
Biodiversity and Conservation **11**: 2189–2216, 2002.

Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions

ANA M.F. BIO[1,*], PIET DE BECKER[2], ELS DE BIE[2], WILLY HUYBRECHTS[2] and MARTIN WASSEN[3]

[1] Environmental Group of the Centre for Modelling Petroleum Reservoirs, CMRP/IST, Tecnical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; [2] Institute of Nature Conservation, Kliniekstraat 25, 1070 Brussel, Belgium; [3] Department of Environmental Sciences, Faculty of Geographical Sciences, Utrecht University, P.O. Box 80115, 3508 TC Utrecht, The Netherlands; [*] Author for correspondence (e-mail: anabio@ist.utl.pt; fax +351-21-8417389)

Received 24 November 2001; accepted in revised form 20 March 2002

Abstract. In ecological modelling, limitations in data and their applicability for predictive modelling are more rule than exception. Often modelling has to be performed on sub-optimal data, as explicit and controlled collection of (more) appropriate data would not be feasible. An example of predictive ecological modelling is given with application of generalized additive and generalized linear models fitted to presence–absence records of plant species and site condition data from four nutrient-poor Flemish lowland valleys. Standard regression procedures are used for modelling, although explanatory and response data do not meet all the assumptions implicit in these procedures. Data were non-randomly collected and are spatially autocorrelated; model residuals retain part of that correlation. The scale of most site-condition records does not match the scale of the response variable (species distribution). Hence, interpolated and up-scaled explanatory variables are used. Data are aggregated from distinct phytogeographical regions to allow for generalized models, applicable to a wider population of river valleys in the same region. Nevertheless, ecologically sound models are obtained, which predict well the distribution of most plant species for the Flemish river valleys considered.

Key words: Autocorrelation, Generalized additive model(ling), Generalized linear model(ling), Phreatophytes, Spatial autocorrelation, Spatial variability

Introduction

Lowland river valleys harbour a large part of biodiversity of the western European lowland (Naiman et al. 1989; Wheeler et al. 1995). They do so because abiotic conditions in these valleys are diverse as a consequence of micro-topography, soil differences and feeding by different water sources: atmospheric water, groundwater and river water (Wassen and Barendregt 1992; De Becker et al. 1999). However, the loss of biodiversity is large in these valleys and, mainly, caused by human interferences that have been leading to deterioration of the valley ecosystems (Petts and Amoros 1996; Rich and Woodruff 1996). Causes for deterioration in valleys of western Europe include: high

atmospheric N-deposition (Erisman and Draaijers 1995); groundwater abstraction before it exfiltrates in the valleys (Schot and Molenaar 1992); prevention of river flooding (Hellberg 1995; Runhaar et al. 1996; Décamps et al. 1988); agricultural drainage systems; application of fertilizers; and pollution of groundwater and surface water by sewage.

Today, the threats are recognized and restoration plans are made, often in the context of integrated water management (Gilbert and Anderson 1998; Nienhuis et al. 1998). For the assessment of the effects of restoration measures hydrological and ecological models have been developed for regional water management authorities and nature conservation organisations. Hydrological models allow us to determine the effects of changes in water management on a certain location in terms of groundwater as well as surface water dynamics and preferably water quality. These models are explicitly spatial, simulating the regional groundwater flow patterns in a catchment (e.g. Batelaan and De Smedt 1994; Batelaan et al. 1995; Van der Aa et al. 2001).

Ecological models link abiotic information (like water availability and quality) to flora and fauna, which depend on these site conditions. Mechanistic ecological models, containing causal relationships derived from experimental studies, are only available for relatively simple and well-studied ecosystems, such as freshwater lakes (Janse et al. 1992; Van Liere and Gulati 1992), and their development is very time consuming and expensive. For the restoration of entire river valleys, generally applicable models valid for a range of ecosystems are required. These ecosystems and their interrelations are so complex that deterministic knowledge is often not available and experimental studies are not feasible. Therefore, empirical models are used to gain insight into these systems and to allow for some form of prediction of environmental and management effects on these systems.

Empirical ecological models are often based on available data that were not explicitly collected for that purpose or on limited data sets especially collected for the purpose of model development (see Witte and Van der Meijden 1992; Barendregt and Nieuwenhuis 1993; De La Ville et al. 1997; Ertsen et al. 1998). In empirical modelling, the functional relationship between a response variable (for instance the presence or absence of a plant species) and one or more explanatory variables (e.g. water table and water quality variables) is generally specified by a regression model (Austin et al. 1994; Bio et al. 1998; Franklin 1998; Guisan and Harrell 2000). Therefore, quantity and quality of data are obviously of utmost importance. An ideal data set for ecological modelling contains enough samples that are representative of and well distributed in the modelled geographical and environmental ranges and that satisfy model assumptions. However, such ideal data sets are rarely found and the urgent need for swift restoration measures presses modellers to do with less than ideal data (see Olde Venterink and Wassen 1997).

Classical statistical inference is based on the assumption of independent observations collected at randomly chosen locations (De Gruijter and Ter Braak 1990). Ecological processes are intuitively spatial and records of spatial dependence in eco-

logical data are numerous (e.g. Rossi et al. 1992; Tilman 1994), as neighbouring samples tend to be more similar than samples further apart. Using standard statistics the presence of spatial autocorrelation in data and in model residuals may render error estimates and associated significance tests unreliable. It may also affect model choice, as variable selection is generally based on explained and residual variance. Nonetheless, these data are generally treated as independent random samples and modelled using classical statistical procedures (e.g. Nicholls 1989; Hill 1991; Buckland and Elston 1993).

Recently, methods have been developed for the modelling of spatial dependence (or autocorrelation) in regression using, for instance, neighbourhood information (Sokal and Oden 1978a,b; Smith 1994; Augustin et al. 1996; Osborne et al. 2001). Geostatistical modelling of residual spatial autocorrelation is an alternative approach in development (Pebesma et al. 2000). However, for prediction at other places or in different conditions, the use of spatial autocorrelation as model term or residual information has serious drawbacks. On the one hand, neighbourhood or other spatial dependence information is not directly available for new data and the assumption that levels of spatial dependence for new sites or conditions are similar to those found at the modelled sites may not be valid. On the other hand, a spatial autocorrelation term in the model will act as an indirect variable accounting for and, most likely, masking part of the effect of several direct, ecologically relevant variables. Just as vegetation records, records of abiotic site conditions tend to be autocorrelated too, and an explanatory variable defining the neighbourhood of a site in terms of a species' occurrence will combine biotic (e.g. species' dispersal ability or inter-species competition) and abiotic (favourable or non-favourable site conditions) information, rendering robust, but less informative and, possibly, less generalizable models. Only part of the spatial autocorrelation in the response variable is likely to be explained by the explanatory variables in the regression model. Assessment of the residual spatial variance can aid model evaluation and highlight shortcomings in explanatory variables or model structure (e.g. Robertson and Freckman 1995; Begg and Reid 1997; Gotway and Stroup 1997; Köhl and Gertner 1997).

In this paper an example of spatial ecological predictive modelling is presented within the limitations imposed by data availability and model purpose given by environmental policymakers. Because results are intended for generalization to a wider population of river valleys in eastern Flanders and to changes in time (e.g. through management), several available data sets were aggregated to cover a wider variety of environmental conditions. The data, collected from 1993 to 1997 in four nutrient-poor Flemish lowland river valleys, consist of presence and absence records for groundwater-dependent plant species and abiotic site conditions describing management, soil, groundwater level and several groundwater chemistry parameters. Biotic data, management and soil were mapped in grids of adjacent regular square cells; groundwater level and chemistry were interpolated from samples collected at a limited number of piezometers within each grid.

Multiple logistic regression modelling is done within the frameworks of generalized linear modelling (GLM) and generalized additive modelling (GAM), using classical regression procedures. The advantage of GLM is that it is parameterised and easily fitted into an environmental management modelling tool. The advantage of GAM is its flexibility and its proximity to the observed data. Model selection takes place using a bi-directional stepwise procedure, starting with a full model. Correlation between explanatory variables is examined. Model evaluation and comparison is based on cross-validation and model discrimination. Spatial autocorrelation in vegetation field records and model residuals is assessed through empirical semivariograms; the residual semivariograms indicate spatial structure not accounted for by the models' explanatory variables.

Materials and methods

Study sites

Data were collected from four different lowland river valleys in Flanders, the northern part of Belgium (Figure 1). All sites are nature reserves managed by a nature conservation organisation. Sites were selected because of their well-preserved, relatively undisturbed and unspoiled abiotic and biotic conditions; a long period (at least 10 years) of constant management; and usually marked hydrological gradients.

The Zwarte Beek site is situated at the western fringe of the Campinian plateau. It comprises an 800 m long section through a narrow valley, situated at approximately 52–56 m above sea level. Zwarte Beek is known for its excellent fen grasslands (mainly *Caricion curto-nigrae*), fen scrub and alder carr. The soil consists of a 7-m thick peat layer, with an abrupt conversion into sandy sediments at the fringes of the valley. The area is intensely fed (ca. 16 l m^{-2} day^{-1}) by nutrient and mineral-poor seepage water. The groundwater table is constant and close to the surface level throughout the year (De Becker and Huybrechts 2000a).

