Alternative approaches to assessing the natural regeneration of Scots pine in a Mediterranean forest

Daniel Moreno-Fernandez, Isabel Cañellas, Ignacio Barbeito Sanchez, Mariola Sánchez-González, Alicia Ledo

Annals of Forest Science (2015) 72:569–583
DOI 10.1007/s13595-015-0479-4

ORIGINAL PAPER

Alternative approaches to assessing the natural regeneration of Scots pine in a Mediterranean forest

Daniel Moreno-Fernández[1,2] · Isabel Cañellas[1,2] · Ignacio Barbeito[3] · Mariola Sánchez-González[1,2] · Alicia Ledo[1,4]

Received: 18 December 2014 / Accepted: 25 March 2015 / Published online: 15 April 2015

Abstract

• ***Key message*** **In modelling regeneration patterns, parametric regression is recommended because it can account for the spatial and temporal correlation present in the data, whereas decision trees allow more complex interactions and can be used to reduce the number of variables.**

• ***Context*** The establishment of seedlings after regeneration fellings is key to guaranteeing the development and persistence of the forest. Depending on the objective pursued, data available or type of forest, a number of different methods have been employed to assess the relationship between seedling establishment and both environmental and stand factors. Most authors have conducted their analyses using parametric regression or point pattern analysis.

• ***Aim*** We analysed the way in which light, stand conditions, edaphic and topographic variables affect the regeneration of *Pinus sylvestris* L. in Central Spain. We used different methods to analyse the same data set. The strengths and weaknesses of each method were discussed.

• ***Methods*** We used two parametric approaches: generalized linear mixed model regression using a negative binomial followed by the variant explanatory variables reduction prior to regression as well as three nonparametric approaches not commonly employed in forest regeneration: nonmetric multidimensional scaling, regression trees and random forests algorithm.

• ***Results*** The parametric regression identified a larger number of variables associated with the regeneration process and the inclusion of a random effect in the model allowing the consideration of the spatial variability among plots. However, decision trees captured the complex interaction among variables, which typical parametric methods were unable to detect.

• ***Conclusion*** Different statistical methods gave similar insights into the underlying ecological process. However, different statistical premises with inference implications can be noticed. This may give misinterpretation of the model depending on the nature of the data. The choice of a given method should be made according to the nature of the data and the achievement of desirable results.

Keywords CHAID · Random forests · PCA · NMDS · Negative binomial · Generalized linear mixed models

Handling Editor: Aaron R Weiskittel

Contribution of the co-authors Daniel Moreno-Fernández: data analysis, analysis of results and writing the manuscript
Isabel Cañellas: experimental design, discussion of the results and coordination of the work.
Ignacio Barbeito: experimental design and discussion of the results.
Mariola Sánchez-González: discussion of the results.
Alicia Ledo: supervision of the analysis, discussion of the results and participation in the writing of the manuscript.

✉ Daniel Moreno-Fernández
danielmorenofdez@gmail.com

Isabel Cañellas
canellas@inia.es

Ignacio Barbeito
ibarbeito@nancy.inra.fr

Mariola Sánchez-González
msanchez@inia.es

Alicia Ledo
alicialedo@gmail.com

[1] INIA-CIFOR, Ctra. A Coruña km 7.5, 28040 Madrid, Spain

[2] Sustainable Forest Management Research Institute, Universidad de Valladolid & INIA, Palencia, Spain

[3] INRA, UMR1092, Laboratoire d'Etude des Ressources Forêt Bois (LERFoB), Centre INRA de Nancy, 54280 Champenoux, France

[4] University of Aberdeen, School of Biological Sciences. Plant and Soil Department, St Machar Drive 23, Aberdeen AB24 3UU, UK

1 Introduction

Natural forest regeneration is a key process in ensuring forest sustainability (Lucas-Borja 2014). However, in the Mediterranean basin, natural regeneration is hampered by the summer drought, high summer temperatures, long intervals between good seed crops, the presence of livestock and the existence of a too dense herb and litter layer (Barbeito et al. 2011; Calama and Montero 2007; Manso et al. 2013; Pardos et al. 2007). Therefore, understanding the dynamics, patterns and factors involved in the success or failure of regeneration and interactions at the seedling level can provide foresters with the fundamental knowledge required for decision-making in forest management (Lucas-Borja 2014).

Regeneration in forest stands has been studied using different statistical methods depending on the aim, data available or type of forest (Table 1). Analyses have been refined as statistical techniques have improved (Zuur et al. 2007). Regeneration studies often involve counts of the number of seedlings per sampling unit. However, a wider range of techniques can be applied to assess seedling abundance. An initial, simple approach is the linear model (del Cerro et al. 2009; Osem et al. 2013). However, regeneration data rarely satisfy the basic assumptions for linear models: normality of errors, linearity of parameters, homogeneity of variance and independence of the covariates (Zuur et al. 2010).

Explanatory spatial covariates that influence recruitment patterns have also been incorporated into different types of recruitment models, as explanatory variables. The covariates most used are soil nutrients and/or soil moisture (Barbeito et al. 2009), light availability (Adili et al. 2013; Fyllas et al. 2008), environmental conditions (Rathbun and Fei 2006) and stand structure (Manso et al. 2013). Vertical and horizontal forest structures determine light availability and, therefore, regeneration dynamics (Catovsky and Bazzaz 2002; Montgomery 2004) and the development of the stand (Boyden et al. 2005; Montes and Cañellas 2007). Some authors have assessed the influence of the retained trees on regeneration through the inclusion of structure influence indices in the model (Barbeito et al. 2011; Kuuluvainen and Pukkala 1989; Paluch 2005).

Regeneration studies often involve the consideration of a lot of explanatory variables that may be highly correlated. Collinearity among covariates increases the risk of inferring that these covariates have no explanatory power (increasing the risk of type I errors), and in addition, it can be especially problematic when ecological signals are weak (Zuur et al. 2010). Collinearity is often assessed through variance inflation factors, pairwise scatterplots comparing covariates or correlation coefficients (Zuur et al. 2010). In order to avoid correlation and to reduce the number of independent variables, scaling techniques can be used to create a dimensionally reduced data set with uncorrelated variables (Borcard and Legendre 2002; Legendre and Legendre 1998; Strobl et al. 2009). The most widely used tools in this regard are principal component analysis (PCA) and factor analysis or nonmetric multidimensional scaling procedure (NMDS). The obtained uncorrelated new covariates obtained can be then used in standard regression methods. The requirement of normality of the data in the PCA analysis is under discussion (Jolliffe 2002; Borcard and Legendre 2002; Zuur et al. 2010). In this respect, NMDS is not sensitive to normality, to outliers and to homoscedasticity assumptions of classical metric scaling (Casini et al. 2004; Legendre and Legendre 1998). In addition, NMDS allows the identification of non-linear gradients of the environmental covariates. However, the disadvantage of using ordination methods is that the original input variables are projected into a reduced set of components or axes, so that

Table 1 Methods used to study natural pine regeneration in Mediterranean mountains

Study	Pine species studied	Methods
Barbeito et al. 2009	*P. sylvestris*	Point pattern
Barbeito et al. 2011	*P. sylvestris*	Generalized linear model after ordination method
Christopoulou et al. 2014	*P. nigra*	Mixed effect model and boosted regression trees
del Cerro et al. 2009	*P. nigra*	General linear model
Fyllas et al. 2008	*P. brutia* and *P. nigra*	Generalized linear model
Gómez-Aparicio et al. 2006	*P. nigra*	General linear model
González-Martínez and Bravo 2001	*P. sylvestris*	General linear model after ordination method
Lucas-Borja et al. 2012	*P. nigra*	General linear model
Montes and Cañellas 2007	*P. sylvestris*	Point pattern analysis
Osem et al. 2013	*P. halepensis*	General linear model
Pardos et al. 2007	*P. sylvestris*	Point pattern analysis and general linear model
Pausas et al. 2004	*P. halepensis*	General linear model
Prévosto et al. 2012	*P. halepensis*	Generalized and general linear model
Spanos et al. 2000	*P. brutia*	Linear and exponential regression

their individual effect is no longer identifiable (Strobl et al. 2009).

