Challenges in modelling the abundance of 105 tree species in eastern North America using climate, edaphic, and topographic variables

Dominic Chambers [a], Catherine Périé [b], Nicolas Casajus [c], Sylvie de Blois [a,*]

[a] McGill School of Environment and Department of Plant Science, Macdonald Campus of McGill University, 21111 Lakeshore, Sainte-Anne-de-Bellevue, Canada H9X 3V9
[b] Direction de la recherche forestière, Ministère des Ressources naturelles et de la Faune, 2700 rue Einstein, Quebec City, Canada G1P 3W8
[c] Conservation of Northern Ecosystems and Centre of Nordic Studies, Quebec University at Rimouski, 300 allée des Ursulines, Rimouski, Canada G5L 3A1

ABSTRACT

Improving predictions of the location of suitable environmental conditions for species using species distribution models (SDM) is at the core of biodiversity/climate change research, but modelling species abundance, rather than distribution, is proving particularly challenging. Using data from more than 200,000 forest plots in eastern North America and Random Forest, we evaluated the performance of species abundance models (SAM) in predicting the relative abundance (measured as importance value) of each of 105 tree species in relation to climate, edaphic, and topographic variables. We calculated the coefficient of determination $\left(R_{SAM}^2\right)$ between observed and predicted abundances as a measure of model performance for each species. We also performed multiple linear regressions to explain variation of R_{SAM}^2 among species using five biogeographical or spatial attributes of species as explanatory variables. Predictive performances of SAM $\left(R_{SAM}^2\right)$ were generally low, ranging from 0.016 to 0.815 (mean = 0.258). Black spruce (*Picea mariana*) had the best predictive model and Florida maple (*Acer barbatum*) and American chestnut (*Castanea dentata*) the worst. Thirty-seven of the 41 best performing species $\left(R_{SAM}^2 \geqslant 0.3\right)$ had climate ranked as the best and/or second best predictor. Species with the best performance tended to be those that could reach dominance, showed aggregated distribution of abundance, and/or had high latitudinal limits in the study area. Climate change is likely to affect patterns of dominance in communities by altering patterns of co-occurrences, but for many species that constitute the bulk of tree diversity, predictions based solely on the current distribution of relative abundances may not be reliable enough to inform conservation or management decisions. Predicting tree abundance in a warming climate using SAM remains a challenge, but it is only by reporting performances across a range of climate and statistical models, regions and species, as well as by highlighting model limitations and strengths, that we will improve the reliability of predictions and in turn better inform forest conservation and management.

1. Introduction

The need to forecast potential changes in species distribution in response to climate warming and other environmental changes has stimulated the development of modelling approaches that use current geographic distribution to estimate suitable environmental conditions for a given species (Araújo et al., 2004; McKenney et al., 2007, 2011; Thuiller, 2003). Based on occurrence data, species distribution models (SDM) have generally reported moderate to high accuracy (Area Under the ROC Curve ⩾0.60) in reproducing the current distribution of trees (McKenney et al., 2007), birds (Brotons et al., 2004; Hu et al., 2010), or amphibians (Lawler et al., 2010). This has increased confidence that SDM can be useful for mapping biodiversity, assessing climate risk, or informing resource management (Millar et al., 2007). Improving confidence in predictions however, requires an understanding of the sources of uncertainty in SDM (Buisson et al., 2010). One of these sources is the modelling approach used; this has led to the development of consensus approaches where predictions are combined across a range of statistical models (Araújo and New, 2007). When different species are modelled using the same statistical approach, differences in accuracy among predictions for the same set of predictors have been attributed to the spatial, geographical, and/or biological attributes of the species themselves (Hanspach et al., 2011; Luoto et al., 2005; Marmion et al., 2009; Syphard and Franklin, 2010). The fact that SDM have been often criticised for not implicitly taking into account processes such as dispersal has not limited their development as an active area of research in ecology, conservation, and management.

* Corresponding author.
  E-mail address: sylvie.deblois@mcgill.ca (S. de Blois).

Improving predictions of the potential location of species is important, but modelling species abundance, rather than occurrence, may prove particularly challenging (Anadón et al., 2010; Iverson et al., 2008; Pearce and Ferrier, 2001). Yet knowing about the relative abundance of a species in a community or landscape is crucial as abundance patterns determine a species' influence on other species or ecosystem processes, its conservation status, or its use as a resource. Information on the presence of a species at a site may mask patterns of low abundance or, on the contrary, patterns of very high abundance in populations of core importance to the species. In spite of recognising the importance of monitoring changes in abundance patterns for biodiversity research (McMahon et al., 2011), there are still relatively few studies addressing patterns of species abundances over broad geographical extent in relation to climate and other environmental gradients, and the available empirical observations show that the spatial distribution of abundances is more complex than previously acknowledged (Sagarin et al., 2006). Climatic suitability, as determined by SDM, positively correlates with abundance for some species (VanDerWal et al., 2009), suggesting that processes like dispersal or physiological constraints on establishment that determine a species' occurrence at a site also determine to some extent its abundance. Species abundance models (hereafter called SAM), however, have lagged behind SDM, partly because of the scarcity of good abundance data across large spatial extent to allow testing environmental drivers of species' abundance for a range of species.