The Vorsdonkbos site is located at the southern fringe of the Demer river valley, approximately 11 m above sea level. This site is a marked seepage zone fed by two

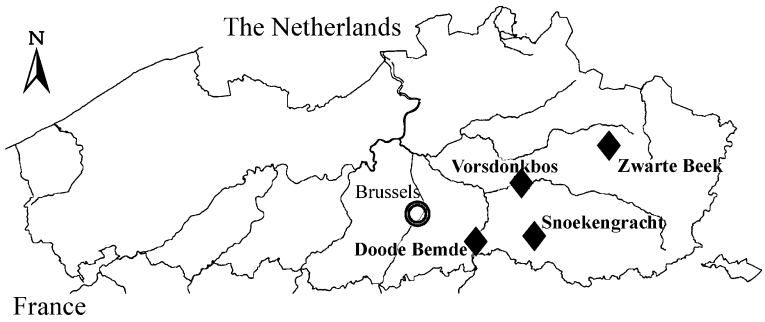

Figure 1. Situation map of Flanders (the northern part of Belgium) with the location of the four study sites.

distinct aquifers. The southern part is extensively supplied (20 l m^{-2} day^{-1}) with mineral-poor groundwater. Here a zone with fragments of fen grasslands (*Caricion curto-nigrae* and *Cirsio-Molinietum*) and oligotrophic woodland (*Sphagno-Alnetum*) is found. In the central and northern part of Vorsdonkbos, which is fed by mineral-rich groundwater, the vegetation changes to fen meadow (*Calthion palustris*), tall herb fen (*Filipendulion*) and mesotrophic alder carr (*Caricion elongatae-Alnetum glutinosae*) (Huybrechts and De Becker 2000).

The Doode Bemde is an alluvial floodplain mire in the valley of the river Dijle, situated at approximately 30 m above sea level. Its soil texture is mainly silt. The area is fed by moderate amounts (3 l m^{-2} day^{-1}) of mineral-rich groundwater. Here, a complete vegetation mosaic is found, ranging from mesotrophic alder carr and reed-beds (Phragmitetalia), over tall sedge swamps (Magnocaricion) and tall herb fen, to fen meadow and somewhat drier Arrhenatheretum grasslands on the natural levees of the river (De Becker and Huybrechts 2000b).

The fourth site, Snoekengracht, situated approximately 57 m above sea level, is similar to the Doode Bemde site, except for a narrower valley and even more mineral-rich seepage water feeding the area (Huybrechts and De Becker 1999).

Mapping of the plant species

During spring and early summer, between 1993 and 1997, plant species were mapped in the four study sites described above on a regular grid of adjacent square cells. Grid cells measured 20 × 20 m, except for the Snoekengracht, where cells measured 10 × 10 m (Table 1). Plant species records were limited to about 85 mainly phreatophyte species (*sensu* Londo 1988). All species that occurred in at least 20 cells in each of the four sample sites were selected for further analysis in this study (Table 2).

Site conditions

Soil type and management regime were recorded during the same period as the plant species (1993–1997) and mapped in the same regular grids. Soil type was derived from hand drillings at grid cell intersections (i.e. one drilling for four grid cells) to a

Table 1. Area, size and number (#) of grid cells and number of piezometers (# Piez.) for the aggregated sample (All) and the four sub-samples.

	Area (ha)	Cells (m)	# Cells	# Piez.
All	78		2587	154
Doode Bemde	21	20 × 20	526	36
Snoekengracht	8	10 × 10	817	36
Vorsdonkbos	25	20 × 20	636	40
Zwarte Beek	24	20 × 20	608	42

Table 2. Frequency of plant species in the aggregated sample (All) and in the four sub-samples.

Species	All ($n = 2587$)	D ($n = 526$)	S ($n = 817$)	V ($n = 636$)	Z ($n = 608$)
Ajuga reptans	375	41	157	106	71
Angelica sylvestris	1382	244	384	380	374
Arrhenatherum elatior	277	123	72	39	43
Caltha palustris	641	70	294	92	185
Carex acuta	326	115	21	71	119
Cirsium palustre	1158	99	213	339	507
Equisetum palustre	643	267	230	55	91
Eupatorium canabinum	431	256	100	43	32
Filipendula ulmaria	1492	356	472	357	307
Galeopsis tetrahit	1011	63	290	366	292
Juncus acutiflorus	611	42	62	101	406
Lotus uliginosus	659	85	69	110	395
Lychnis flos-cuculi	619	97	263	129	130
Lycopus europaeus	1073	200	128	404	341
Lythrum salicaria	983	201	220	287	275
Mentha aquatica	330	106	83	64	77
Phalaris arundinacea	472	166	85	144	77
Scirpus sylvaticus	613	72	135	162	244

D = Doode Bemde, S = Snoekengracht, V = Vorsdonkbos, Z = Zwarte Beek; n = number of records in (sub-)samples.

depth of 1 m and classified using a simplified set of four major texture types: sand, silt, clay and peat. The management regime was classified per grid cell into three categories: yearly mowing in early summer eventually followed by grazing or mowing of the aftermath; cyclic mowing (once every 5–10 years) or no mowing at all during the last 5–10 years; and, no management for at least 10 years.

Groundwater quality and regime variables were determined from samples collected in networks of piezometers (Table 1, see also an example of the distribution of a piezometer network for the site of the Doode Bemde given in Figure 2a). To obtain reliable data on the chemical composition of the groundwater, four sampling tours were carried out in spring and autumn, in two consecutive years during the 1993–1997 period. Ground water pH, K^+, $Fe_{(tot)}$, Mg^{2+}, Ca^{2+}, SO_4^{2-}, Cl^-, NO_3^-, NH_4^+, $H_2PO_4^-$, as well as the ionic ratio (IR $= 100[1/2Ca^{2+}]/[1/2Ca^{2+} + Cl^-]$; see Van Wirdum 1990) were determined for each sampling tour and the results averaged (for details on the analysis procedure see De Becker et al. 1999). In this study, the average groundwater depth (below surface) is the only groundwater regime variable considered for model building. Average groundwater depth was determined for every piezometer, for the duration of 2 years, using fortnightly measurements.

To obtain groundwater-variable estimates for all sample grid cells and on grid cell scale, the averaged estimates at the piezometer point locations were spatially interpolated using block kriging (Cressie 1993; Pebesma 1997; De Becker et al. 2001). Kriging is an interpolation method based on the similarity, or spatial autocorrelation, between sites located at given distances from each other. Interpolation is done follow-

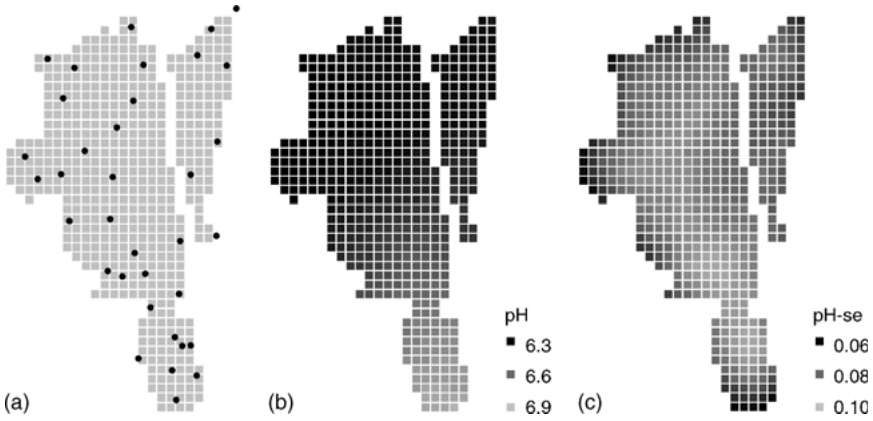

Figure 2. Doode Bemde: sample grid with location of the piezometers (a), kriging estimates for pH (b), and kriging error for pH (c). Grid cells are 20 × 20 m in size.

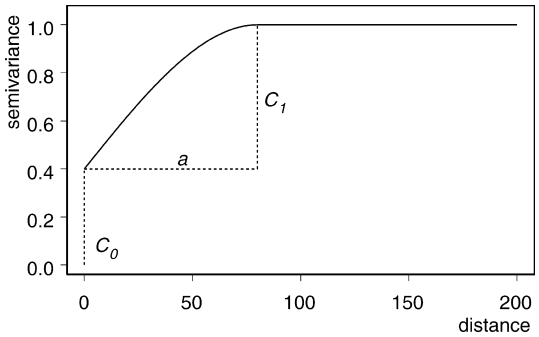

Figure 3. Example of a typical semivariogram (spherical semivariogram model); a = range, C_0 = nugget, C_1 = partial sill, the sill = $C_0 + C_1$.

ing a semivariogram model, fitted to empirical semivariances. Features of a typical semivariogram are presented in Figure 3. A description of semivariograms is given in the Spatial autocorrelation in data and residuals section.

