Olivier Bouriaud, Romania; Vladimir Seben, Slovak Republic; Gal Kusar, Slovenia; J. Villanueva and Antoni Trasobar, Spain; G oran Kempe, Sweden; Bill Mason and Shona Cameron, United Kingdom; Igor Buksha, Ukraine. Finally, we like to thank an anonymous reviewer for his expert comments on the statistics, and the suggestion to use Bhattacharyya distance for validation.

References

- Bhattacharyya A (1943) On a measure of divergence between two statistical populations defined by their probability distributions. *Bull Calcutta Math Soc* 35:99–109
- Bohn U, Gollub G, Hettwer C (2000) Map of the natural vegetation of Europe. Federal Agency for Nature Conservation, Bonn
- Diggle PJ, Ribeiro PJ Jr (2007) Model-based geostatistics. Springer series in statistics. Springer, New York
- EEA (2006) European forest types, categories and types for sustainable forest management reporting and policy. EEA Technical report, vol 9/2006
- EEA (2007) Environmentally compatible bio-energy potential from European forests. European Environmental Agency
- FAO (1992) The digitized soil map of the world—notes. World Soil Resources Report, vol 67 (2–7), Release 1.1. Rome
- FAO (2007) State of the world’s forests 2007. Food and Agricultural Organization of the United Nations, Rome
- Hasenauer HL (ed) (2006) Sustainable forest management. Growth models for Europe. Springer, Berlin
- Hidalgo PJ, Mar n JM, Quijada J, Moreira JM (2008) A spatial distribution model of cork oak (*Quercus suber*) in southwestern Spain: a suitable tool for reforestation. *For Ecol Manag* 255:25–34
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978
- Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
- Kuhlmann T, Verhoog D, Verkerk H, Lindner M, Kaae B, Hasler B, Nielsen TS, Caspersen HO, Jansson T (2008) Land use policy scenarios in six target sectors. In: Helming K, Wiggering H (eds) SENSOR report series, vol 2008/3. ZALF, Germany
- McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London
- MCPFE Liaison Unit Warsaw, UNECE, FAO (2007) State of Europe’s forests 2007, the MCPFE report on sustainable forest management in Europe
- Nabuurs GJ, Masera O, Andrasko K, Benitez-Ponce P, Boer R, Dutschke M, Elsiddig E, Ford-Robertson J, Frumhoff P, Karjalainen T, Krankina O, Kurz WA, Matsumoto M, Oyhantcabal W, Ravindranath NH, Sanchez MJS, Zhang X (2007) Forestry. In: Metz B, Davidson OR, Bosch PR, Dave R, Meyer LA (eds) Climate change 2007: mitigation. Contribution of working group III to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA
- P aivinen R, van Brusselen J, Schuck A (2009) The growing stock of European forests using remote sensing and forest inventory data. *Forestry* 82:479–490
- Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2007) GenStat for windows (10th Edition) introduction. VSN International, Hemel Hempstead
- Pearman PB, Randin CF, Broennimann O, Vittoz P, van der Knaap WO, Engler R, Lay GL, Zimmermann NE, Guisan A (2008) Prediction of plant species distributions across six millennia. *Ecol Lett* 11(4):357–369
- Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimatic envelope models useful? *Glob Ecol Biogeogr* 12:361–371
- Roekaerts M (2002) The biogeographical regions map of Europe. Basic principles of its creation and overview of its development
- Schelhaas MJ, Schuck A, Varis S, Zudin S (2003) Database on forest disturbances in Europe (DFDE)—technical description
- Schelhaas MJ, Eggers J, Lindner M, Nabuurs GJ, Pussinen A, P aivinen R, Schuck A, Verkerk PJ, van der Werf DC, Zudin S (2007) Model documentation for the European forest information scenario model (EFISCEN 3.1)
- Schuck A, P aivinen R, H ame T, van Brusselen J, Kennedy P, Folving S (2003) Compilation of a European forest map from Portugal to the Ural mountains based on earth observation data and forest statistics. *For Policy Econ* 5(2):187–202
- Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. *Remote Sens Environ* 62:77–89
- Sykes MT, Prentice IC, Cramer W (1996) A bioclimatic model for the potential distributions of north European tree species under present and future climates. *J Biogeogr* 23:203–233
- Thuiller W, Vayreda J, Pino J, Sabate S, Lavorel S, Gracia C (2003) Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). *Glob Ecol Biogeogr* 12:313–325
- Tr oltzsch K, van Brusselen J, Schuck A (2009) Spatial occurrence of major tree species groups in Europe derived from multiple data sources. *For Ecol Manag* 257(1):294–302
- UN-ECE (2005) European forest sector outlook study, main report. ECE/TIM/SP/20. United Nations, Geneva
- Verkerk PJ, Lindner M, Zanchi G, Zudin S (2011) Assessing impacts of intensified biomass removal on deadwood in European forests. *Ecol Indic* 11:27–35
- Walvoort DJ, de Gruijter JJ (2001) Compositional kriging: a spatial interpolation method for compositional data. *Math Geol* 33:951–966
- Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester