Chin. Geogra. Sci. 2013 Vol. 23 No. 3 pp. 331–343
doi: 10.1007/s11769-013-0590-0

Springer Science Press
www.springerlink.com/content/1002-0063

Predictive Vegetation Mapping Approach Based on Spectral Data, DEM and Generalized Additive Models

SONG Chuangye[1], HUANG Chong[2], LIU Huiming[3]

(1. *State Key Laboratory of Vegetation and Environmental Change*, *Institute of Botany*, *Chinese Academy of Sciences*, *Beijing* 100093, *China*; 2. *State Key Laboratory of Resources and Environmental Information System*, *Institute of Geographic Sciences and Natural Resources Research*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*; 3. *Satellite Environment Centre*, *Ministry of Environmental Protection*, *Beijing* 100094, *China*)

Abstract: This study aims to provide a predictive vegetation mapping approach based on the spectral data, DEM and Generalized Additive Models (GAMs). GAMs were used as a prediction tool to describe the relationship between vegetation and environmental variables, as well as spectral variables. Based on the fitted GAMs model, probability map of species occurrence was generated and then vegetation type of each grid was defined according to the probability of species occurrence. Deviance analysis was employed to test the goodness of curve fitting and drop contribution calculation was used to evaluate the contribution of each predictor in the fitted GAMs models. Area under curve (AUC) of Receiver Operating Characteristic (ROC) curve was employed to assess the results maps of probability. The results showed that: 1) AUC values of the fitted GAMs models are very high which proves that integrating spectral data and environmental variables based on the GAMs is a feasible way to map the vegetation. 2) Prediction accuracy varies with plant community, and community with dense cover is better predicted than sparse plant community. 3) Both spectral variables and environmental variables play an important role in mapping the vegetation. However, the contribution of the same predictor in the GAMs models for different plant communities is different. 4) Insufficient resolution of spectral data, environmental data and confounding effects of land use and other variables which are not closely related to the environmental conditions are the major causes of imprecision.

Keywords: vegetation mapping; Generalized Additive Models (GAMs); SPOT; Receiver Operating Characteristic (ROC); Generalized Regression Analysis and Spatial Predictions (GRASP); Huanghe River Delta

Citation: Song Chuangye, Huang Chong, Liu Huiming, 2013. Predictive vegetation mapping approach based on spectral data, DEM and Generalized Additive Models. *Chinese Geographical Science*, 23(3): 331–343. doi: 10.1007/s11769-013-0590-0

1 Introduction

Remote sensing image classification is the main approach to achieve vegetation map at large scale (Sanders *et al*., 2004). Traditional classification methods based on the spectral features, such as supervised classification and unsupervised classification, were widely used to classify the remote sensing image (Yang and Zhou, 2001). However, the similarities of spectral signature among plant communities make the delineation of distinct vegetation patches based on remote sensing image a difficult task (Treitz *et al*., 1992). Expert system was developed to surmount the limitation of using spectral features solely to classify the remote sensing image (Joseph and Gary, 2004). Expert system incorporates the spectral features, environmental variables and expert knowledge in remote sensing image classification (Joseph and Gary, 2004; Zhang and Zhu, 2011). Therefore, theoretically, expert system is a perfect classification method, however, uncertainties of knowledge acquisition and quantization limit the application of expert system (Joseph and Gary, 2004).

Received date: 2012-04-18; accepted date: 2012-09-21
Foundation item: Under the auspices of National Natural Science Foundation of China (No. 41001363)
Corresponding author: SONG Chuangye. E-mail: songcy@ibcas.ac.cn

Environmental response models are another way to map vegetation (Guisan and Zimmermann, 2000). According to the ecological niche theory, each plant community occupies a geographical space, which is determined by terrain, soil, temperature, precipitation, radiation and so on (Song, 2001). Based on the ecological niche theory, a ′balanced′ or ′quasi-balanced′ relationship was assumed to exist between vegetation and environment. Based on this theory, statistical models were employed to define environmental response models and then predict the distribution of vegetation (Guisan and Zimmermann, 2000; Ferrier and Guisan, 2006). However, it is insufficient for environmental response models to map actual vegetation, because the actual vegetation is greatly affected by disturbances such as land use change (Guisan and Zimmermann, 2000; White *et al*., 2001; Dirnbök *et al*., 2002; Dullinger *et al*., 2003), or historically determined distribution patterns (Dirnbök *et al*., 2001). Consequently, environmental response models are usually used to predict potential rather than actual vegetation distribution (Guisan and Zimmermann, 2000).

In conclusion, the similarities of spectral features among different plant communities limit the capacity of remote sensing image classification, and also it is insufficient for environmental response models to map actual vegetation. On the other hand, plant communities which can not be reliably delineated by spectral attributes may be separated with the help of abiotic environmental variables, and vice versa (Ferrier *et al*., 2002; Ferrier and Guisan, 2006). In other words, both abiotic environmental variables and spectral attributes have limitations for mapping vegetation, they should be viewed as mutual complementation rather than competitive sources of information for vegetation mapping (Ferrier *et al*., 2002; Miller *et al*., 2007). Therefore, integrating environmental variables and spectral variables based on statistical models might be a feasible way to map vegetation.

Nonlinear models based on Gaussian curve, such as Detrended Canonical Analysis (DCA) and Canonical Correspondence Analysis (CCA), are very popular in the study of the relationship between vegetation and environment (Ohmann and Gregory, 2002; Dirnböck *et al*., 2003). However, the relation between vegetation and environment is so complex that it can not be completely fitted for unimodel, moreover, dimensionality reduction of DCA and CCA will cause information loss (Zhang, 1995). Generalized Additive Models (GAMs) are nonparametric extension of Generalized Linear Models (GLMs) and allow for both continuous and factor variables (Hastie and Tibshirani, 1990). Smoothed function derived from the explanatory variables was employed to build a model in GAMs, instead of pre-establishing a parametric model, which overcomes the limitation of pre-establishing parametric models. Other non-parametric models, such as Classification and Regression Tree (CART) and Artificial Neural Network (ANN), also have the ability to fit robust models to ecological data and may be better suited for modeling interactions. However, the methods of selecting significant variables in CART and ANN are not based on appropriate statistical distribution and the response curve shapes can not be observed, resulting in their failure in the ecological interpretability (Lehmann *et al*., 2002). Therefore, GAMs were widely used to model the responses of vegetation to environmental variables in ecological studies (Austin, 2002; Lehmann *et al*., 2002; Shen and Zhao, 2007; He *et al*., 2008; Wen *et al*., 2008).

In this study, our objectives are: 1) to examine the relationship between vegetation and environmental variables, as well as spectral attributes, based on GAMs; 2) to map the distribution of dominant plant communities based on GAMs; and 3) to evaluate the uncertainties of GAMs in vegetation mapping.

2 Data and methods

2.1 Study area

The Huanghe (Yellow) River Delta (36°55′–38°12′N, 118°07′–119°18′E) located in the northern Shandong Province of the eastern China is selected as the study area of this research (Fig. 1). It lies on the south side of the Bohai Sea. The region is characterized by temperate, semi-humid continental monsoon climate. The mean annual temperature ranges from 11.50℃ to 12.48℃, with the warmest monthly temperature of 26.68℃ in July and the coldest of 4.18℃ in January. The mean annual precipitation is 590.9 mm and mean annual evaporation is over 1500 mm. The maximum monthly rainfall is 227.0 mm in July and the minimum is 1.7 mm in January (Zang, 1996). Within the delta, the underground water table is high and the water is saline. The entire area is mainly covered by wet and saline soil. Meadow, especially the halophytic meadow, dominated by Chinese tamarisk (*Tamarix chinensis* Lour.), Seepweed (*Suaeda salsa* (Linn.) Pall.) and reed (*Phragmites australis* (Cavanilles) Trinius ex Steudel), is the typical

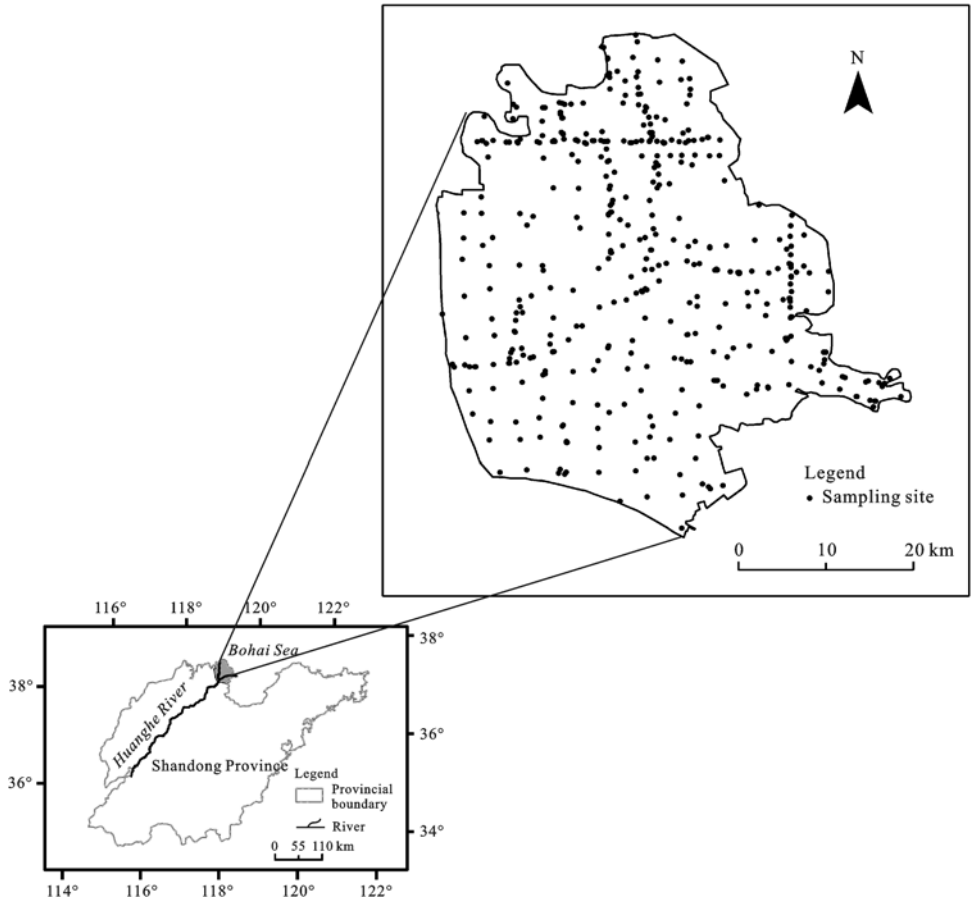

Fig. 1 Location of Huanghe River Delta and distribution of vegetation sampling sites

vegetation in this area (Fig. 2).

2.2 Data and processing

2.2.1 Vegetation survey data

Data were collected during the growing season in 2006, 2007 and 2008. The sampling strategy was random and the size of quadrat varied from 100 m^2 (10 m × 10 m) for shrub species to 1 m^2 (1 m × 1 m) for herbaceous species. Variables recorded include species name, density, height and coverage of all herbal and shrub plants. The cover was estimated according to Braun-Blanquet′s 5-level ordinal scale (Braun-Blanquet, 1933). Geogra-