For the same site variable, empirical semivariograms showed different shapes of spatial autocorrelation in the four sample sites. Therefore, spatial analysis and interpolation were done separately for each site and site variable. Semivariances were calculated averaged over all directions, as anisotropy (different form of spatial autocorrelation in different geographical directions) could not be distinguished given the limited number of sample points (piezometers). The models considered for the empirical semivariograms were: linear, spherical, exponential and Gaussian (for a detailed description of the semivariogram models used, see De Becker et al. 2001). They were fitted using weighted least squares estimation (Cressie 1985). Empirical semivariograms and their models were obtained using the S+SpatialStats module (version 1.0, Mathsoft 1996) of the statistical package S-Plus. Site conditions were estimated for the whole grid, per sub-sample, by means of block kriging (using Surfer, Golden

Table 3. Summary of the kriging estimates and standard errors for the water chemistry variables collected at piezometers and interpolated for the four sub-sample grids.

	Estimate			Standard error		
Variable	Minimum	Maximum	Mean	Minimum	Maximum	Mean
Mean	−1.18	0.17	−0.25	0.03	0.30	0.13
PH	5.01	7.29	6.52	0.06	0.51	0.19
Cl	8.33	137.89	29.20	3.32	35.55	13.17
Ca	10.44	221.25	85.79	5.36	36.12	16.19
Fe	0.25	58.61	10.96	2.08	16.00	6.31
K	0.50	10.73	2.04	0.25	1.71	0.70
Mg	1.73	28.92	8.38	0.49	4.10	1.58
NO_3	0.22	8.10	0.72	0.11	2.23	0.96
NH_4	0.13	7.71	0.38	0.10	1.63	0.36
PO_4	0.03	0.88	0.17	0.02	0.26	0.12
SO_4	2.50	162.14	39.12	5.23	49.18	17.72
IR	9.19	80.81	53.43	3.84	11.82	7.18

Software Inc. 1999) with block sizes equal to the respective grid cell sizes (Pebesma 1997). These block estimates were subsequently used as predictor values for regression modelling. Kriging variance was studied to get some insight into the reliability in this part of the model input (Table 3; see also example given in Figure 2).

The estimates for the water quality parameters Ca, Fe, K, Mg, NO_3, NH_4 and SO_4 show skewed distributions. In the present regression modelling, these variables are used both untransformed and taking their logarithms (base 10) which render more normal distributions.

Some of the groundwater chemistry estimates are strongly correlated. Correlation is strongest between Ca and Mg (Spearman's rank coefficient = 0.96), pH and Ca (0.89), pH and Mg (0.86), Cl and PO_4 (0.79), and pH and IR (0.76). Strong correlation between explanatory variables in a regression model may lead to problems of collinearity (or multicollinearity). When an explanatory variable is nearly a linear combination of other explanatory variables in the model, their individual influence on the response variable becomes difficult to distinguish. Coefficient estimates become unstable and show extremely large standard errors and insignificant coefficients may be assigned to relevant predictors (De Veaux and Ungar 1994). Collinearity can be dealt with through *a priori* variable selection or elimination, or combination of correlated variables into new explanatory terms. We chose to use all variables as candidate predictors in the stepwise model selection procedure, to avoid *a priori* elimination of possibly relevant variables and trusting that the stepwise procedure would include highly correlated variables in the same regression equation only if both are significant despite possible collinearity. To assess possible collinearity problems, model coefficients and their errors are checked for irregularities and approximate variance inflation factors (VIF) are calculated for the final 18 regression models (De Veaux and Ungar 1994; Dallal 2001). VIF $= 1/(1 - R^2)$, with a coefficient of determination

(R^2) obtained from the regression of each explanatory variable (the way it appears in the model; i.e. in its linear or quadratic form) against all other explanatory variables present in the model (also using linear or quadratic terms as modelled). A VIF greater than 10 is often considered to point at possible collinearity problems (Dallal 2001). Notice that the computed variance inflation factors are approximate. We used exclusively linear models (with first and second-order model terms) for VIF calculation, as R^2 is not readily available for GAM.

Model selection

The probability of occurrence of each plant species is calculated with respect to the combined effect of site conditions using multiple logistic regression (Hosmer and Lemeshow 1989). Two model frameworks are considered: GLM (Nelder and Wedderburn 1972; McCullagh and Nelder 1989), applied in numerous ecological studies (e.g. Austin et al. 1984; Margules et al. 1987; Zimmermann and Kienast 1999); and GAM (Hastie and Tibshirani 1990; Yee and Mitchell 1991), applied in more recent studies (e.g. De Swart et al. 1994; Huntley et al. 1995; Austin and Meyers 1996; Bio et al. 1998; Austin 1999). Both enable ecologists to model species response to a wide range of environmental data using a link function (here: the logit) between response and predictor variables. GAM form an extension of GLM. While GLM fit functions linear in their parameters, allowing for linear and polynomials response shapes, GAM are more flexible permitting both linear and complex additive response shapes, as well as a combination of the two within the same model. Additive models include a variety of smooth functions (or smoothers) that estimate the response for each abiotic variable – or set of variables – dependent on the responses observed for neighbouring values on the predictor gradient. The response curve is hence more data- than model-driven (Hastie and Tibshirani 1990). However, environmental managers and policymakers have the tendency to favour GLM, because they are parameterised and easily fit into an environmental management modelling tool and because they suggest simple defined forms of generalized relationships between species and their environment.

To simplify the multiple regression models in this study and to enable the inclusion of all available site conditions in the 'full' models used as starting point for the stepwise selection procedure, we restrict each species–predictor response to a curve using a maximum of 2 degrees of freedom (df). Although the GAM's smooth functions are constrained to a complexity comparable to a quadratic function, they show more flexibility (e.g. asymmetry) in response curve shapes than the GLM's polynomial (see example in Figure 4).

Models are built using a stepwise selection procedure to select relevant explanatory variables and the complexity of the species' response shapes to each variable. We opted for a bi-directional stepwise model selection procedure, starting with a full model and alternately omitting and re-introducing one model component at each step (Leathwick 1998; Pearce and Ferrier 2000). Candidate predictor variables are:

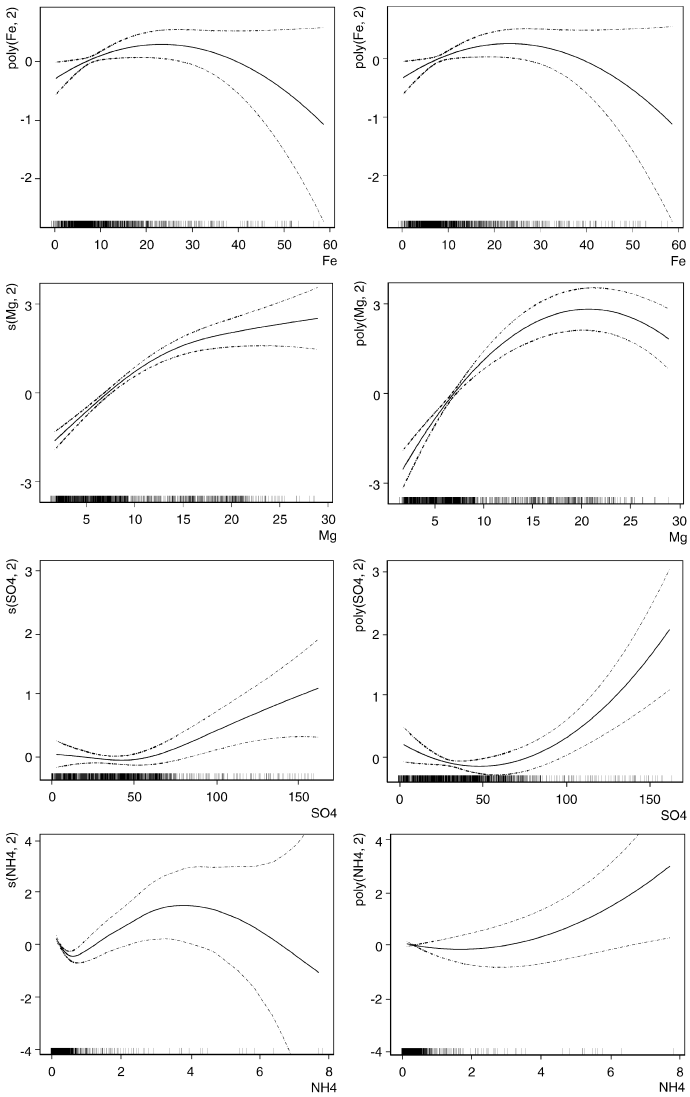

Figure 4. Comparison of smooth (left column) and second-order polynomial response shapes (right column), respectively from GAM and GLM obtained for *Lychnis flos-cuculi*. The dashed lines are approximate 95% point-wise confidence intervals; tickmarks show the location of observations along the variable range; and *y*-axes are scaled to present the partial effect of the respective variable.

management type (treat), soil type (soil), an interaction term between mean groundwater level and soil type (mean:soil), pH, Ca, Cl, Fe, Mg, SO_4, IR, NO_3, NH_4, PO_4 and K. Mean groundwater level is used in interaction with soil, because water levels measured in the piezometers imply different water availability for plants, depending on soil capillarity. Each selection starts with a model containing all abiotic site conditions, with groundwater chemistry variables fitted by a curve using 2 df; i.e. a quadrat-

ic function for the GLM and a cubic smoothing spline of 2 df for the GAM. At each step, each model term is omitted or simplified from either a polynomial (for GLM) or smoothing spline (for GAM) to a linear term, or a new (or previously omitted) candidate predictor term is added.