This interpretability problem can be avoided by using decision or regression trees, such as classification and regression trees (CART; Breiman et al. 1984), conditional inference trees (Hothorn et al. 2006) or chi-squared automatic interaction detection (CHAID; Kass 1980). Besides allowing the use of multiple variables in the data set, these nonparametric procedures are capable of identifying interactions between the variables which are too complex to be captured by parametric regression models (Strobl et al. 2009). This can be especially useful in the field of ecology. Forest data have been analysed using regression trees (CHAID, Álvarez-Álvarez et al. 2011; conditional inference trees, Barbeito et al. 2012), but their use in regeneration studies is scarce (Fei 2010). However, simple tree models are vulnerable to small changes in the learning data (Strobl et al. 2008). Machine learning methods, such as the random forests algorithm (Breiman 2001) or boosted regression trees (Elith et al. 2008), solve the problem of instability by averaging an ensemble of trees into a more robust composite model (Cutler et al. 2007; Strobl et al. 2009).

Natural, rather than artificial, regeneration is the preferred option in Mediterranean areas. The shelterwood system is commonly employed to ensure the establishment of new Scots pine seedlings (*Pinus sylvestris* L.) in Central Spain (Cañellas et al. 2000). However, regeneration is sometimes hampered by excessively short regeneration periods or by a number of abiotic and biotic factors, such as competition from grass or the large fluctuation in cone production over the years (Barbeito et al. 2011; Cañellas et al. 2000; González-Martínez and Bravo 2001) which is especially notable in masting species such as *Pinus pinea* L. (Calama et al. 2011; Manso et al. 2013). With regard to the abiotic factors, the survival of *P. sylvestris* seedlings is closely linked to climate. In particular, summer drought can cause complete seedling mortality during the first summer after the regeneration fellings (Pardos et al. 2007). In addition, light availability has been shown to be relevant for regeneration on a small scale, with moderate light conditions being optimum to maximize the regeneration density of Scots pine in the Central Mountain Range in Central Spain (Barbeito et al. 2009; Pardos et al. 2007). Topsoil variables, such as soil moisture and nutrients, as well as soil preparation have also been identified as important factors in forest regeneration (Barbeito et al. 2011).

The aim of this study is to employ and compare five different statistical techniques in order to evaluate the influence of environmental factors on natural regeneration, using as a case study of Scots pine forest in Central Spain. The techniques considered are (i) a generalized linear mixed model (GLMM), (ii) data reduction using the PCA ordination method prior to GLMM, (iii) ordination using the NMDS, (iv) CHAID algorithm and (v) random forests algorithm and the conditional inference trees. Additionally, we aimed to answer the following question: What is the most suitable method to evaluate tree recruitment and factors implicated? We compared the results obtained using the different methods, then assessed and discussed the strengths and weakness of each methodology. We calculated the goodness of fit for each methodology to assess how well the different models fit the observations. Measures of goodness of fit summarize the discrepancy between the observed values and the values expected under a statistical model (Maydeu-Olivares and García-Forero 2010). The importance of the methodology used in regeneration studies will be evident if not all the approaches identify the same biotic and abiotic variables that are associated with the regeneration abundance. Additionally, owing to the presence of complex interactive effects among the underlying variables and to correlations among the potential covariates used to represent those variables, it is expected that the parametric modelling strategies considered in this study would fail to identify the importance of covariates detected as important by decision trees.

2 Materials and methods

2.1 Study site

This study was conducted at the mountain forest of Navafría (41° 00′ N, 3° 48′ W), located on a north-facing slope in the Central Mountain Range in Spain. The altitude of this Scots pine forest ranges from approximately 1200 to 2200 m, and the altitudinal position of the timberline is around 1800 m. Annual rainfall exceeds 730 mm, and mean annual temperature is around 9.9 °C. These mountains are composed of granite and gneiss, with fairly homogeneous soils throughout the pinewoods, predominantly humic cambisol-type soils or leptosols at the higher sites (Forteza et al. 1988) according to the FAO taxonomic soil classification of 1989. The site index of these stands ranges between 21.4 and 29.9 m (Montes et al. 2007). Scots pine forms pure stands in the middle and high altitudes of the forest, whereas in the lower parts of the forest, this species grows alongside *Quercus pyrenaica* Willd. Some shrubs, such as *Genista florida* L., *Cytisus scoparius* (L.) Link and *Cistus laurifolius* L., appear in patches of forest which are at the reinitiation stage of the stand development.

The Navafría forest is managed using a permanent block plan with a 100-year rotation period. The management plans date from the end of the nineteenth century. The forest is divided into three working groups and these in turn into three wood production circles. Each wood production circle is made up of five blocks (García López 1994). Natural regeneration is achieved through the uniform shelterwood system over a 20-year regeneration period. Artificial regeneration is only employed in exceptional circumstances as a supplementary measure or aid. The silvicultural management applied consists

of an intensive thinning from below regime from years 30 to 80 (removing subcanopy trees) or mixed (removing dominated and co-dominant trees). The canopy is then opened, leaving a density of 200–250 trees ha^{-1} at the beginning of the regeneration period. Additional trees are subsequently removed in a uniform manner throughout the regeneration block in two or three harvests over the regeneration period, leaving a low residual tree density. Thus, the seedlings become established under the protection of the older trees (Loftis 1990). The remaining trees (30–40 trees ha^{-1}) are finally harvested at 100 years of age. In order to facilitate the establishment of the regeneration, soil preparation operations, subsoil and blade scarification are carried out before the final harvest. Through this management approach, natural regeneration of Scots pine throughout the forest is achieved successfully.

We installed a linear transect along a contour line in each of the five regeneration blocks. The five transects were installed after the first regeneration fellings and prior to the final felling, in the middle of the regeneration period. Square plots of 64 m^2 in size were established for regeneration inventory purposes at 50-m intervals along these transects, making a total of 45 plots across the five blocks (Table 2). Each plot was subdivided into four square subplots where seedlings (individuals shorter than 130 cm) were counted. Additionally, all the trees and the stumps within a 15-m radius from the centre of each regeneration plot were mapped. Measurements included diameter at breast height (dbh) of the trees, total tree height and diameter of the stumps at ground level. The topsoil characteristics of each plot were also recorded, consisting of a mixture of samples taken from a depth of 20 cm and collected at the centre of each subplot to obtain the following topsoil variables: soil pH (in distilled water); percentage of sand, lime and silt according to the International Society of *Soil* Science; the percentage of oxidizable organic matter (Walkley and Black method); available P (Olsen method); colloidal K, Ca, Mg and Na (all atomic absorption); the electrical conductivity (in distilled water); and the stoniness (percentage of stones in the uppermost 20 cm of the soil) (Table 2). We took a hemispherical photograph at the centre of each plot at a height of 1.30 m above the ground to measure light availability (Table 2). The Global Site Factor (GSF) was then calculated using HemiView Canopy Analysis software (Delta-T Devices Ltd.). The GSF is the proportion of total radiation under a plant relative to that in the open, ranging from 0 (completely closed canopy) to 1 (completely open canopy). We also measured plot slope, expressed as a percentage and the altitude (Table 2). We considered the same altitude for all the plots in each transect.

In order to assess the influence of local stand structure, two indices for the measured trees were calculated: the influence potential (IPOT) and the index of influence (INF; Woods and Acer 1984) of the retained trees at a given point. IPOT is based on the concept of ecological field theory (Wu et al. 1985), empirically modified by Kuuluvainen and Pukkala (1989):

$$\mathrm{IPOT}_j = 1 - \mathrm{GPOT}_j \quad (1)$$

where

$$\mathrm{GPOT}_j = \prod_{i=1}^{n_j}\left(1-\beta_{ij}\right) \text{ and } \beta_{ij} = \left(\mathrm{dbh}_{ij} \Big/ \max(\mathrm{dbh})\right)\exp\left(-b_{ij}\cdot s_{ij}\right)$$

β_{ij} is the potential influence of tree i at plot centre j, s_{ij} is the distance from tree i to the plot centre and dbh_{ij} is the diameter of tree i at the plot centre j. Max(dbh) is the maximum diameter at breast height in the data set; 95 cm in our data set. The parameter b_{ij} was replaced by $1/(a\ h_{ij})$, where h_{ij} is the height (m) of tree i at plot centre j and a is a parameter to be estimated. Alternative values of a from 0 to 1 by 0.1 based on the correlation between regeneration density and IPOT were tested. The highest correlation between a and regeneration density was found for the value a=0.4. IPOT ranges from 0 (no competition) to 1 (maximum competition).

Moreover, the INF competition index was defined as

$$\mathrm{INF}_j = \sum_{i=1}^{n} \mathrm{dbh}_{ij}\exp\left(-2s_{ij}^2\mathrm{dbh}_{ij}^{-1}\right) \quad (2)$$

where INF_j represents the index of influence in a plot centre j, and s_{ij} and dbh_{ij} are defined above. INF acquires low values in gaps and high values in dense overstory patches (Paluch 2005).