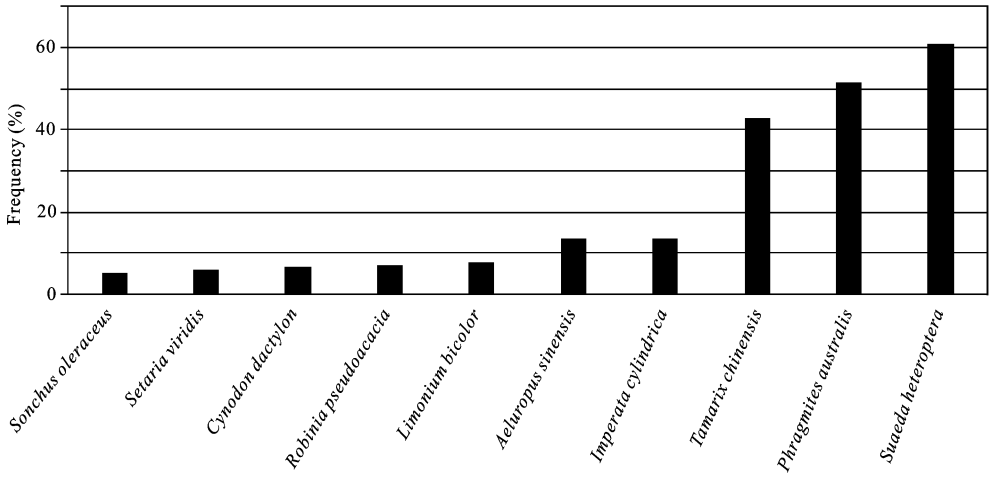

Fig. 2 Main plant species in Huanghe River Delta. Frequency is percent of quadrats in which a plant species occurs

phical coordinates of quadrats were recorded by using a GPS (±10 m). In total, 483 quadrats of vegetation were collected (Fig. 1) and the set of 483 quadrats representing the ultimate modeling input data were marked on the SPOT-5 image and digitized thereafter.

Due to the errors of geometrical correction and imprecision of GPS coordinates of quadrats, we set quadrats near the center of vegetation patch, and kept the quadrats at least 50 m away from the boundary. This will insure that the quadrat-derived variables could represent the spectral and environmental characteristics of this vegetation patch.

As the response variables of GAMs are presence-absence (1/0) data, we supposed that if the importance value (IV) of *T. chinensis*, *S. salsa* or *P. australis* is higher than other species in the same quadrat, then this quadrat was assigned with 1, otherwise, the quadrat was assigned with 0. The importance value (IV) for each species in each quadrat was obtained by the sum of its relative density, coverage and height.

$$IV = C_r + D_r + H_r \quad (1)$$

where C_r is the relative coverage; D_r is the relative density, and H_r is the relative height of each plant species.

2.2.2 Environmental data

Previous studies proved that topography, soil and ground water table have great effects on the vegetation distribution in the Huanghe River Delta (Song *et al*., 2009). However, we did not get the data of ground water table. But, ground water table is significantly related with topography, as well as the nearest distance to the Huanghe River and the nearest distance to the coastline (Song *et al*., 2009). Therefore, altitude (ALT), slope (SP), the nearest distance to the coastline (DC) and the nearest distance to the Huanghe River (DR) were selected as predictors in this research. Altitude and slope were derived from DEM (with 50 m horizontal resolution) which came from the digitization of 1∶10 000 topographical map of the Dongying City in Shandong Province. The nearest distance to the Huanghe River and the nearest distance to the coastline were calculated by the ″Near″ tool in ArcGIS 9.3.

Totally, 585 soil samples (0–30 cm) were collected in the study area. Soil salinity (SALT), soil organic matter (SOM) and soil total nitrogen (TN) and pH were measured according to Liu (1996). At last, the measured soil variables were log-transformed to conform to the normal distribution and interpolated spatially by ordinary Kriging method in ArcGIS 9.0. Table 1 is the prediction errors which are used to evaluate the effects of the interpolation by ordinary Kriging.

2.2.3 Spectral data

SPOT-5, a multi-spectral image with high-resolution of 10 m (acquired on September 7, 2005) was used as the basic spectral data in this study. The image contains four bands of information: visible green (VG) (0.50–0.59 μm), visible red (VR) (0.61–0.68 μm), near infrared (NIR) (0.78–0.89 μm), and short infrared wave band (SIR) (1.58–1.75 μm).

Geometric correction for SPOT-5 image was done by SPOT5-Orbit Pushbroom model which was included in Erdas image 9.3. Twenty-seven ground control points were recorded by using differential GPS (TrimbleGeoXT) and the DEM with 50 m resolution were used to diminish the geometric error and the effects of topography on the SPOT-5 image. Bilinear interpolation was employed to resample the image and the root mean squared error of the geometric correction is less than 5 m. The SPOT-5 image was not subjected to atmosphere correction, one reason was that only one scene was used in this research, the other was that there were no atmosphere data at the time of image acquisition (Song *et al*., 2000).