The model with the lowest value for the Bayesian Information Criteron BIC (Akaike 1978; Schwarz 1978) is kept. Selection stops when no predictor addition or omission would cause a lower BIC value. We opted for BIC as selection criterion, because it showed to be more reliable and parsimonious (though sometimes too strict; Akaike 1977; Buckland et al. 1997; Bio 2000, pp. 27–41) in the selection of relevant variables than the frequently used Akaike's Information Criterion AIC (Akaike 1977; Raftery 1986).

To keep the stepwise procedure workable, four separate selection series take place per plant species: GLM and GAM with untransformed site variables; and GLM_{log} and GAM_{log} using the logarithm for variables with skewed sample distribution. The most accurate of the four models is subsequently retained as the species' final regression model. Accuracy is determined by 10-fold cross-validation (Fielding and Bell 1997). Therefore, data are randomly split into 10 approximately equal-sized groups. Each group is used as an independent validation set, to evaluate the performance of each of the four regression models fitted to the remaining 9/10 of the data. The residual deviance of the predictions on the validation data is computed and averaged for the 10 validation groups. The model with the lowest average cross-validation deviance is selected.

Model evaluation

The four models obtained through stepwise selection for each plant species are compared in terms of discrimination – i.e. how well prediction distinguishes species presence from species absence. Discrimination measures are derived from contingency tables, which are constructed using thresholds based on species' prevalence (i.e. frequency of occurrence). Thresholds are chosen to yield equal proportions of wrongly predicted presences and wrongly predicted absences. For each model, this results in equal measures of sensitivity (proportion of correctly predicted presences), specificity (proportion of correctly predicted absences) and correct classification rate, CCR (Fielding and Bell 1997; Manel et al. 2001). We also determine Cohen's kappa, $\hat{\kappa}$, a discrimination measure for the proportion of correctly predicted presences and absences that accounts for chance effects (Cohen 1960; Hudson and Ramm 1987).

Regional differences

Data were collected from four separate lowland river valleys, differing in their groundwater quality and level, and in soil and management conditions. Each of these subsamples covers different, partially overlapping, ranges of environmental gradients (Figure 5). Previous empirical modelling on the sub-samples separately demonstrated

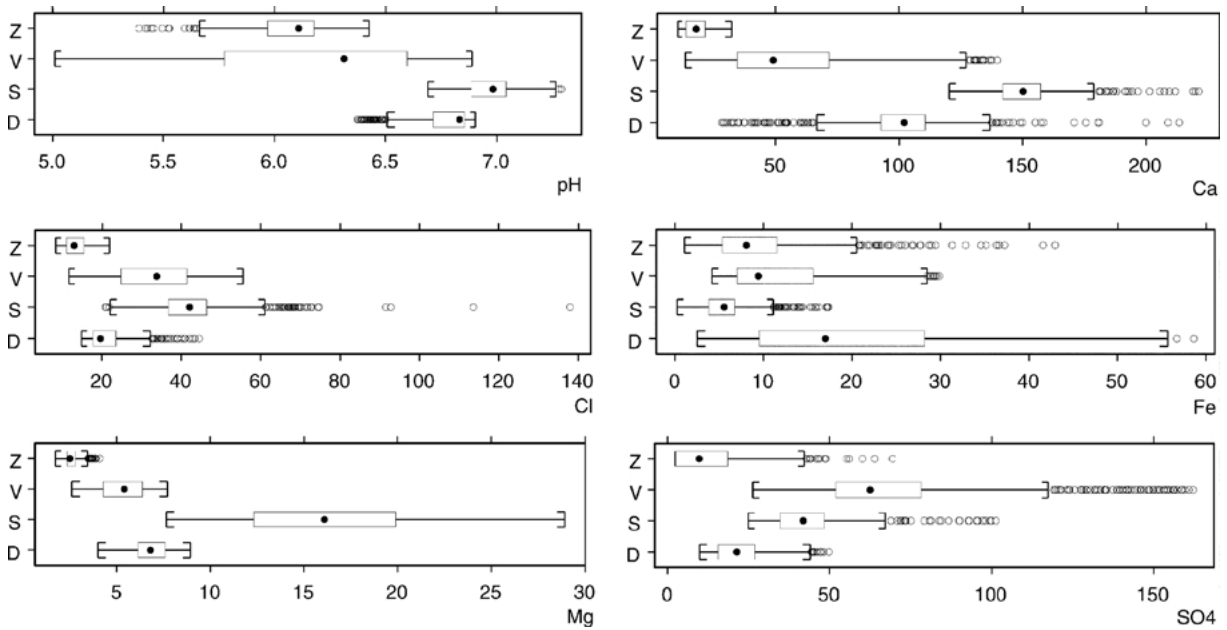

Figure 5. Gradients of the interpolated site condition for the four river valleys composing the data. Box limits are set at quartiles of the ranges; D = Doode Bemde, S = Snoekengracht, V = Vorsdonkbos, Z = Zwarte Beek.

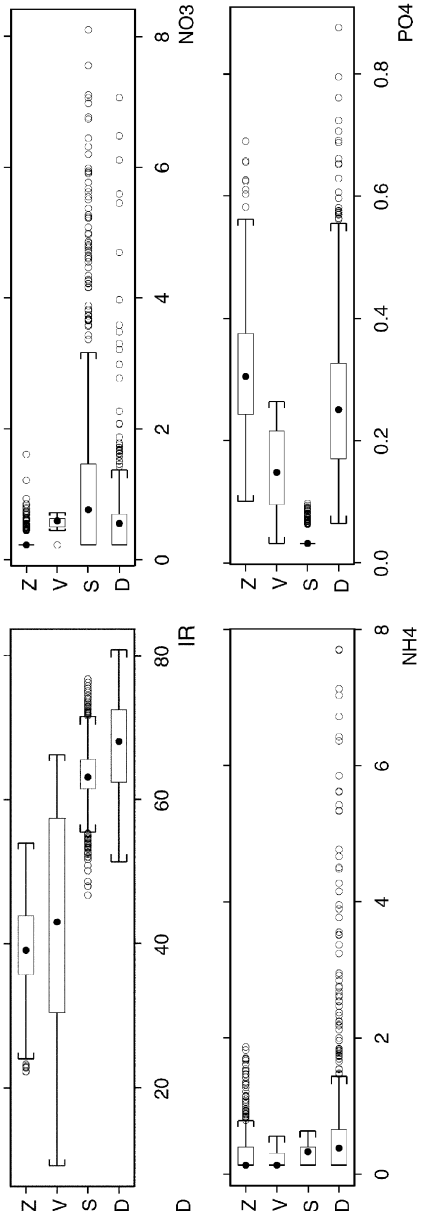

Figure 5. Continued.

that a species' distribution is explained by different site conditions in the different valleys (De Becker et al. 2001). These results can point at differences in species response, different forms of pseudo-correlation (i.e. the species is not directly correlated to the explanatory variables but to other non-modelled variables that are themselves correlated to the explanatory variables), and differences in site condition ranges and, hence, impact. An explanatory variable is significant only if it shows enough variation, a relationship to the response variable and a range that is (at least partly) critical to species' distribution. For different ranges and combinations of site conditions, different variables can become critical for the species and, hence, explain a significant part of species' variation in regression models. Given the aim of this study – the development of models applicable to Flemish (or similar) lowland river valleys in general – the four sub-samples were aggregated into one data set, covering a wide range of site conditions. To assess sub-sample specific variation not accounted for by the empirical models, a factor coding for the sub-samples is added *a posteriori* to the final models and its effect tested.

Spatial autocorrelation in data and residuals

To assess possible spatial autocorrelation in field data and model residuals, their empirical semivariograms are computed. Semivariograms provide information about the spatial autocorrelation between data pairs with increasing distance between them. The average semivariance is calculated for each pair of data within a given distance class and semivariances are subsequently plotted as a function of the classes (or lags) considered. In a typical semivariogram (Figure 3) the semivariance increases with increasing lag distance till it reaches a plateau, the sill. The distance, at which the sill is reached – called range (a) – indicates the maximum distance of spatial autocorrelation between pairs of individual samples. The frequently observed positive value at the origin is called the nugget effect (C_0). It is attributed to the sum of residual, spatially uncorrelated noise (such as measurement error) and spatial variation on spatial scales smaller than the smallest lag considered. The difference between the sill and the nugget is called the partial sill (C_1), or structural or spatial variance (Rossi et al. 1992), and provides a measure for the spatial autocorrelation in data. C_1 corresponds to the amount of the total variance (the sill) that can be attributed to structural spatial autocorrelation (given the lag distances considered). The sill is the sum of C_0 and C_1.

Empirical semivariograms are calculated for the plant species presence–absence records. The resulting semivariograms, here applied to binary data, are called indicator or probabilistic semivariograms (Rossi et al. 1992; Burrough and McDonnell 1998). They reflect the variance in presences and absences at different sample distances. Semivariograms are plotted for a maximum distance of 200 m and lags of multiples of 20 m.