The target variable in our case study was the number of seedlings per plot, i.e. count data. This is an important attribute that should be considered in the method selection. All the measured variables (edaphic, light, topographic and stand condition variables; Table 2) were included as explanatory variables in the following fitted models.

2.2 Parametric regression: GLMM using a negative binomial

Generalized linear mixed models (GLMMs) extend linear regression mixed models to accommodate non-constant variance through a variance function, non-linear relationships and certain types of non-normally distributed errors (Bolker 2008; Hardin and Hilbe 2012; Venables and Ripley 2002). There are three steps in GLMMs: (1) choose a distribution function for the response variable, which is a function from the exponential family; (2) define a linear predictor function specifying the covariates; and (3) link the predictor function and the mean of the distribution (Zuur et al. 2007). The connection of the mean of the distribution function to the linear predictor is done using link functions, e.g. logit, log or probit

Table 2 Summary of the main characteristics of the study area

Variables	Mean	Minimum	Maximum	Standard deviation
Number seedlings per subplot (4×4 m)	17	1	125	24
Edaphic variables				
pH	6.04	4.37	8.57	1.50
Electric conductivity (mS/cm)	0.11	0.03	0.48	0.09
Sand (%)	65.85	55.68	81.96	5.44
Silt (%)	13.91	5.64	23.64	3.74
Clay (%)	20.26	9.68	29.40	4.20
Organic matter (%)	10.86	4.75	21.97	3.58
Available P (mg/kg)	7.7	4.3	20.5	4.2
Available K (mg/kg)	181	87	379	61
Available Ca (mEq/100 g)	4.5	1.9	9.1	1.8
Available Mg (mEq/100 g)	0.85	0.25	2.50	0.45
Available Na (mEq/100 g)	0.06	0.01	0.16	0.03
Stoniness (%)	19.60	2.50	100	15.10
Light variable				
Global Site Factor (0–1)	0.62	0.22	0.94	0.16
Topographic variables				
Slope (%)	25	5	65	10
Elevation (m)	1540	1381	1730	138
Competition variables				
IPOT (0–1)	0.68	0.11	0.97	0.23
INF	39.24	0.01	152.80	33.41
Number of retained trees	6	0	15	4
Number of stumps	4	0	17	4

Stoniness percentage of stones in the uppermost 20 cm of the soil, *Global Site Factor* the proportion of total radiation under a plant, *IPOT* influence potential (Wu et al. 1985; Kuuluvainen and Pukkala 1989), *INF* index of influence (Woods and Acer 1984)

(Venables and Ripley 2002). Count data may be assumed to follow a Poisson distribution or a negative binomial distribution. However, in the Poisson distribution, a single parameter quantifies both the mean and the variance, so the use of Poisson distribution is reduced to data that matches the requirement of having a value of variance similar to the mean value (Crotteau et al. 2014; Lawless 1987). When the variance values are higher than the mean, the negative binomial distribution is more appropriate than the Poisson distribution (Bliss and Fisher 1953; Venables and Ripley 2002). Although the negative binomial distribution cannot strictly be included in the exponential family, negative binomial models can be fitted using a small extension of the GLMMs (Bolker 2008). The negative binomial distribution is characterized by a dispersion parameter, θ, which relates the mean (μ) of a response variable y and its variance: $V(y)=\mu+(\mu^2/\theta)$ (Bolker 2008; Crotteau et al. 2014). This parameterization of the variance, termed NB2 by Hardin and Hilbe (2012) or restricted negative binomial by Crotteau et al. (2014), assumes that the value of θ is constant across factors.

Observations recorded in forestry and ecology frequently present spatial or temporal correlation: several measurements are taken from the same subject or the sampling is carried out hierarchically. If these correlations are ignored, the assumption of independence of the observations and the error terms is violated and the standard error and p values of the covariates might be affected (Zuur et al. 2007). The correlation among measurements can be taken into account by including random effects in the model (Fisher 1918; Henderson et al. 1959; Zuur et al. 2009). Pure spatial and temporal correlation can be addressed through semivariograms or correlograms, respectively (Zuur et al. 2009). Nevertheless, in our study, subplots are nested within the plots and the plots in turn nested within the transect. Hence, a plot random intercept and a plot nested within transect random intercept effects were included in the model. Thus, the full negative binomial mixed model can be written as follows:

$$y = \text{negative binomial}(\mu, \theta) \tag{3}$$

$$\log(\mu) = X\beta + Zu \tag{4}$$

where X is the matrix of the covariates, β is the vector of the unknown (although estimable) parameters of fixed effects, Z is

the random-effects design matrix and u is the vector of random effects with mean zero and variance to be estimated. After fitting the model, we used the McFadden's pseudo-R^2 to assess the goodness of fit (Domenich and McFadden 1975):

$$\text{Mc Fadden}'\text{s } R^2 = 1 - \frac{-2\log L_{\text{Full}}}{-2\log L_{\text{Null}}} \quad (5)$$

where L_{Full} and L_{Null} are the likelihood of the full model and the likelihood of the null model. McFadden's pseudo-R^2 values can range between 0 and 1. Nevertheless, the resulting values are not equivalent to the classical R^2 values. Indeed, values of McFadden's pseudo-R^2 between 0.2 and 0.4 should be taken to represent a very good model fit (Louviere et al. 2000).

To avoid over parameterization, the inclusion of the random effects, the covariates (Table 2) and the interaction of the variables in the model was carried out via a forward stepwise procedure according to the likelihood ratio test using a [Pr>(Chi)]=0.05 as a confidence level. This analysis was conducted by using the "glmmadmb" function of the "glmmADMB" package (Fournier et al. 2012) in R 3.0.2. (R Core Team 2013).

2.3 Including uncorrelated variables in the GLMM. PCA as ordination method

In the former section (2.2), the explanatory variables were included sequentially to avoid collinearity. A second option to deal with this problem is presented in this section. The measured covariates represent a multivariate data set, a collection of sites positioned in a space where each variable defines one dimension (Borcard et al. 2011). In this case, an ordination method can be used to reduce the number of variables to a new set of uncorrelated variables, which can be therefore included as explanatory variables in the GLMM fitting. The PCA is a classic multivariate technique (Legendre and Legendre 1998). It reduces the dimensionality of a group of variables into a new set of uncorrelated variables, called the principal components. The principal components retain all of the variable information, but weight linear combinations of the original variables (Abdi and Williams 2010; Jolliffe 2002; Legendre and Legendre 1998). The first principal component axis data set is the line that goes through the most explanatory dimension describing a multinomial distribution. The following axes are orthogonal to one another and explain successively less (Abdi and Williams 2010; Borcard and Legendre 2002). The PCA is an eigenvector-based method and requires the construction of a disperse matrix S, which assesses the distance among observations. The mathematical details are beyond the scope of this study, but this information can be found in Jolliffe (2002) and Legendre and Legendre (1998). The requirement of normality of the original variables in the PCA analysis has been widely discussed, and some authors state that normality is not required (Jolliffe 2002; Zuur et al. 2010), whereas others argue that data must be normally distributed (Borcard and Legendre 2002).

We carried out two PCAs, one using edaphic variables and the other using the competition and light variables. The eigenvectors and the scores of the principal components were calculated after applying an orthogonal varimax rotation (Legendre and Legendre 1998). The PCA was conducted using the "principal" function of the "psych" package (Revelle 2013) in R. Seedling abundance was then modelled using Eq. (1). In this case, the predictors were the scores of the first and second principal components of each PCA and the topographic variables (slope and altitude). Two or three components are commonly used following ordination methods (Barbeito et al. 2011; González-Martínez and Bravo 2001) because those components typically explain a large amount of the total variance. The criteria described in the previous section (2.2) were also used for the inclusion of covariates. As a measure of goodness of fit, we used the statistics described in the previous section.

2.4 Ordination using NMDS

NMDS (Shepard 1962) is another ordination method. As with PCA, it is also an ordination technique, because it reduces a multivariable set into orthogonal univariate axes. However, unlike PCA and other ordination methods, NMDS is not based on an eigenvalue solution (Legendre and Legendre 1998). This method requires the construction of a community dissimilarity distance matrix (Legendre and Legendre 1998). For this purpose, any measure of association can be used, not only correlation, which sometimes is not the most appropriate measure (Zuur et al. 2007). Several distance measures can be employed to build the dissimilarity distance matrix, e.g. Bray-Curtis, Manhattan or Jaccard. Measure selection depends on the availability and nature of the data. To calculate those dissimilarities, numerical optimization methods have to be used. NMDS is an iterative method. The iterative procedure attempts to position the objects in the three dimensions so as to minimize a stress function (scaled from 0 to 1), which measures how far the distances in the reduced-space configuration are from the original distances (Borcard et al. 2011; Legendre and Legendre 1998). The mathematical details can be found in Legendre and Legendre (1998).