When SAM have been used with climate data, they have produced mixed results. SAM have been apparently successful in modelling bird abundance in Africa in relation to climatic variables (Huntley et al., 2011), but less so for trees in North America (Canham and Thomas, 2010; Iverson et al., 2008); their performance may vary with taxa or regions. For trees, failure to incorporate edaphic variables along with climatic variables or failure to capture the northern limit of species (Canham and Thomas, 2010), or the lack of confidence in models built with only a portion of the species' range (Iverson et al., 2008), have all been invoked to explain poor performance of abundance models. It remains unclear whether predictive performance would be improved by extending datasets to include latitudinal limits or more environmental predictors. More importantly, it is unclear whether or how tree abundance patterns vary with climate at broad spatial scale (Canham and Thomas, 2010; Stegen et al., 2011). For animal species, examination of the relationship between climatic suitability predicted from presence/absence data and abundance showed that sites with low suitability consistently had low abundance, whereas suitability predicted the upper limit of abundance better than mean abundance (VanDerWal et al., 2009). Community processes such as competition can reduce species abundance at a site, but ecological theory also predicts that environmental conditions will have an impact on competitive outcomes and therefore on species abundance. All this suggests that species that can reach dominance in part of their range would be better modelled by SAM. There have been generally few attempts at explaining modelling variations in the predictive performance of SAM to identify the attributes of species that perform best or, conversely, of those that perform poorly.

To understand the relationship between tree abundance patterns and environmental predictors in a context of environmental changes, we tested how well we could predict the relative abundance of 105 tree species using abundance data from more than 200,000 forest plots across eastern North America in relation to climatic, edaphic, and topographic variables. We integrated ecological information across eastern Canada and the USA, thereby capturing the northern limit of tree species. We used the statistical modelling approach, Random Forest, that consistently performed best in exploratory analyses using a range of statistical models.

Finally, we went a step further by explaining variations in SAM performances using biogeographical and spatial attributes of species as explanatory variables to facilitate generalizations across species and regions.

2. Materials and methods

2.1. Study area

This study is part of a larger research project known as CC-Bio for "Effect of Climate Change on Quebec Biodiversity" (Berteaux et al., 2010). The focus of CC-Bio is on the impacts of climate change on the biodiversity of the province of Quebec, Canada. The extent of the study area, however, was defined to take into account the expected shift north of climatic envelope ($\sim$ +5 °C isotherm) as well as available data on species, climate, and edaphic conditions. The study area ranges in latitude from 30°30′N to 53°00′N and in longitude from 93°00′W to 60°30′W covering an area of more than 2,567,000 km² (Fig. 1).

Because of its size, the study area is characterized by strong climatic, geological, and topographical variability. It contains several climatic zones ranging from temperate in most of its range, to arid in the southwest of the Great Basin, to a maritime climate along the Atlantic coast. Annual mean temperature increases gradually from −5 °C in the north to 20 °C in the south, whereas annual total precipitation ranges from 670 to 2000 mm (USDA Forest Service, 2010).

Physiographic regions include the Canadian Shield (Laurentian Upland, North), the Great Lake plains, the Appalachian Highlands (North East), the Atlantic Plains (East), and the Interior Plains (West). The elevation ranges from sea level to 1250 m (mean = 300 m, SD = 165 m) (Canadian Council on Geomatics, 2003; USGS, 2010).

The study area was covered by a grid that contains 6418 cells of 400 km² (20 km × 20 km) each. Each cell was considered as a sampling unit and contained information on tree abundance, climatic (temperature and precipitation), topographic (elevation) and edaphic variables (drainage and surface deposit). The grid in Canada was a convenient extension of the grid used by Prasad et al. (2006) in eastern USA, making it possible to merge datasets from different jurisdictions (see details below). Although climatic data can be obtained at finer or coarser scales, the cell size that we used was considered adequate to capture broad climatic variations over such a large study area.

2.2. Data and analysis

There were two main steps in the analysis (Fig. 2):

- Step (1) model tree abundance, measured as importance value in a cell (the response variable), for each species (n = 105) using climatic, edaphic, and topographic data as explanatory variables. Calculate a measure of predictive performance $\left(R^2_{\text{SAM}}\right)$ for each modelled species.
- Step (2) explain the variation among species in the predictive performance of SAM by conducting a Multiple Linear Regression (MLR) using R^2_{SAM} from step 1 as the response variable and five biogeographical and spatial attributes describing each species as explanatory variables.

2.2.1. Step 1: model tree abundance for each species
2.2.1.1. Data collection.
2.2.1.1.1. Tree importance value. The importance value (IV) was used as a measure of the relative density and dominance of each tree species (n = 105). In the American portion of the study area, IV for a given species was obtained online from the USDA Forest