Analogously, semivariograms are calculated for the GLM and GAM residuals. However, in order to meet the stationarity assumptions for semivariogram analysis

(Cressie 1993) we require residuals with a constant variance. Logistic regression residuals have non-constant variances that depend on the predicted values. Therefore, standardised (Pearson) residuals are used (Albert and McShane 1995), obtained by dividing the residuals on the response scale – i.e. the differences between observations (1 or 0) and model predictions – by the square root of their variance. Under correct model assumptions Pearson residuals have zero mean and unit variance, resulting in semivariograms with approximate unit sill.

To enable comparison between the semivariograms obtained from the model residuals and those obtained for the species field records, the latter are computed from Pearson residuals of a model containing an intercept only. The resulting semivariograms have a shape, range and relative partial sill identical to those fitted for the presence and absence records and approximate unit sill under correct model assumptions. The degree of correlation left unexplained by the regression models is assessed comparing the partial sills for field data and model residuals.

Results

Regression models

Most of the final regression models are generalized additive with species' responses fitted by first-order functions and smoothing splines. Only seven species are better modelled by GLM, with first-order and second-order (cubic) response functions (Table 4). Most final models are based on non-transformed explanatory variables. Two species are better described by $\log_{10}$-transformed water chemistry variables; one with a GAM, the other with a GLM. An example of differences between GAM and GLM is given in Figure 4, in which the splines fitted in the GAM for *Lychnis flos-cuculi* are compared to second-order polynomials in a GLM for the same species and explanatory variables. The flexible shape of the cubic smoothing spline, despite of its restriction to 2 df, and its mostly smaller confidence intervals are evident.

The most frequently selected explanatory variables are management type (treat), pH and SO_4 (15 times each), followed by Mg and the interaction term mean:soil (14×), Ca, Fe, NH_4 and ionic ratio (13×), NO_3 (12×), PO_4 (11×), Cl and K (10×) and, finally, soil (6×). Second-order polynomial or smooth responses are dominant for SO_4 (13 out of 15 times), Mg (12 out of 14), Ca (11 out of 13), PO_4 (8 out of 11) and Fe (7 out of 13). All of the species' final regression models contain at least one pair of (linearly) strongly correlated site conditions (e.g. magnesium and calcium saturation), indicating how important that particular complex is for species' distribution. The correlation of individual explanatory variables with the combination of all other model terms, expressed by the VIF, points at possible collinearity problems, mainly with IR and pH. However, a likelihood ratio test shows that each model term within each final model is significant at $\alpha = 0.01$. Furthermore, no strange behaviour

Table 4. Final regression models and model evaluation parameters.

	Model	Terms[a]	Thr.	CCR	$\hat{\kappa}$	% *D*
A. reptans	GLM	Soil + treat + p(pH) + Cl + p(Mg) + p(NO_3) + p(SO_4) + **IR** + mean:soil	0.154	0.70	0.25	15
A. sylvestris	GAM	Soil + treat + s(pH) + s(Fe) + K + Mg + NO_3 + NH_4 + s (PO_4) + s(SO_4)	0.544	0.72	0.44	18
A. elatior	GAM	Treat + **pH** + Cl + s(Ca) + Fe + Mg + s(PO_4) + s(SO_4) + ***IR*** + mean:soil	0.125	0.80	0.37	33
C. palustris	GAM	Treat + **pH** + s(Cl) + s(Ca) + s(Fe) + K + s(Mg) + NO_3 + PO_4 +s(SO_4) + ***IR*** + mean:soil	0.263	0.74	0.41	25
C. acuta	GAM	Treat + pH + s(Cl) + s(Ca) + PO_4 + s(SO_4) + ***IR*** + mean:soil	0.154	0.73	0.27	18
C. palustre	GLM	Treat + p(Ca) + K + p(Mg) + p(NO_3) + NH_4 + p(PO_4) + p(SO_4) + p(IR) + mean:soil	0.420	0.77	0.53	29
E. palustre	GAM_{log}	Treat + **pH** + s(Cl) + s(Fe) + s(K) + s(Mg) + s(NO_3) + s(NH_4) + PO_4 + s(SO_4) + **IR** + mean:soil	0.232	0.79	0.51	31
E. canabinum	GLM	Soil + treat + **pH** + p(Ca) + p(Mg) + NH_4 + p(SO_4) + p(IR) + mean:soil	0.138	0.80	0.45	37
F. ulmaria	GAM	Soil + **pH** + **Cl** + s(Ca) + Fe + K + s(Mg) + NH_4 + s(PO_4) + s(SO_4) + ***IR***	0.588	0.69	0.38	18
G. tetrahit	GAM	Treat + pH + s(Ca) + Fe + NO_3 + s(NH_4) + SO_4 + mean:soil	0.400	0.70	0.39	20
J. acutiflorus	GAM	Treat + s(pH) + NO_3 + s(NH_4) + s(PO_4) + s(SO_4) + s(IR)	0.247	0.87	0.67	51
L. uliginosus	GLM_{log}	Treat + p(pH) + **Ca** + p(Fe) + p(K) + p(Mg) + p(NO_3) + NH_4 + p(PO_4) + mean:soil	0.242	0.83	0.61	41
L. flos-cuculi	GAM	Soil + treat + **Ca** + s(Fe) + s(Mg) + NO_3 + s(NH_4) + s(SO_4) + mean:soil	0.233	0.72	0.36	19
L. europaeus	GLM	Cl + p(Ca) + Fe + K + p(Mg) + p(NO_3) + p(PO_4) + SO_4 + p(IR) + mean:soil	0.473	0.76	0.52	27
L. salicaria	GAM	Treat + **pH** + Cl + s(Fe) + K + s(NO_3) + NH_4 + s(SO_4) + **IR** + mean:soil	0.395	0.65	0.29	12
M. aquatica	GLM	Treat + p(pH) + p(Cl) + p(Ca) + Fe + p(Mg) + NO_3 + NH_4 + p(IR) + mean:soil	0.137	0.74	0.30	24
P. arundinacea	GLM	Soil + **pH** + p(Cl) + p(Ca) + Fe + K + p(Mg) + p(NH_4) + ***IR***	0.182	0.74	0.35	18
S. sylvaticus	GAM	Treat + s(pH) + s(Ca) + s(Fe) + K + s(Mg) + s(NH_4) + s(PO_4) + s(SO_4) + mean:soil	0.257	0.76	0.44	22

Selected predictor variables (Terms) in stepwise selection sequence, threshold (Thr.), CCR, Cohen's $\hat{\kappa}$ and percentage explained deviance (% *D*) are given; explanatory variables with a VIF $>$ 10 are printed in bold, those with VIF $>$ 20 in bold italics.

[a] p(•) = polynomial using 2 df, i.e. a quadratic response shape; s(•) = cubic smoothing spline using 2 df.

of coefficients is observed (during modelling and cross-validation), suggesting that coefficients were successfully estimated.

Most models discriminate well, with CCR ranging between about 65 and 87% (Table 4). Cohen's $\hat{\kappa}$ points at poor discrimination for *A. reptans*, *C. acuta*, *L. salicaria*, *M. aquatica* and *P. arundinacea*. *J. acutiflorus* and *L. uliginosus* are the best modelled species, in terms of CCR, $\hat{\kappa}$ and percentages of explained deviance. The model for *L. salicaria* performs worst, with the lowest CCR and little explained deviance.

As an example of model performance, Figures 6 and 7 show mapped predictions for *L. flos-cuculi* and *M. aquatica* in comparison to the presence–absence field data used for modelling. Agreements between observations and predictions, as well as differences in agreement between sub-samples, are visible. Predictions for *L. flos-cuculi*, for instance, appear to be poor for the Zwarte Beek, with many high prediction values in grid cells which the species did not occur in and numerous low prediction values in cells which the species did occur in. This is confirmed by the discrimination measures for that particular sub-sample and species (Table 5). *Lychnis flos-cuculi* is best predicted for the Snoekengracht, *Mentha aquatica* for the Doode Bemde.

Regional differences

For 13 of the 18 plant species the factor distinguishing between sub-samples is significant ($\alpha = 0.01$) when added to the final regression model; for *L. flos-cuculi* the regional factor is only significant at $\alpha = 0.05$; and for *A. reptans*, *A. elatior*, *L. uliginosus* and *P. arundinacea* it is not significant at $\alpha = 0.05$. When significant, adding this factor decreases the cross-validation deviances for all species; CCR and the proportions of explained deviance change little (0–3%).

Spatial autocorrelation in data and model residuals

Empirical semivariograms for plant species' field data and for the residuals of the final regression models are presented in Figure 8. All species show spatial autocorrelation in their presence and absence records. Most model-residual plots are flatter than the field data plots, indicating that some of the spatial autocorrelation is explained by model predictors. Some species show similar semivariograms for species data and model residuals. Notice, that all plots are based on Pearson residuals (for the plant data using a logistic regression model without explanatory variables) and should therefore display an approximate unit sill. Sills bigger and smaller than one point at, respectively, over and under-dispersion in the data. For some species (e.g. *C. palustre*) model residuals display more variance than the field data, for others (e.g. *L. flos-cuculi*) less.