We selected the Bray-Curtis empirical dissimilarity distance matrix (Bray and Curtis 1957) to reduce the observed 19 variables (Table 2) in three dimensions or axes. We selected this distance because it is one of the most appropriate metric distance for community ecology data (Legendre and Fortin 1989), and it has been used in recruitment pattern assessment before (Ledo et al. 2015). We then plotted the Shepard diagram (Borcard et al. 2011; Legendre and Legendre 1998) to

compare the original (empirical) distances with the ordination distances in the three-dimensional space. The goodness of fit of the ordination is measured as the R^2 of both a linear and a monotone (nonparametric) regression of the NMDS distances on the original ones (Borcard et al. 2011).

We grouped the number of seedlings for each subplot into a categorical variable with two levels: one corresponding to the fourth quartile (Q4) and the other to the first, second and third quartiles (Q123) of the distribution of the number of seedlings. We chose this division because the number of seedlings was relatively homogeneous in the first, second and third quartiles of the distribution, but increased greatly in the fourth quartile. We plotted the centroids of the quartiles to interpret the environmental variables with respect to the quartiles. We drew an ordination hull to enclose and cluster the environmental variables into the two levels (Q123 and Q4). Ellipses for the 95 % confidence areas of the class centroids were also calculated (Oksanen 2013). Furthermore, smooth surfaces of covariates to ordinations can be fitted using generalized additive models (Cayuela et al. 2006; Oksanen et al. 2013). NMDS was conducted using the function "metaMDS" implemented by the R package "vegan" (Oksanen et al. 2013).

2.5 Decision tree automatic classification method: CHAID algorithm

CHAID is a nonparametric procedure that splits a data set (root node) into groups, classes or segments that differ with respect to the response variable (Kass 1980; Kleppin et al. 2008). CHAID does not necessarily produce dichotomous categories, and therefore, a node can be split into more than two categories (Álvarez-Álvarez et al. 2011; van Diepen and Franses 2006). No distributional assumptions of the data are required in the CHAID, which implies a great advantage over other methods such as the GLMM parametric approach that we followed in Section 2.2. The CHAID mode operates as follows: in a first step, continuous variables are divided into a number of categories with approximately equal numbers of observations, since the CHAID procedure operates on nominal variables (Álvarez-Álvarez et al. 2011; van Diepen and Franses 2006). CHAID then splits the data for the specific predictors which exhibit the strongest degree of association with the response variable (Álvarez-Álvarez et al. 2011; Kleppin et al. 2008; van Diepen and Franses 2006). Since CHAID is a forward stepwise approach, it is possible that a better segmentation solution can be found by using another variable at an earlier stage (Vriens 2001). Mathematical details of the method are in van Diepen and Franses (2006) and Kass (1980).

The root node in the present study is the seedling density. We used a significance level of 5 % of the chi-squared test of independence to decide which categories of each predictor to merge and which predictor to split. As the chi-squared test is only approximately chi-squared distributed, a large sample size is required (van Diepen and Franses 2006). CHAID was carried out using SPSS (IBM Corp 2012).

We evaluated the goodness of fit using the classic R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (5)$$

where y_i is the observed density, $\hat{y}_i$ is the estimated seedling density, $\bar{y}$ is the mean of the density and n is the number of observations.

2.6 The random forests algorithm and conditional inference tree

The random forests algorithm (Breiman 2001) is a nonparametric technique that combines the prediction of many independent decision tree models into a robust composite model (Cutler et al. 2007), thereby improving the accuracy of the prediction (James et al. 2013). To achieve this, the random forests procedure generated 500 bootstrapped data sets from the original data set. The procedure then constructs, by default, 500 bagged regression trees on the bootstrapped training samples and aggregates the bagged trees for prediction (James et al. 2013; Strobl et al. 2009). The remaining observations not used to fit a given bagged tree, referred to as out-of-bag (OOB) observations, are used to compute the prediction accuracy of the given tree (Strobl et al. 2009). In addition, the algorithm permutes a random sample of the predictors for determining each split in each tree, producing diverse uncorrelated trees (James et al. 2013; Strobl et al. 2009). Thus, the preference for certain predictors due to their scale of measurement or importance is avoided (Barbeito et al. 2012). The preference for a given predictor could result in the bagged trees being quite similar to each other and, hence, highly correlated predictions (James et al. 2013). The predictor importance is the difference in prediction accuracy before and after permuting, averaged over all trees (Breiman 2001; Strobl et al. 2009). A conditional permutation scheme for the computation of the variable importance measure was used in order to account for the correlation among the variables (Strobl et al. 2008). The variable importance measure is based upon the mean decrease in accuracy of predictions on the out-of-bag samples when a given variable is excluded from the model (James et al. 2013). According to Grömping (2009), the prediction of the random forests algorithm can be estimated via mean square error from the OOB data (OOB-MSE) as

$$\text{OOB-MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i\text{OBB}})^2 \quad (6)$$

where $\hat{y}_{i\mathrm{OBB}}$ denotes the average prediction for the ith observation from all trees for which this observation has been OOB. As with linear regression, with the average of sum of squares $\mathrm{SST} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, OOB-$R^2$ can be obtained as 1−OOB-MSE/SST.

Finally, we built an unbiased tree based on non-parametrical conditional inference procedures for testing independence between response and each predictor. The split selection criterion is based on conditional inference tests or permutation test (Hothorn et al. 2006). We used a significance level of 5 % as a splitting node criterion in tree construction. The goodness of fit of the conditional inference tree was evaluated thorough the R^2 described in the CHAID section. We employed the functions cforest and ctree of the R package "party" (Hothorn et al. 2006; Strobl et al. 2008) to perform the random forests analysis and the conditional inference tree.

3 Results

3.1 Parametric regression: GLMM using a negative binomial

We assessed the influence of environmental variables on the number of seedlings per subplot through a negative binomial regression with a log link. Seedling distribution was explained through various models in which the different covariates were sequentially included (Table 3). The most accurate model (in terms of lower log-likelihood) describing seedling abundance included the content of Na, P and the number of stumps within a 15-m radius (Table 3). The Na (estimation coefficient±standard errors=−12.90±3.20, $p<0.0001$) and P (estimation coefficient±standard errors=−0.07±0.029, $p=0.012$) content showed a negative association, while stump density had a positive association (estimation coefficient±standard errors=0.09±0.028, $p=0.0001$) on regeneration density (Table 3). Altitude, GSF, the logarithm of electric conductivity and stoniness improved the model only very slightly (0.05 <[Pr>(Chi)]<0.1) in terms of log-likelihood, so these were not included in the model. GSF (estimation coefficient±standard errors=1.18±0.71, $p=0.0941$) and altitude (estimation coefficient±standard errors=0.016±0.0008, $p=0.0579$) were positively correlated with seedling density, whereas the logarithm of electric conductivity (estimation coefficient±standard errors=−0.35±0.20, $p=0.0832$) and stoniness (estimation coefficient±standard errors=−0.01±0.01, $p=0.0657$) was negatively correlated. As regards the interactions among variables, we did not find a significant effect of any of the three interactions tested (Table 3). The model including the plot random effect performed better, in terms of likelihood, than the model without random effects. This fact pointed out the importance of considering the spatial correlation within the plots in the model (Table 3). McFadden's pseudo-R^2 was 0.022, indicating poor model fit (Louviere et al. 2000).

3.2 Including uncorrelated variables in the GLMM. PCA as ordination method

The edaphic variables were reduced to two principal components, the first component was mainly related to the available Ca and Mg, whereas the second component was related to the pH and the electric conductivity (Table 4). Together, the two components explained 32 % of the total variance. Concerning the light and competition variables, the PCA reduced the number of variables to two principal components. The number of trees was related to the first component and the number of stumps to the second (Table 4). In this PCA, the two components explained 48 % of the total variance.

We found that the second component of the PCA of the light and competition variables (related to the number of stumps) was positively associated (estimation coefficient± standard errors=0.342±0.119, $p=0.004$) in terms of likelihood ratio test (([Pr>(Chi)]<0.05); Table 5) to the seedling density. We did not find a significant effect ([Pr>(Chi)]>0.05) of the two other components or the two topographic variables (slope and altitude) on the number of seedlings. In this regression, the variance of the plot random intercept was 0.45. Regardless of the goodness of fit, McFadden's pseudo-R^2 was 0.01 indicating a quite poor model fit (Louviere et al. 2000).

