Ecological Indicators 26 (2013) 1–4

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Short communication

Inferring nitrogen deposition from plant community composition

Richard J. Payne[a,*], Simon J.M. Caporn[a], Carly J. Stevens[b,c], Jacky A. Carroll[a,d], Jill L. Edmondson[a,e], David J. Gowing[c], Nancy B. Dise[a]

[a] School of Science and the Environment, Manchester Metropolitan University, Chester St., Manchester M1 5GD, United Kingdom
[b] Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, United Kingdom
[c] Department of Earth, Environment and Ecosystems, Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom
[d] Penny Anderson Associates Ltd., Park Lea, 60 Park Road, Buxton, Derbyshire SK17 6SN, United Kingdom
[e] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

ARTICLE INFO

Article history:
Received 27 April 2012
Received in revised form 7 August 2012
Accepted 17 October 2012

Keywords:
Nitrogen
Bioindication
Pollution
Ellenberg indices
Transfer functions
Biodiversity

ABSTRACT

Chronically elevated reactive nitrogen deposition has a severe impact on many ecosystems, and there is widespread interest in the possibility of using plant community composition to estimate the level of nitrogen deposition and consequent impacts. Existing approaches use a variety of simple measures including functional type ratios, Ellenberg numbers, and diversity indices. We propose an alternative approach in which species–environment models are constructed using national datasets designed to capture broad-scale deposition patterns. We construct models using partial least squares, weighted average, and maximum likelihood Gaussian logit regression for two British semi-natural habitats, and test how well they predict N deposition by cross-validation. We find that performance is good with R^2 values up to 0.7, and suggest that such models could be a useful addition to the bioindication toolbox.

1. Introduction

Deposition of reactive nitrogen (N) largely derived from intensive agricultural and industrial activity is an increasingly urgent conservation concern. A wealth of evidence links N deposition to loss of biodiversity, plant community change and degradation of ecosystem services (Bobbink et al., 1998, 2010). In developing countries N deposition is increasing rapidly and constitutes a clear threat to biodiversity hot-spots and protected areas (Bleeker et al., 2011; Phoenix et al., 2006). While N deposition is stabilising, or even falling in much of the industrialised world, ecosystems carry a legacy of past deposition which will not be quickly reversed. National- and international-scale models of N deposition (Jonson et al., 1998; Smith et al., 2000; Fagerli and Aas, 2008) represent the large-scale distribution of pollution reasonably accurately, but cannot show the local-scale impacts of point sources such as individual industrial or agricultural units. It is these local-scale impacts which are usually the concern of practical conservation management, where the interest is often in the impacts of a specific polluter on an individual designated site. Protected areas are preferentially located in topographically complex regions (Joppa and Pfaff, 2009) where large-scale deposition models perform less effectively (Sutton et al., 2004). Considerable research attention has therefore focussed on the identification of bioindicator approaches. Bioindicators can be used to identify both the level of pollution and the impacts of pollution exposure, although this distinction is rarely made explicit. Among the many approaches to bioindication of nitrogen pollution (Sutton et al., 2004), studies have investigated the potential of plant community-based bioindicators using the occurrence or abundance of indicator species, or derived indices such as Ellenberg values, plant functional type ratios and diversity measures (Pitcairn et al., 2002, 2003; Stevens et al., 2009). Although results have often shown significant relationships with N deposition, the strength of this relationship is variable.

We propose an alternative concept in which relationships between plant species composition and N deposition are modelled using national vegetation datasets and N deposition models. These species–environment models can be calibrated using sites better suited to national-scale models (away from point sources and complex topography), and then applied to predict deposition in situations where national-scale deposition models are less appropriate. The approach essentially uses vegetation–N relationships to down-scale national deposition models. In this paper we test the concept that the vegetation composition of a set of plant communities that fulfil certain criteria can be used to develop models to predict the cumulative N deposition at other sites comprising a similar vegetation type.

* Corresponding author. Tel.: +44 0161 2476198.
E-mail address: r.payne@mmu.ac.uk (R.J. Payne).

1470-160X/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2012.10.013

Table 1
Model performance for grassland and heathland data showing model structure and R^2, root mean squared error of prediction (RMSEP) and maximum bias (Max Bias) determined by jack-knife cross-validation. Model results illustrated in Fig. 1 marked '*'.

	R^2_{jack}	$RMSEP_{jack}$	Max $Bias_{jack}$
Grasslands			
PLS (3 component)	0.43	548	623
ML*	0.64	464	1440
WA (inverse deshrinking)	0.46	526	677
Heathlands			
PLS (1 component)*	0.70	367	695
ML	0.62	456	586
WA (inverse deshrinking)	0.61	421	759

2. Materials and methods

We model the relationship between species abundance and N deposition for two semi-natural UK vegetation types using three alternative regression approaches. We use UK vegetation datasets of average species cover for acid grasslands (%), and frequency (occurrence per quadrat) for heather moorlands. The acid grassland dataset encompasses 64 sites of UK National Vegetation Classification (NVC: Rodwell, 1992) type U4, (*Festuca ovina–Agrostis capillaris–Galium saxatile* grassland) sampled in 2002–3 (Stevens et al., 2004, 2006). The heathlands dataset (NVC type H12, *Calluna vulgaris–Vaccinium myrtillus* heath) combines the data of Edmondson et al. (2010) and Caporn et al. (2009) giving 36 sites sampled in 2005 and 2006. While the grasslands data includes all plant species the heathlands data includes bryophytes alone. All studies used five quadrats per site.