Figure 6. GAM predictions for *Lychnis flos-cuculi*. Grid cells are presented full size for species presence and smaller for species absence in the field data; grey shade codes predicted probabilities. (a) Doode Bemde, (b) Snoekengracht, (c) Vorsdonkbos and (d) Zwarte Beek; grid cell sizes are 20×20 m for (a), (c), (d) and 10×10 m for (b).

Figure 7. GLM predictions for *Mentha aquatica*. Grid cells are presented full size for species presence and smaller for species absence in the field data; grey shade codes predicted probabilities. (a) Doode Bemde, (b) Snoekengracht, (c) Vorsdonkbos and (d) Zwarte Beek; grid cell sizes are 20 × 20 m for (a), (c), (d) and 10 × 10 m for (b).

Table 5. CCR and Cohen's $\hat{\kappa}$ for *Lychnis flos-cuculi* and for *Mentha aquatica* in each of the sub-samples.

Species	Sub-sample	CCR	$\hat{\kappa}$
L. flos-cuculi	Doode Bemde	0.75	0.35
	Snoekengracht	0.74	0.45
	Vorsdonkbos	0.77	0.39
	Zwarte Beek	0.65	0.24
M. aquatica	Doode Bemde	0.68	0.33
	Snoekengracht	0.83	0.41
	Vorsdonkbos	0.72	0.22
	Zwarte Beek	0.70	0.20

The thresholds considered are 0.233 and 0.137, respectively.

For most species, empirical semivariograms computed for the four sub-samples separately show different shapes (see example in Figure 9) – evidence for geographic differences in spatial dependency structures.

Notice that we did not fit lines (models) to the empirical semivariograms plotted in Figures 8 and 9. Semivariogram models would facilitate quantification of the spatial variance component (using the difference between the sill, at maximum range and the nugget variance). Yet, some caution with the interpretation of these plots is in place. Because standardized residuals are used, we do not know how much of the species' nugget is explained by the models; we just know the proportion of nugget and structural variance. It is therefore difficult to quantify accurately the amount of autocorrelation accounted for by the model and these semivariograms constitute more of a general indication for it.

Discussion

In general, the here developed regression models seem ecologically sound and predict species' distribution in Flemish river valleys well, despite of discrepancies between data quality and model assumptions. This study demonstrates that predictive modelling using standard statistical regression procedures can be reasonably successful with GLM or GAM in the presence of: non-homogeneous aggregated data; data that are spatially autocorrelated; partly interpolated and partly measured explanatory variables; explanatory variables and response variables collected at different scales; and correlated explanatory variables. However, model application and inference should be handled with care, as assumptions of independent, error-free explanatory variables and independent errors are clearly not met.

More than half of the species are better modelled by GAM than by GLM, with data-driven smooth response shapes instead of second-order polynomials. These results are in agreement with others found in literature, indicating that species' responses are often complex and difficult to fit using simple symmetric response shapes (e.g. Austin

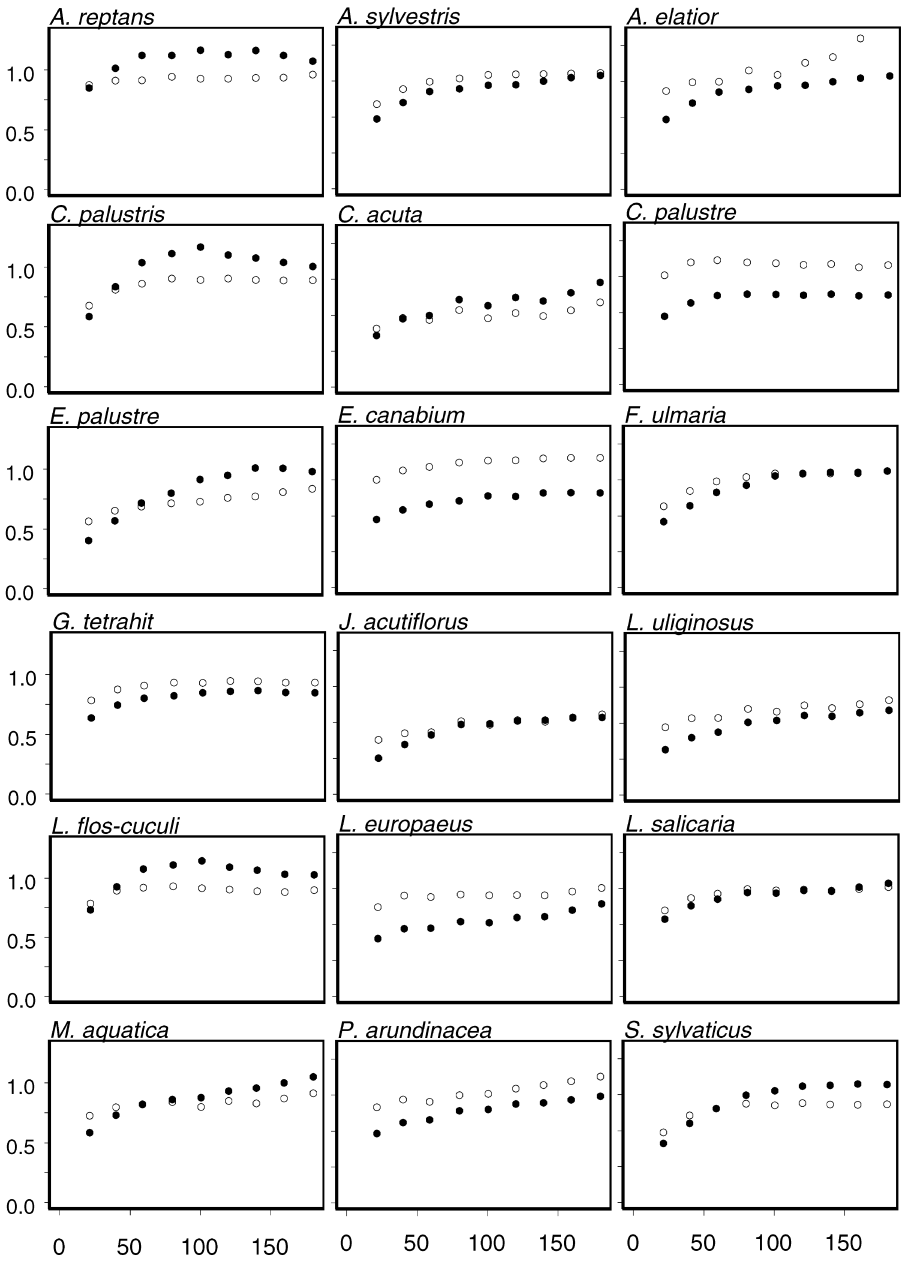

Figure 8. Empirical semivariograms for the plant species data (full circles) and model residuals (empty circles), calculated from Pearson residuals of models containing the intercept only and Pearson residuals of the final regression models, respectively; semivariances are presented on the *y*-axes and distances in metres on the *x*-axes.

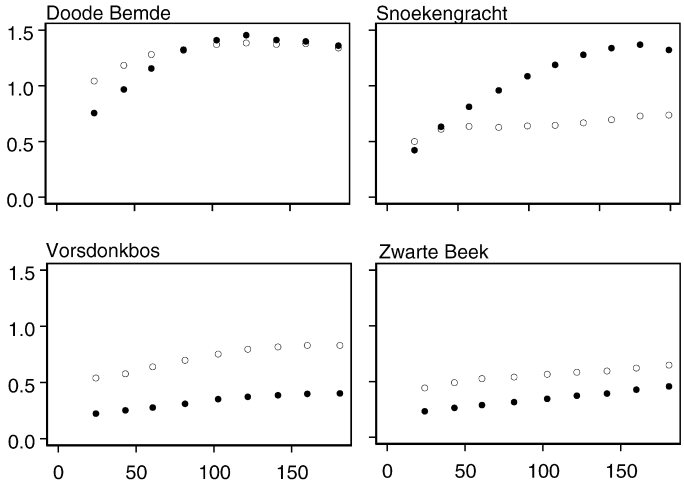

Figure 9. Empirical semivariograms for *Lychnis flos-cuculi* computed separately for the four sampled valleys. Species field data (full circles) and model residuals (empty circles) were calculated from Pearson residuals of models containing the intercept only and Pearson residuals of the final regression models, respectively; semivariances are presented on the *y*-axes and distances in metres on the *x*-axes.

and Meyers 1996; Bio et al. 1998; Leathwick 1998). Unlike the responses considered in GLM, smoothed responses are generally not parameterised, but recalculated for the data on which the model was developed (the training data) whenever necessary. Consequently, the use of GAM in the modelling environment as management tool will require more computer space than the use of GLM, due to the attachment of the original dataset. One could argue that complex response shapes could also be fitted using GLM with higher-order polynomials. But these tend to reveal spurious and unrealistic response shapes (Austin et al. 1990; Huisman et al. 1993; Bio et al. 1998).

Ecologically sensible regression models are obtained containing sets of site conditions relevant to the studied species (Van der Molen 1999). Nearly all regression models contain soil, management and one or more variables indicating acidity/calcium saturation and nutrient richness of the water as predictors. These variables are regarded as key-factors for vegetation development in river valleys (Grootjans 1985; Wassen 1990; De Mars 1996; Bootsma 2000). In this article a limited number of so-called ‘generalist’ phreatophytic species are studied, because these are present in all of the four study sites, despite the differences in site conditions. Cartographic representation of model results (e.g. Figures 6 and 7) shows spatially consistent (not random or scattered) species distribution patterns. Evaluation of these results based on expert knowledge of the site conditions and the ecological behaviour of the modelled plant species reveals no aberrant or unexpected prediction results. Similar results were obtained in an earlier study, for less ‘generalist’ species (De Becker et al. 2001), suggesting that this type of regression models is suitable for species’ distribution prediction.