3.3 Ordination with NMDS

The NMDS (Section 2.4) reduced the dimensionality of the 19 variables to three dimensions; one related to the light variables and to the indices of the influence of retained trees on regeneration, the second related to the soil variables, and the last related to the number of stumps. The R^2 of the linear ordination and the monotone of regression of the NMDS distances on the original ones were 0.88 and 0.98, respectively. The stress value was minimized to 0.14. The NMDS diagram shows that the entire hull of Q4 was enclosed by the hull of Q123 (Fig. 1). The ellipses for the 95 % confidence areas of both classes were almost overlapped, indicating only very slight differences between classes. IPOT and Mg centroids were enclosed by Q123 but not by Q4. The gradient (via contours) of the projected smooth surfaces of IPOT and Mg influenced regeneration density (Fig. 1). The projected smooth surfaces of the INF and the number of trees followed the same pattern as the IPOT, parallel with the IPOT projected smooth surfaces; whereas the surfaces of the number of stumps was also parallel with respect to the IPOT projected smooth surfaces but showed a positive association with the regeneration density. Moreover, both the contours of the

Table 3 Summary of the selection process of the environmental variables and the fitting statistics of the forward stepwise regression

Model	Fixed effects	Random effect	Log-likelihood	Compared to model	Pr>(Chi)
(0)	Intercept	–	−543.56	–	–
(1)	Intercept	Plot	−512.74	(0)	<0.001
(2)	Intercept	Plot/transect	−512.40	(1)	0.4124
(3)	Na	Plot	−508.77	(1)	0.0048
(4)	Na, P	Plot	−506.45	(3)	0.0312
(5)	Na, P, altitude	Plot	−504.71	(4)	0.0663
(6)	Na, P, GSF	Plot	−505.10	(4)	0.0999
(7)	Na, P, slope	Plot	−505.33	(4)	0.1340
(8)	Na, P, pH	Plot	−506.44	(4)	0.8993
(9)	Na, P, IPOT	Plot	−505.24	(4)	0.1201
(10)	Na, P, INF	Plot	−505.31	(4)	0.1305
(11)	Na, P, organic matter	Plot	−506.29	(4)	0.5668
(12)	Na, P, Mg	Plot	−505.67	(4)	0.2114
(13)	Na, P, Ca	Plot	−505.29	(4)	0.1272
(14)	Na, P, K	Plot	−506.34	(4)	0.6375
(15)	*Na, P, stumps*	*Plot*	*−501.28*	*(5)*	*0.0013*
(16)	Na, P, stumps, trees	Plot	−501.21	(15)	0.7043
(17)	Na, P, stumps, Ln(electric conductivity)	Plot	−499.83	(15)	0.0889
(18)	Na, P, stumps, clay	Plot	−501.27	(15)	0.8580
(19)	Na, P, stumps, sand	Plot	−501.26	(15)	0.8415
(20)	Na, P, stumps, silt	Plot	−501.28	(15)	0.9383
(21)	Na, P, stumps, stoniness	Plot	−449.62	(15)	0.0681
(22)	Na, P, stumps, Na*stumps	Plot	−501.28	(15)	0.9643
(23)	Na, P, stumps, Na*P	Plot	−501.23	(15)	0.7567
(24)	Na, P, stumps, stumps*P	Plot	−501.24	(15)	0.7773

The chosen model is in italics

Table 4 Eigenvectors of the environmental variables obtained from the PCAs

Group of variables	Variables	Component 1	Component 2
Edaphic variables	pH	−0.10	*0.95*
	Electric conductivity	−0.04	*0.94*
	Sand (%)	0.11	0.03
	Silt (%)	0.18	−0.04
	Clay (%)	0.01	0.00
	Organic matter	0.02	−0.01
	K	0.10	0.02
	P	−0.14	−0.11
	Mg	*0.85*	−0.17
	Ca	*0.96*	−0.03
	Na	0.34	0.17
	Stoniness (%)	−0.02	−0.14
Light and competition variables	GSF	−0.40	0.36
	IPOT	0.40	−0.36
	INF	0.24	−0.10
	Number of trees	*0.89*	−0.11
	Number of stumps	−0.11	*0.95*

The eigenvectors of the main variables related to each component are in italics

Table 5 Summary of the selection process of the principal components and the fitting statistics of the forward stepwise regression

Model	Fixed effects	Random Effect	Log-likelihood	Compared to model	Pr>(Chi)
(0)	Intercept	–	−543.56	–	–
(1)	Intercept	Plot	−512.74	(0)	<0.001
(2)	Intercept	Plot/transect	−512.40	(1)	0.412
(3)	PC1 com	Plot	−512.55	(1)	0.540
(4)	*PC2 com*	*Plot*	*−508.86*	*(1)*	*0.005*
(5)	PC2 com+PC1 ed	Plot	−508.20	(4)	0.252
(6)	PC2 com+PC2 ed	Plot	−507.81	(4)	0.147
(7)	PC2 com+altitude	Plot	−508.86	(4)	0.950
(8)	PC2 com+slope	Plot	−508.07	(4)	0.209

The chosen model is in italics

PC1 ed component related to the Ca and Mg, *PC2 ed* component related to the pH and the electric conductivity, *PC1 com* component related to the number of trees, *PC2 com* component related to the number of stumps

projected smooth surfaces of the Ca and the Na followed the same pattern than the projected smooth surfaces of the Mg, parallel to the latter and displaying a negative relationship with the regeneration density. The rest of the other variables overlapped each other and were very close to the centroids of Q4 and Q123. Therefore, we assumed that there was no association with the regeneration density; hence, they are not shown in Fig. 1.

3.4 Decision tree automatic classification method: CHAID algorithm

The CHAID procedure split the data into two main branches according to the Na concentration. Thus, this variable played the most statistically important role ($p<0.0001$) in the subplot seedling density (Fig. 2). The lowest concentrations of Na (≤0.030 mEq/100 g) were related to the highest number of seedlings. In the case of higher concentrations of Na (>0.030 mEq/100 g), the algorithm split the data into six branches according to the number of retained trees ($p<0.0001$). The distribution of the data in these six branches did not follow a linear pattern. The R^2 of the CHAID was 0.23, pointing to a poor model fit.

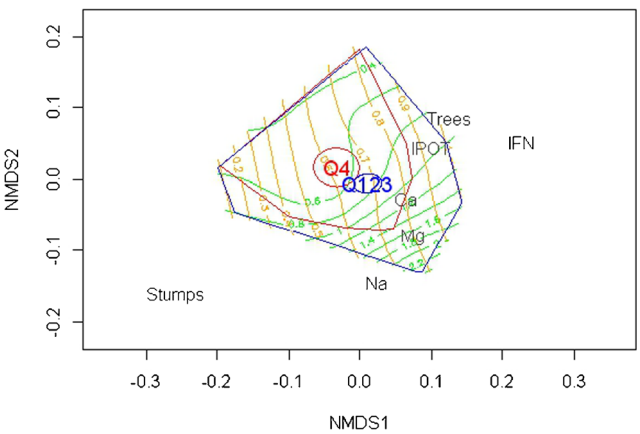

Fig. 1 NMDS ordination of the variables showing the centroids of the quartiles of the number of seedlings. Variables within Q4 hull are not shown to facilitate the visualization. *Red ellipse* and *red hull* Q4 (fourth quartile of the number of seedlings). *Blue ellipse* and *blue hull* Q123 (first, second and third quartiles of the number of seedlings). *Yellow smooth surfaces* IPOT gradient. *Green surfaces* Mg gradient

3.5 Random forests algorithm and conditional inference tree

Across all the trees considered in the random forests algorithm (Section 2.6), the content of sodium and the number of stumps were the two most important variables since they reached the greatest values of conditional variable importance (Fig. 3). The OOB-MSE and the OOB-R^2 of the random forests algorithm were 434.77 and 0.17, respectively.

The first node of the conditional inference tree split the data according to the content of Na ($p<0.001$) (Fig. 4). The largest number of seedlings was related to the lowest concentrations of Na (≤0.040 mEq/100 g). For the highest concentrations of Na (>0.040 mEq/100 g), the branch was divided according to the number of stumps ($p=0.002$), which were positively related to the number of seedlings per subplot (Table 6). The R^2 of the conditional inference tree was 0.17, indicating a poor model fit.