In both datasets, N deposition is the strongest environmental correlate with community composition, and appears to be a key agent of vegetation change (Payne et al., 2011, unpublished). We use modelled cumulative N deposition between 1900 and the year of sampling rather than modelled current N deposition, reflecting an increasing weight of opinion that cumulative deposition better represents how ecosystems respond to N pollution (Duprè et al., 2010; De Schrijver et al., 2011). We apply the scaling factors of Fowler et al. (2004) to output from the Centre for Ecology and Hydrology CBED model (Smith et al., 2000) to calculate cumulative nitrogen deposition since 1900 on a 5 km × 5 km grid basis. The cumulative N deposition range was 430–2856 kg N ha^{-1} (mean: 1742, sd: 720) for the grassland sites and 459–3067 kg N ha^{-1} (mean: 1886, sd: 671) for the heathlands.

We test three regression techniques based on two contrasting concepts of how species abundance may respond to N deposition. The simplest concept assumes that species respond linearly to N deposition: an increase in N deposition produces an increase or decline in each species. Multiple linear regression performs poorly for ecological data with large number of species whose abundances are strongly correlated (e.g. Ter Braak and van Dam, 1989). We test an alternative approach: partial least squares (PLS) regression. PLS attempts to extract a minimal number of latent factors or components from a training set which explain the variability in the environmental data (Geladi and Kowalski, 1986). PLS has been applied in several previous ecological studies (e.g. Charman, 1997) and has been used for the bioindication of nitrogen deposition with metabolic finger-print data (Gidman et al., 2006).

The assumption of a linear relationship between species abundance and cumulative N deposition may be valid if impacts are due to direct toxicity, species are at the edge of their environmental tolerances or where there is a limited range of deposition values. However, N is an essential nutrient for plants so an alternative hypothesis is that, for many species, small inputs may be beneficial but larger additions deleterious, producing a unimodal response. We therefore also test two regression techniques which assume a unimodal response of species abundance to N deposition.

In maximum likelihood (ML) Gaussian logit regression, the relationship between an environmental variable and abundance of each species is modelled as a Gaussian curve. Maximum likelihood estimation is used to determine the value of the environmental variable with the highest probability of being associated with a particular community composition; this estimate is the model prediction (for details of computation see Birks, 1995). The method has been shown to perform well with simulated data and real ecological datasets (Ter Braak and Looman, 1986), but it is relatively complex and computationally intensive.

An alternative unimodal technique is weighted average (WA) regression, in which it is assumed that a species will be most abundant in a site with environmental conditions close to the species optimum; a reasonable approximation of the species optimum is therefore made by calculating the average environmental values of all the sites in which the species occurs, weighted by the abundance of the species in those sites. An estimate of the environmental variable for an unknown site is provided by a weighted average of the optima of all species present. As this procedure serves to compress the environmental gradient a de-shrinking regression is applied to remove this compression. WA is less statistically rigorous than ML but is computationally simpler and often has superior performance in practise (Birks et al., 1990; Ter Braak and van Dam, 1989).

Assumptions of all these models include the independence of samples, lack of confounding secondary gradients and the presence of a direct (or indirect but linear) relationship between the species and the environmental variable of interest. These assumptions, and the consequences of their violation, are discussed in greater depth by Birks (1990, 1995, 1998), Belyea (2007) and Ter Braak and Prentice (1988).

We applied all three techniques (PLS, ML and WA) to both of the vegetation datasets. We assessed model performance statistically by applying the model to the same dataset used to construct the model. To avoid overly optimistic estimates of performance if the same data are used to both build and test models we used jack-knife ('leave-one-out') cross-validation in which models are successively constructed using $n-1$ samples with the remaining sample serving as a test. Performance statistics used are the R^2 between observed and predicted values, the root mean squared error of prediction (RMSEP) and the Maximum Bias (cross-validated values are denoted R^2_{jack}, $RMSEP_{jack}$ and Max $Bias_{jack}$). These three measures provide distinct but complementary information about the performance of models: R^2 gives a measure of the overall strength of relationship between observed and predicted values, RMSEP gives a measure of average errors along the gradient and Maximum Bias gives a measure of maximum mean error for any one tenth of the gradient. Models were developed using C^2 (Juggins, 2003).

3. Results

All models produced RMSEP values below the standard deviation of the nitrogen deposition data, so all can be considered to have predictive power despite the limited size of the training sets. RMSEP values suggest that these models may be able to predict cumulative N deposition with a mean error as low as 367 kg N ha^{-1}, 15% of the range captured by the vegetation dataset (Table 1).

For the acid grasslands data the best-performing model in terms of R^2 and RMSEP is ML and in terms of maximum bias is a 3-component PLS model. For the heathlands data the best-performing model is a single-component PLS model for R^2 and RMSEP and ML for maximum bias. Model performance with the heathland data is superior to that of the grasslands, despite the smaller dataset size