Model selection leads to multiple regression models containing most of the available predictor variables for each species. In a previous study with GLM applied to the same data, but using a forward model selection procedure (De Becker et al. 2001), fewer variables were selected. Given the strict selection criterion (BIC) used in the present study and the significance of the model terms in our final models, we believe that this discrepancy is not due to lack of parsimony in the present selection procedure, but that forward stepwise selection does overlook some relevant explanatory variables.

Final models are chosen from the four stepwise selected models, based on their performance assessed through cross-validation. Cross-validation applied to model selection itself and testing of all possible explanatory variable combinations would have given better insight into the predictive performance of a wider range of possible models, but turned out to be too laborious for this study.

Strong linear correlations between pairs of site conditions did not necessarily cause collinearity problems in the multiple regression models, where most explanatory variables are fitted by smoothed responses or second-order polynomials. Critical variance inflation values are mainly observed for IR and pH, and only for explanatory variables modelled as linear terms. However, we found no strange behaviour in model coefficients, suggesting these were acceptably fit. Anyway, collinearity does not affect the ability of a regression equation to predict a response. It would pose a problem if the purpose of the study were estimation of the contribution of individual predictors. Our model selection goal is to determine whether a site-condition variable has predictive capability in the presence of others. Hence, possible correlation can be disregarded if that predictor turns out to be significant despite being correlated with other predictors (De Veaux and Ungar 1994; Dallal 2001).

Plant species data were mapped in grids of adjacent regular square cells in selected study areas spread over different phytogeographical regions of Flanders. Sampling the vegetation on a grid covering the whole area of interest provides the most complete sample possible and reduces the chance of sampling a limited part of plant species' response ranges (De Becker et al. 1999), which could hamper the general applicability of the models. However, field data exhibit spatial autocorrelation between records of each grid and spatial variation between grids.

Most of the abiotic data (groundwater level and chemistry) were collected at a limited number of point locations within each grid; hence, at a much smaller sampling scale (or support) and with extensive un-sampled surface in between. They had to be spatially interpolated and up-scaled (to grid-cell size) to match the vegetation, management and soil data. Using estimated site conditions as explanatory variables violates the assumption of GLM and GAM that explanatory variables are measured without error. Estimation errors in the interpolated explanatory variables are present and apparently high for some variables. Their effect could be assessed using Monte Carlo methods, sampling from the distribution of explanatory variable estimates at each grid cell; an exercise too laborious and time consuming for the present study.

Kriging estimates of the site conditions are obviously autocorrelated, according to the spatial model used for interpolation. As explanatory variables, they explain part of the spatial autocorrelation observed in the vegetation field data. Residual autocorrelation can be attributed to spatial variables relevant to species' distribution that are not included in our models; to differences in (scales of) spatial dependence in biotic and abiotic data; or to poor spatial interpolation of the point data. Collection of water quality data in each grid cell would reduce this input error. But, even then, sampling at the grid-cell-covering vegetation scale would not be financially or even technically feasible.

A factor coding for the four sampled valleys is, most of the times, significant when added to the final regression model. This points at regional differences in species' distribution that are not explained by our models. There may be differences in species' response to the explanatory variables due to valley-specific pseudo-correlations with non-modelled variables. Notice that the effect of the valley-coding factor within a model is the attribution of a different intercept for each valley; response shapes and predictor terms are not changed by it. Valley-specific model selection would probably lead to greater improvements in model discrimination and prediction. In a previous study (De Becker et al. 2001) we found that regional models, selected individually for each sub-sample, included different relevant site conditions and different species-condition response shapes. The difference in grid cell size for Snoekengracht is also a possible source of error in the aggregated sample.

This study presents an application of GLM and GAM in the development of predictive models of the type requested as nature-management tools. We observe that, in practice, models have to suit model purpose as well as possible even if data do not fully support model assumptions. Shortcomings, if not removable, should be assessed and, at least, communicated to the final user, just as model applicability and credibility. The models presented are, for instance, valid for nutrient-poor river valleys only, as model input data do not include nutrient rich situations. For model inference, data characteristics (autocorrelation, input errors, possible collinearity) have to be considered. So far, the predictive power of these models could not be examined on other regions. Validation against data collected elsewhere – i.e. an extrapolation in space – is a next step to be taken to see how far the applicability of these empirical models reaches.

Acknowledgements

The authors would like to thank the nature conservation organisation Natuurpunt for the use of the four nature reserves. Particular thanks go to the site managers Georges Buelens (Snoekengracht), Luk Vervoort (Vorsdonkbos) and Willy Van Look (Zwarte Beek) for their generous cooperation. We are grateful to Vincent Rodenburg for reading through the manuscript.

References

Akaike H. 1977. On entropy maximization principle. In: Krishnaiah P.R. (ed) Applications of Statistics. North-Holland Publishing Company, Amsterdam, pp. 27–41.

Akaike H. 1978. A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30A: 9–14.

Albert P.S. and McShane L.M. 1995. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. Biometrics 51: 627–638.

Augustin N.H., Mugglestone M.A. and Buckland S.T. 1996. Anautologistic model for the spatial distribution of wildlife. Journal of Applied Ecology 33: 339–347.

Austin M.P. 1999. The potential contribution of vegetation ecology to biodiversity research. Ecography 22: 465–484.

Austin M.P. and Meyers J.A. 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecology and Management 85: 95–106.

Austin M.P., Cunningham R.B. and Fleming P.M. 1984. New approachesto direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio 55: 11–27.

Austin M.P., Nicholls A.O., Doherty M.D. and Meyers J.A. 1994. Determining species response functions to an environmental gradient by means of a beta-function. Journal of Vegetation Science 5: 215–228.

Austin M.P., Nicholls A.O. and Margules C.R. 1990. Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species. Ecological Monographs 60: 161–177.

Barendregt A. and Nieuwenhuis J.W. 1993. ICHORS, hydro-ecological relations by multidimensional modelling of observations. In: Hooghart J.C. and Posthumus C.W.S. (eds) The Use of Hydro-ecological Models in The Netherlands. CHO-TNO, Delft, The Netherlands.

Batelaan O. and De Smedt F. 1994. Regionale grondwater stroming rond een aantal kwelafhankelijke natuurgebieden. Intituut voor Natuurbehoud, Brussels, Belgium.

Batelaan O., De Smedt F., De Becker P. and Huybrechts W. 1995. Characterisation of regional groundwater discharge area by combined analysis of hydrochemistry, remote sensing and groundwater modelling. In: Dillon P. and Simmers I. (eds) Shallow Groundwater Systems. International Contributions to Hydrogeology, Vol 18. IAH, Balkema, Rotterdam, The Netherlands, pp. 75–87.

Begg G.S. and Reid J.B. 1997. Spatial variation in seabird density at a shallow sea tidal mixing front in the Irish Sea. ICES Journal of Marine Science 54: 552–565.

Bio A.M.F. 2000. Does vegetation suit our models? Data and model assumptions and the assessment of species distribution in space. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands.

Bio A.M.F., Alkemade R. and Barendregt A. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. Journal of Vegetation Science 9: 5–16.

Bootsma M. 2000. Stress and recovery in wetland ecosystems. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands.

Buckland S.T. and Elston D.A. 1993. Empirical models for the spatial distribution of wildlife. Journal of Applied Ecology 30: 478–495.

Buckland S.T., Burnham K.P. and Augustin N.H. 1997. Model selection: an integral part of inference. Biometrics 53: 603–618.

Burrough P.A. and McDonnell R.A. 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford, UK.

Cohen J. 1960. A. coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37–46.

Cressie N. 1985. Fitting variogram models by weighted leastsquares. Mathematical Geology 17: 563–586.

Cressie N.A.C. 1993. Statistics for Spatial Data. Revised edition. John Wiley and Sons, New York.

Dallal G.E. 2001. Collinearity. http://www.tufts.edu/∼gdallal/collin.htm.

De Becker P. and Huybrechts W. 2000a. Vallei van de Zwarte Beek – Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium.

De Becker P. and Huybrechts W. 2000b. De Doode Bemde – Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium.

De Becker P., De Bie E., Huybrechts W., Bio A.M.F. and Wassen M. 2001. Ontwikkeling van een hydro-ecologisch model voor vallei-ecosystemen in Vlaanderen, VLITORS. Intituut voor Natuurbehoud, Brussels, Belgium.

De Becker P., Hermy M. and Butaye J. 1999. Ecohydrological characterization of a groundwater-fed alluvial floodplain mire. Applied Vegetation Science 2: 215–228.

Décamps H., Fortuné M., Gazelle F. and Pautou G. 1988. Historical influence of man on the riparian dynamics of a fluvial landscape. Landscape Ecology 1: 163–173.

De Gruijter J.J. and Ter Braak C.J.J. 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Mathematical Geology 22: 407–415.