4 Discussion

4.1 Factors underlying Scots pine regeneration in Central Spain

A minimum of 2000 ha^{-1} (4 seedlings per subplot) denotes a sufficient regeneration density for Scots pine (e.g. Hyppönen et al. 2013). In our study, we found that at least 4 seedlings per subplot (2500 seedlings ha^{-1}) were present in 99 % of the subplots suggesting regeneration success and good management practices. Therefore, we can state that the ranges of the

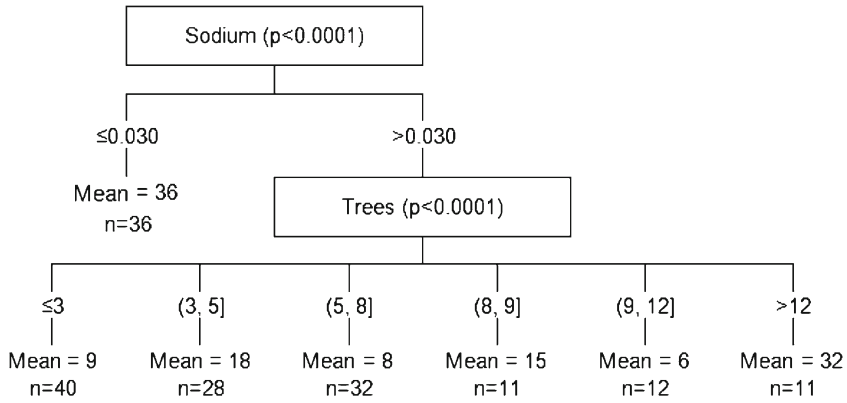

Fig. 2 Regression tree as a result of the CHAID algorithm for seedling density. *Mean* indicates the mean number of seedlings per subplot and *n* the number of subplots per node

measured variables are suitable for achieving natural regeneration. The current values of these variables are within a range that does not limit the regeneration process in the studied forest. Thus, some of the relationships found between the environmental variables and the regeneration density may be arbitrary rather than causal.

The different methods pointed to similar, although not identical, factors underlying the regeneration process (Table 6). The models obtained using the different methods had low goodness of fit values. This indicates that the low association between the environmental variables and the seedling density may be due to the lack of limiting factors in the regeneration process.

The parametric regression, the CHAID, the random forests algorithm and the conditional inference tree found a negative association between available Na and the seedling density. This negative relationship can be explained by the salt sensitivity of Scots pine (Bravo-Oviedo and Montero 2008). However, the sodium concentrations in the study area are lower than those found in other Scots pine forests (Bravo and Montero 2001). Furthermore, there was a positive relationship between the number of stumps and the seedling density according to the GLMM, the GLMM after PCAs reduction, the random forests algorithm and the conditional inference tree. The number of stumps is related to recently cut trees, so seedling density is directly related to canopy openness indicating that higher intensities in the regeneration fellings favoured the establishment of seedlings. In addition, the extraction of the wood resulted in a soil preparation effect, which facilitated the establishment of the Scots pine seedlings (Barbeito et al. 2011). The CHAID found an interaction between the highest levels of Na and the remaining trees. It indicates that the Na is the most important factor driving the regeneration, and when its concentration increases over the optimum, other factors, the remaining trees also appeared as a driver. The NMDS placed the two classes (the fourth quartile and the first, second and third quartiles of the distribution of the number of seedlings) quite close to each other reporting a weak negative relationship between the number of remaining trees and the competition indices with the regeneration density. The remaining trees play an important role in the regeneration, as Scots pine seedlings prefer moderate light conditions in Mediterranean areas (Barbeito et al. 2009; Pardos et al. 2007) rather than intensive light conditions as in Northern countries (Valkonen 2000). Furthermore, in Mediterranean areas, the success of Scots pine regeneration depends on summer droughts, so silviculture operations in our study area is also aiming at reducing the competition for resources thorough the soil preparation and thinnings (Barbeito et al. 2011).

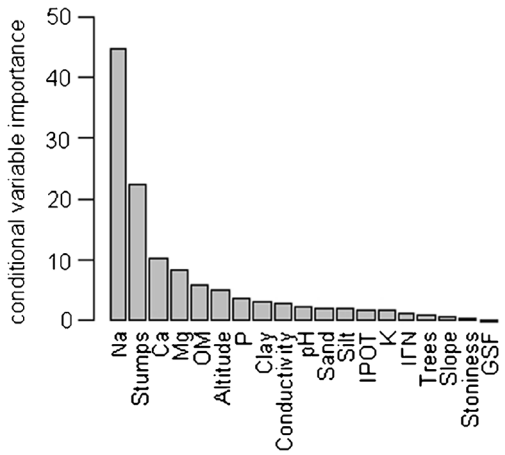

Fig. 3 Conditional variable importance measured following the *permutation principle of the mean* decrease in accuracy importance in the random forests algorithm. Larger values of conditional variable importance indicate more importance in the random forests model

4.2 Methods comparison

The variables used to predict the seedling recruitment were highly correlated. Including this in a regression model is a problem because it increases the risk of type I errors (Zuur et al. 2007). The PCA solved this problem reducing the data into uncorrelated components. Nevertheless, the full-dimensional GLMM had higher values of the goodness of fit statistics (McFadden's pseudo-R^2) than the GLMM including the components of the PCAs. In addition, the reduced

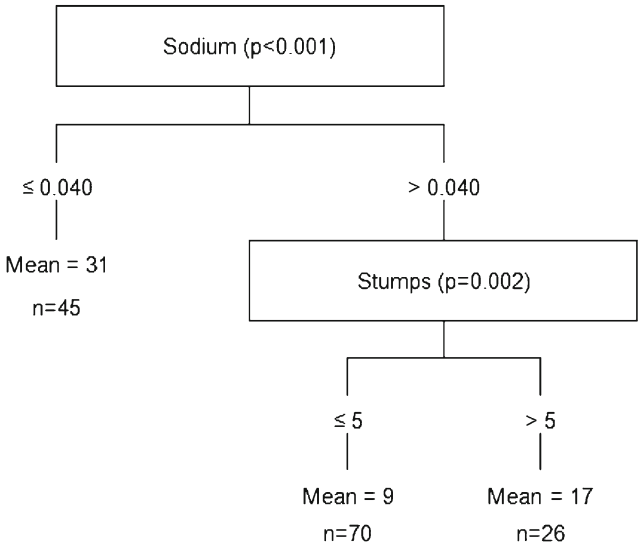

Fig. 4 Implementation of conditional inference trees for seedling density into the defined theory of conditional inference procedures. *Mean* indicates the mean number of seedlings per subplot and *n* the number of subplots per node

components present another flaw, which is that the effect of the individual covariates is not fully identifiable (Strobl et al. 2009).

NMDS could also have been used to reduce the dimensionality of the data for the parametric regression. However, since the NMDS is based directly on the dissimilarities, it does not provide correlations between derived axis scores and the variables, whereas techniques based on the eigenvalues and eigenvectors allow the researcher to relate, at least in part, the original variables to the components (Quinn and Keough 2001). Therefore, if the objective of the ordination technique is to reduce the data to use the scores of the axes in a regression model, the eigenvalue techniques may be more useful, since the correspondence to the component-variables is maintained. However, eigenvalue techniques are based only on correlation or covariate coefficients and limited to Euclidean distances, and this may not be the most appropriate measure of association (Legendre and Legendre 1998; Zuur et al. 2007). In addition, the method can also proceed with missing distance estimates as long as there are enough measures left to position each object with respect to a few of the others (Legendre and Legendre 1998). Furthermore, the NMDS allowed to fit non-linear relationships between the covariates to ordinations using generalized additive models, whereas other methods, such as PCA, imply linear relationships and it cannot always be appropriate (Oksanen 2013).

As regards the regression trees, the R^2 of the CHAID was larger than that of the conditional inference tree. Additionally, the CHAID has the main advantage of splitting each node into more than two branches, i.e. it is not a binary partitioning method (Kass 1980; Kleppin et al. 2008; van Diepen and Franses 2006) as the conditional inference tree (Hothorn et al. 2006). However, since the CHAID is a forward stepwise method, i.e. not all the variables are considered simultaneously but sequentially, there is the possibility that a better segmentation solution can be found using another variable at an earlier stage (Vriens 2001). Thus, CHAID cannot guarantee a single optimal solution (Perreault and Barksdale 1980). In relation to this, the bias in the variable selection is solved by the conditional inference trees using the permutation test principle (Hothorn et al. 2006). Both decision trees, the CHAID and the conditional inference tree, found second-order interactions. In addition, decision trees may outperform the classical approaches if there is a non-linear and complex interaction between the variables and the results of the decision trees are easier to interpret than those of other regression models (James et al. 2013; Strobl et al. 2009). Nevertheless, the main weakness of simple tree models is their instability to small changes in the learning data (Strobl et al. 2008). The random forests algorithm solves the problem of instability by averaging an ensemble of trees into a more robust composite model (Cutler et al. 2007; Strobl et al. 2009). However, the R^2 of the CHAID was larger than that of the random forests algorithm.