De La Ville N., Cousins S. and Bird C. 1997. Habitat Suitability Analysis Using Logistic Regression and GIS to Outline Potential Areas for Conservation of the Grey Wolf (*Canis lupus*). Presented at the Conference: 'GIS Research UK', University of Leeds, Leeds, 9–11 April.

De Mars H. 1996. Chemical and physical dynamics of fen hydro-ecology. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands.

De Swart E.O.A.M., Van der Valk A.G., Koehler K.J. and Barendregt A. 1994. Experimental evaluation of realized niche models for predicting responses of plant species to a change in environmental conditions. Journal of Vegetation Science 5: 541–552.

De Veaux R.D. and Ungar L.H. 1994. Multicollinearity: a tale of two nonparametric regressions. In: Cheesman P. and Oldford R.W. (eds) Selecting Models from Data: AI and Statistics IV. Lecture Notes in Statistics 89. Springer, New York, pp. 393–402.

Erisman J.W. and Draaijers G.P.J. 1995. Atmospheric Deposition in Relation to Acidification and Eutrophication. Elsevier, Amsterdam.

Ertsen A.C.D., Bio A.M.F., Bleuten W. and Wassen M.J. 1998. Comparison of the performance of species response models in several landscape units in the province of Noord-Holland, The Netherlands. Ecological Modelling 109: 213–223.

Fielding A.H. and Bell J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38–49.

Franklin J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science 9: 733–748.

Gilbert O.L. and Anderson P. 1998. Habitat Creation and Repair. Oxford University Press, Oxford, UK.

Gotway C.A. and Stroup W.W. 1997. A generalized linear model approach to spatial data analysis and prediction. Journal of Agricultural, Biological, and Environmental Statistics 2: 157–178.

Grootjans A.P. 1985. Changes of groundwater regime in wet meadows. Ph.D. Thesis, University of Groningen, Groningen, The Netherlands.

Guisan A. and Harrell F.E. 2000. Ordinal response regression. Journal of Vegetation Science 11: 617–626.

Hastie T.J. and Tibshirani R.J. 1990. Generalized Additive Models. Chapman & Hall, London.

Hellberg F. 1995. Entwicklung der Grünland vegetation bei Wiedervernässung und periodischer Überflutung. Vegetationsökologische Untersuchungen in nordwestdeutschen Überflutungspoldern. Dissertationes Botanicae, Band 243. Cramer Verlag, Stuttgart, Germany.

Hill M.O. 1991. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. Journal of Biogeography 18: 247–255.

Hosmer D.W. Jr and Lemeshow S. 1989. Applied Logistic Regression. John Wiley and Sons, New York.

Hudson W.D. and Ramm C.W. 1987. Correct formulation of the Kappa coefficient of agreement. Photogrammetric Engineering and Remote Sensing 53: 421–422.

Huisman J., Olff H. and Fresco L.F.M. 1993. A hierarchical set of models for species response analysis. Journal of Vegetation Science 4: 37–46.

Huntley B., Berry P.M., Cramer W. and McDonald A.P. 1995. Modelling present and future ranges of some European higher plants using climate response surfaces. Journal of Biogeography 22: 967–1001.

Huybrechts W. and De Becker P. 1999. De Snoekengracht – Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium.

Huybrechts W. and De Becker P. 2000. Vorsdonkbos-Turfputten – Ecohydrologische Atlas. Institute of Nature Conservation, Brussels, Belgium.

Janse J.H., Aldenberg T. and Kramer P.R.G. 1992. A mathematical model of the phosphorus cycle in Lake Loosdrecht and simulation of additional measures. Hydrobiologia 233: 119–136.

Köhl M. and Gertner G. 1997. Geostatistics in evaluating forest damage surveys: considerations on methods for describing spatial distributions. Forest Ecology and Management 95: 131–140.
Leathwick J.R. 1998. Are New Zealand's Nothofagus species in equilibrium with their environment? Journal of Vegetation Science 9: 719–732.
Londo G. 1988. Nederlandse freatofyten. Pudoc, Wageningen, The Netherlands (incl. English summary).
Manel S., Williams H.C. and Ormerod S.J. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38: 921–931.
Margules C.R., Nicholls A.O. and Austin M.P. 1987. Diversity of Eucalyptus species predicted by a multivariable environmental gradient. Oecologia 71: 229–232.
Mathsoft 1996. S+SpatialStats User's Manual. Version 1.0. Mathsoft, Seattle, Washington.
McCullagh P. and Nelder J.A. 1989. Generalized Linear Models. 2nd Edition. Chapman & Hall, London.
Naiman R., Decamps H. and Fournier F. 1989. Role of Land/InlandWater Ecotones, Landscape Management and Restoration. UNESCO, Paris.
Nelder J.A. and Wedderburn R.W.N. 1972. Generalized linear models. Journal of the Royal Statistical Society A 135: 370–384.
Nicholls A.O. 1989. How to make biological surveys go further with generalised linear models. Biological Conservation 50: 51–75.
Nienhuis P.H., Leuven R.S.E.W. and Ragas A.M.J. (eds) 1998. New Concepts for Sustainable Management of River Basins. Backhuys Publishing, Leiden, The Netherlands.
Olde Venterink H. and Wassen M.J. 1997. A comparison of six models predicting vegetation response to hydrological habitat change. Ecological Modelling 101: 347–361.
Osborne P.E., Alonso J.C. and Bryant R.G. 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. Journal of Applied Ecology 38: 458–471.
Pearce J. and Ferrier S. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modelling 128: 127–147.
Pebesma E.J. 1997. Gstat User's Manual. Available at:http://www.geog.uu.nl/gstat/.
Pebesma E.J., Duin R.N.M. and Bio A.M.F. 2000. Spatial interpolation of sea bird densities on the Dutch part of the North Sea. Department of Physical Geography, Faculty of Geographical Sciences, Utrecht University, Utrecht, The Netherlands.
Petts G.E. and Amoros C. (eds) 1996. Fluvial Hydrosystems. Chapman & Hall, London.
Raftery A.E. 1986. Choosing models for cross-classifications. American Sociological Review 51: 145–146.
Rich T.C.G. and Woodruff E.R. 1996. Changes in vascular plant floras of England and Scotland between 1930–1960 and 1987–1988: the BSBI Monitoring Scheme. Biological Conservation 75: 217–229.
Robertson G.P. and Freckman D.W. 1995. The spatial distribution of nematode trophic groups across a cultivated ecosystem. Ecology 76: 1425–1432.
Rossi R.E., Mulla D.J., Journel A.G. and Franz E.H. 1992. Geostatistical tools for modelling and interpreting ecological spatial dependence. Ecological Monographs 62: 277–314.
Runhaar J., Van Gool C.R. and Groen C.L.G. 1996. Impact of hydrological changes on nature conservation areas in the Netherlands. Biological Conservation 76: 269–276.
Schot P.P. and Molenaar A. 1992. Regional changes in groundwater flow patterns and effects on groundwater composition. Journal of Hydrology 130: 151–170.
Schwarz G. 1978. Estimating the dimension of a model. The Annals of Statistics 6: 461–464.
Smith P.A. 1994. Autocorrelation in logistic regression modelling of species distributions. Global Ecology and Biogeography Letters 4: 47–61.
Sokal R.R. and Oden N.L. 1978a. Spatial autocorrelation in biology. 1. Methodology. Biological Journal of the Linnean Society 10: 199–228.
Sokal R.R. and Oden N.L. 1978b. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10: 229–249.
Tilman D. 1994. Competition and biodiversity in spatially structured habitats. Ecology 75: 2–16.
Van der Aa N.G.F.M., Trepel M., Van Gaans P.F.M., Bleuten W. and Kluge W. 2001. Modelling water flow and fluxes of a valley mire for use in restoration. Landnutzung und Landentwicklung 42: 72–78.
Van der Molen D. 1999. The role of eutrophication models in water management. Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands.

Van Liere L. and Gulati R.D. (eds) 1992. Restoration and recovery of eutrophic lake ecosystems in the Netherlands. Hydrobiologia 233: 283–287.

Van Wirdum G. 1990. Vegetation and Hydrology of a Floating Rich-Fen. Datawyse, Maastricht, The Netherlands.

Wassen M.J. 1990. Water flow as a major landscape ecological factor in fen development. Ph.D. Thesis, Utrecht University, Utrecht, The Netherlands.

Wassen M.J. and Barendegt A. 1992. Topographic position and water chemistry of fens in a Dutch river plain. Journal of Vegetation Science 3: 447–456.

Wheeler B.D., Shaw S.C., Fojt W.J. and Robertson R.A. 1995. Restoration of Temperate Wetlands. John Wiley, Chichester, UK.

Witte J.P.M. and Van der Meijden R. 1992. Verspreiding ennatuurwaarden van ecotoopgroepen in Nederland. Landbouwuniversiteit Wageningen and Rijksherbarium/Hortus Botanicus Leiden, Wageningen/Leiden, The Netherlands.

Yee T.W. and Mitchell N.D. 1991. Generalized additive models inplant ecology. Journal of Vegetation Science 2: 587–602.

Zimmermann N.E. and Kienast F. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. Journal of Vegetation Science 10: 469–482.