Forest regeneration studies in Mediterranean areas have often been conducted using parametric regression or point pattern analysis (Table 1). Despite all the assessed models performed in a similar way and given the similar ecological

Table 6 Significant variables involved in the regeneration process according to the alternative approaches employed

Method	Edaphic and topographic variables	Competition and stand variables
GLMM	Na (−); P (−)	Number of stumps (+)
GLMM after PCA	n.s.	Component related to the number of stumps (+)
NMDS	n.s.	n.s.
CHAID	Na (−)	Number of remaining trees (non-linear)
Random forest and conditional inference tree	Na (−)	Number of stumps (+)

The effect of the variables is in parentheses

n.s. no statistically significant variables identified

results, the election of the statistical technique must be done according to the characteristics and structure of the data. Ecological data often presents both spatial and temporal dependence among the observations. The hierarchical structure of our design indicates spatial dependence among measurements. The GLMM was the only method which allows the spatial correlation between subplots to be considered entering random effects (Fisher 1918; Henderson et al. 1959; Zuur et al. 2007). In studies with multiple hierarchies, like ours, random effects are especially important (Ten Have et al. 1999). We found a significant effect of the plot random effect indicating high correlation among subplots within the same plot. The model with the plot random effect captured more variability than the model without the random part. The other methods employed did not consider this correlation which can provide biased results and hide the effect of some explanatory covariates. Model misspecification is a frequent mistake disregarded in ecological modelling. Additionally, this method also identified the largest number of variables associated with seedling density. Thus, in studies with nested data, the parametric regression model is the most useful statistical approach. In studies with a large set of predictors, decision trees and the random forests algorithm through the conditional variable importance and ordination methods can be used to select and reduce the number of variables to be included in the model. If complex interactions are expected and the data set is large, then decision trees are advisable (Strobl et al. 2009). Whether the researcher had to analyse data sets from a hierarchal design with complex interactions among variables, the parametric regression models could account both issues.

Acknowledgments We wish to thank Adam Collins and Dr. Nicholas Devaney (National University of Ireland, Galway) for revising the English grammar and everybody who participated in the field work, especially Ángel Bachiller. We also thank two anonymous reviewers who provided good comments to improve the paper.

Funding Funding was provided by the Spanish Ministry of Economy and Competitiveness (project AGL2010-21153-C02-01) and Madrid Regional Government (project BOSSANOVA). The Ministerio de Educación, Cultura y Deporte (Ministry of Education, Culture and Sport) funded the corresponding DMF's PhD studies thorough the FPU programme.

References

Abdi H, Williams L (2010) Principal component analysis. Wiley Interdiscip Rev Comput Stat 2:433–459

Adili B, El Aouni MH, Balandier P (2013) Unravelling the influence of light, litter and understory vegetation on *Pinus pinea* natural regeneration. Forestry 86:297–304. doi:10.1093/forestry/cpt005

Álvarez-Álvarez P, Khouri EA, Cámara-Obregón A, Castelo-Dorado F, Barrio-Anta M (2011) Effects of foliar nutrients and environmental factors on site productivity in *Pinus pinaster* Ait. stands in Asturias (NW Spain). Ann For Sci 68:497–509. doi:10.1007/s13595-011-0047-5

Barbeito I, Fortin M-J, Montes F, Cañellas I (2009) Response of pine natural regeneration to small-scale spatial variation in a managed Mediterranean mountain forest. Appl Veg Sci 12:488–503. doi:10.1111/j.1654-109X.2009.01043.x

Barbeito I, LeMay V, Calama R, Cañellas I (2011) Regeneration of Mediterranean *Pinus sylvestris* under two alternative shelterwood systems within a multiscale framework. Can J For Res 41:341–351. doi:10.1139/X10-214

Barbeito I, Dawes MA, Rixen C, Senn J, Bebi P (2012) Factors driving mortality and growth at treeline: a 30-year experiment of 92 000 conifers. Ecology 93:389–401

Bliss CI, Fisher RA (1953) Fitting the negative binomial distribution to biological data. Biometrics 9:176–200

Bolker BM (2008) Ecological models and data in R. Princeton University Press, New Jersey, 382 pp

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol Modell 153:51–68. doi:10.1016/S0304-3800(01)00501-4

Borcard D, Guillet F, Legendre P (2011) Numerical ecology with R. Springer, New York

Boyden S, Binkley D, Shepperd W (2005) Spatial and temporal patterns in structure, regeneration, and mortality of an old-growth ponderosa pine forest in the Colorado Front Range. For Ecol Manag 219:43–55. doi:10.1016/j.foreco.2005.08.041

Bravo F, Montero G (2001) Site index estimation in Scots pine (*Pinus sylvestris* L.) stands in the High Ebro Basin (northern Spain) using soil attributes. Forestry 74:395–406

Bravo-Oviedo A, Montero G (2008) Descripción de los caracteres culturales de las principales especies forestales de España. In: Serrada R, Montero G, Reque JA (eds) Compendio de selvicultura aplicada en España. INIA & Ministerio de Educación y Ciencia, Madrid, 1178 pp

Bray J, Curtis J (1957) An ordination of upland forest communities of southern Wisconsin. Ecol Monogr 27:325–349

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

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall, New York

Calama R, Montero G (2007) Cone and seed production from stone pine (*Pinus pinea* L.) stands in Central Range (Spain). Eur J For Res 126:23–35. doi:10.1007/s10342-005-0100-8

Calama R, Mutke S, Tomé J, Gordo J, Montero G, Tomé M (2011) Modelling spatial and temporal variability in a zero-inflated variable: the case of stone pine (*Pinus pinea* L.) cone production. Ecol Modell 222:606–618. doi:10.1016/j.ecolmodel.2010.09.020

Cañellas I, García FM, Montero G (2000) Silviculture and dynamics of *Pinus sylvestris*. Investig Agrar Sist y Recur For. Fuera de serie: 233–253

Casini M, Cardinale M, Arrhenius F (2004) Feeding preferences of herring (*Clupea harengus*) and sprat (*Sprattus sprattus*) in the southern Baltic Sea. ICES J Mar Sci 61:1267–1277. doi:10.1016/j.icesjms.2003.12.011

Catovsky S, Bazzaz F (2002) Feedbacks between canopy composition and seedling regeneration in mixed conifer broad-leaved forests. Oikos 98:403–420

Cayuela L, Golicher DJ, Benayas JMR, González-Espinosa M, Ramírez-Marcial N (2006) Fragmentation, disturbance and tree diversity conservation in tropical montane forests. J Appl Ecol 43:1172–1181. doi:10.1111/j.1365-2664.2006.01217.x

Christopoulou A, Fyllas NM, Andriopoulos P, Koutsias N, Dimitrakopoulos P, Arianoutsou M (2014) Post-fire regeneration patterns of *Pinus nigra* in a recently burned area in Mount Taygetos, Southern Greece: the role of unburned forest patches. For Ecol Manag 327:148–156. doi:10.1016/j.foreco.2014.05.006

R Core Team (2013) R: a language and environment for statistical computing

IBM Corp (2012) IBM SPSS Statistics for Windows

Crotteau JS, Ritchie MW, Varner JM (2014) A mixed-effects heterogeneous negative binomial. For Sci 60:275–287

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess K, Gibson J, Lawler J (2007) Random forests for classification in ecology. Ecology 88: 2783–2792. doi:10.1890/07-0539.1

Del Cerro A, Lucas ME, Martínez E, López FR, Andrés M, García FA, Navarro R (2009) Influence of stand density and soil treatment on the Spanish Black Pine (*Pinus nigra* Arn. ssp. *salzmannii*) regeneration in Spain. Investig Agrar Sist y Recur For 18:167–180

Domenich T, McFadden D (1975) Urban travel demand: a behavioural approach. North-Holland Publishing Co, Amsterdam, 215 pp

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813. doi:10.1111/j.1365-2656.2008.01390.x

Fei S (2010) Applying hotspot detection methods in forestry: a case study of chestnut oak regeneration. Int J For Res 2010:1–8. doi:10.1155/2010/815292

Fisher R (1918) The correlation between relatives on the supposition of Mendelian Inheritance. Philos Trans R Soc Edinburgh 52:399–433. doi:10.1017/S0080456800012163

Forteza J, Lorenzo L, Najac N, Cuadrado S, Ingelmo F, Hernández J, Prat L, Muñez MC, Macarro MC, Rivas MD, García A (1988) Mapa de suelos de Castilla y León, escala 1:500 000. Consejería de Fomento, Junta de Castilla y León, Valladolid

Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, Nielsen A, Sibert J (2012) AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim Methods Softw 27:233–249. doi:10.1080/10556788.2011.597854

Fyllas NM, Dimitrakopoulos PG, Troumbis AY (2008) Regeneration dynamics of a mixed Mediterranean pine forest in the absence of fire. For Ecol Manag 256:1552–1559. doi:10.1016/j.foreco.2008.06.046

García López J (1994) Short description of the Navafría pine forest and its management history. In: Montero G, Elena R (eds) Mountain silviculture. Investigación Agraria: Sistemas y recursos forestales. Fuera de serie n°3, pp 309–320

Gómez-Aparicio L, Valladares F, Zamora R (2006) Differential light responses of Mediterranean tree saplings: linking ecophysiology with regeneration niche in four co-occurring species. Tree Physiol 26: 947–958

González-Martínez SC, Bravo F (2001) Density and population structure of the natural regeneration of Scots pine (*Pinus sylvestris* L.) in the High Ebro Basin (Northern Spain). Ann For Sci 58:277–288. doi:10.1051/forest:2001126

Grömping U (2009) Variable importance assessment in regression: linear regression versus random forest. Am Stat 63:308–319. doi:10.1198/tast.2009.08199

Hardin J, Hilbe J (2012) Generalized linear models and extensions, third edit. Strata Press, Lake Drive

Henderson C, Kempthorne O, Searle S, von Krosigk C (1959) The estimation of environmental and genetic trends from records subject to culling. Biometrics 15:192–218. doi:10.2307/2527669

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15:651–674

Hyppönen M, Hallikainen V, Niemelä J, Rautio P (2013) The contradictory role of understory vegetation on the success of Scots pine regeneration. Silva Fenn 47:1–19

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R. doi: 10.1007/978-1-4614-7138-71

Jolliffe I (2002) Principal component analysis. Springer, New York

Kass G (1980) An exploratory technique for investigating large quantities of categorical data. J R Stat Soc 29:119–127

Kleppin L, Pesch R, Schröder W (2008) CHAID Models on boundary conditions of metal accumulation in mosses collected in Germany in 1990, 1995 and 2000. Atmos Environ 42:5220–5231. doi:10.1016/j.atmosenv.2008.02.058

Kuuluvainen T, Pukkala T (1989) Simulation of within-tree and between-tree shading of direct-radiation in a forest canopy: effect of crown shape and sun elevation. Ecol Modell 49:89–100

Lawless JF (1987) Negative binomial and mixed Poisson regression. Can J Stat 15:209–225

Ledo A, Cayuela L, Manso R, Condés S (2015) Recruitment patterns and potential mechanisms of community assembly in an Andean Cloud Forest. J Veg Sci. doi:10.1111/jvs.12287

Legendre P, Fortin M-J (1989) Spatial pattern and ecological analysis. Vegetatio 80:107–138

Legendre P, Legendre L (1998) Numerical ecology, 2nd edn. Elsevier, Amsterdam

Loftis DL (1990) A shelterwood method for regenerating Red Oak in the southern Appalachians. For Sci 36:917–929

Louviere J, Hensher A, Swait D (2000) Stated choice methods. Cambridge University Press, New York

Lucas-Borja ME (2014) Climate change and forest natural regeneration in Mediterranean mountain areas. For Res 3:2–3. doi:10.4172/2168-9776.1000e108

Lucas-Borja ME, Fidalgo Fonseca T, Lousada JL et al (2012) Natural regeneration of Spanish black pine [*Pinus nigra* Arn. ssp. salzmannii (Dunal) Franco] at contrasting altitudes in a Mediterranean mountain area. Ecol Res 27:913–921. doi:10.1007/s11284-012-0969-x

Manso R, Calama R, Madrigal G, Pardos M (2013) A silviculture-oriented spatio-temporal model for germination in *Pinus pinea* L. in the Spanish Northern Plateau based on a direct seeding experiment. Eur J For Res 132:969–982. doi:10.1007/s10342-013-0724-z

Maydeu-Olivares A, García-Forero C (2010) Goodness-of-fit testing. Int Encycl Educ 7:190–196

Montes F, Cañellas I (2007) The spatial relationship between post-crop remaining trees and the establishment of saplings in *Pinus sylvestris* stands in Spain. Appl Veg Sci 10:151. doi:10.1658/1402-2001(2007)10[151:TSRBPR]2.0.CO;2

Montes F, Pita P, Rubio A, Cañellas I (2007) Leaf area index estimation in mountain even-aged *Pinus sylvestris* L. stands from hemispherical photographs. Agric For Meteorol 145:215–228. doi:10.1016/j.agrformet.2007.04.017

Montgomery R (2004) Effects of understory foliage on patterns of light attenuation near the forest floor. Biotropica 36:33–39

Oksanen J (2013) Multivariate analysis of ecological communities in R: vegan tutorial

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens, MHH, Wagner H (2013) vegan: Community Ecology Package. vegan Community Ecol Packag

Osem Y, Yavlovich H, Zecharia N, Atzmon N, Moshe Y, Schiller G (2013) Fire-free natural regeneration in water limited *Pinus halepensis* forests: a silvicultural approach. Eur J For Res 132: 679–690. doi:10.1007/s10342-013-0704-3

Paluch JG (2005) The influence of the spatial pattern of trees on forest floor vegetation and silver fir (*Abies alba* Mill.) regeneration in uneven-aged forests. For Ecol Manag 205:283–298. doi:10.1016/j.foreco.2004.10.010

Pardos M, Montes F, Aranda I, Cañellas I (2007) Influence of environmental conditions on germinant survival and diversity of Scots pine (*Pinus sylvestris* L.) in central Spain. Eur J For Res 126:37–47. doi:10.1007/s10342-005-0090-6

Pausas JG, Ribeiro E, Vallejo R (2004) Post-fire regeneration variability of *Pinus halepensis* in the eastern Iberian Peninsula. For Ecol Manag 203:251–259. doi:10.1016/j.foreco.2004.07.061

Perreault W, Barksdale H (1980) A model-free approach for analysis of complex contingency data in survey research. J Mark Res 17:503–515

Prévosto B, Amandier L, Quesney T, de Boisgelin G, Ripert C (2012) Regenerating mature Aleppo pine stands in fire-free conditions: site preparation treatments matter. For Ecol Manag 282:70–77. doi:10.1016/j.foreco.2012.06.043

Quinn G, Keough M (2001) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge

Rathbun SL, Fei S (2006) A spatial zero-inflated Poisson regression model for oak regeneration. Environ Ecol Stat 13:409–426. doi:10.1007/s10651-006-0020-x

Revelle W (2013) psych: Procedures for Personality and Psychological Research

Shepard R (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27:125–140

Spanos IA, Daskalakou EN, Thanos CA (2000) Postfire, natural regeneration of *Pinus brutia* forests in Thasos island, Greece. Acta Oecol 21:13–20. doi:10.1016/S1146-609X(00)00107-7

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307. doi:10.1186/1471-2105-9-307

Strobl C, Malley J, Tutz G (2009) An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol Methods 14:323–348. doi:10.1037/a0016973.An

Ten Have TR, Kunselman AR, Tran L (1999) A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. Stat Med 18:947–960

Valkonen S (2000) Effect of retained Scots pine trees on regeneration, growth, form and yield of forest stands. Investig Agrar Sist y Recur For. Fuera de serie:121–145

Van Diepen M, Franses PH (2006) Evaluating chi-squared automatic interaction detection. Inf Syst 31:814–831. doi:10.1016/j.is.2005.03.002

Venables W, Ripley R (2002) Modern applied statistics with S, 4th edn. Springer, New York

Vriens M (2001) Market segmentation: analytical developments and application guidelines. Millward Brown Intelli-Quest. Techn Overview Ser

Woods KD, Acer K (1984) Patterns of tree replacement : canopy effects on understory pattern in hemlock—northern hardwood forests. Vegetatio 56:87–107

Wu H-I, Sharpe PJH, Walker J, Penridge LK (1985) Ecological field theory: a spatial analysis of resource interference among plants. Ecol Modell 29:215–243. doi:10.1016/0304-3800(85)90054-7

Zuur AF, Ieno EN, Smith G (2007) Analyzing ecological data. Springer, New York

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York, 574 pp

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14. doi:10.1111/j.2041-210X.2009.00001.x