In this research, besides Digital Number (DN) values in all four wavebands (DN_{VG}, DN_{VR}, DN_{NIR} and DN_{SIR}), Normalized Difference Vegetation Index (NDVI), Ratio Vegetation index (RVI), Soil Adjusted Vegetation Index (SAVI) and Difference Vegetation Index (DVI)

Table 1 Prediction errors for measured soil variables

Error	TN	SOM	pH	SALT
Root-mean-square error	0.05705	0.62610	0.40950	0.73910
Average standard error	0.05907	0.65380	0.49840	1.65600
Mean standardized error	–0.02193	–0.01316	0.01005	0.01124
Root-mean-square standardized error	1.01800	1.21300	0.83270	0.60220

Notes: TN is abbreviation of soil total nitrogen; SOM is abbreviation of soil organic matter; SALT is abbreviation of soil salinity

were selected as predictors in this research. NDVI, RVI, SAVI and DVI were calculated according to the following equations (Zhao, 2003):

$$NDVI = (DN_{NIR} - DN_{VR}) / (DN_{NIR} + DN_{VR}) \quad (2)$$

$$RVI = DN_{NIR} / DN_{VR} \quad (3)$$

$$SAVI = (DN_{NIR} - DN_{VR})(1 + L) / (DN_{NIR} + DN_{VR} + L) \quad (4)$$

$$DVI = DN_{NIR} - DN_{VR} \quad (5)$$

where L was soil adjusted coefficient and it was assigned a value of 0.5 in this research (Zhao, 2003).

The pixels containing each of the 483 quadrats and the eight pixels surrounding the quadrat-pixel (the pixel where a quadrat was located) were all identified and the DN_{VG}, DN_{VR}, DN_{NIR}, DN_{SIR}, NDVI, RVI, DVI and SAVI values were extracted. To make sure that the DN_{VG}, DN_{VR}, DN_{NIR}, DN_{SIR}, NDVI, RVI, DVI and SAVI could represent the spectral characteristics of the vegetation, we stipulated that if the deviation between the derived spectral variables and the average spectral variables exceeds two-fold standard deviation, this derived variable was considered as outlier and would be deleted. The mean value of DN_{VG}, DN_{VR}, DN_{NIR}, DN_{SIR}, NDVI, RVI, DVI and SAVI of the quadrat-pixel and the pixels surrounding the quadrat-pixel were used to build the GAMs model.

Since our model was applicable only to vegetation covered areas where quadrat data were available, we masked out non-vegetation areas (e.g., water, urban, agriculture, and beach) from our GAMs predictions.

2.3 Methods

2.3.1 Predictors selection for GAMs fitting

Total 16 predictors were collected in this research, part of the predictors are highly correlated (Table 2). To diminish the impacts of the correlation on the performance of the fitted GAMs model, we set 0.8 as the maximum correlation allowed between predictors, if higher correlation was found, predictors were withdrawn from candidate predictors (the order of elimination of correlated variables was given by the order of column number in selected predictors). Then, the GAMs model selected in this research was fitted by using only the not highly correlated predictors.

2.3.2 Generalized Regression Analysis and Spatial Predictions

Most analyses were performed using Generalized Re-

Table 2 Correlation analyses of collected predictors

Predictor	SP	ALT	DN_{SIR}	DN_{VR}	DN_{NIR}	DN_{VG}	DR	DC	DVI	NDVI	RVI	SAVI	SOM	SALT	pH	TN
SP	1.00															
ALT	0.33	1.00														
DN_{SIR}	0.13	0.44	1.00													
DN_{VR}	0.01	0.21	0.70	1.00												
DN_{NIR}	0.06	0.26	0.35	–0.14	1.00											
DN_{VG}	0.03	0.24	0.69	0.97	–0.09	1.00										
DR	–0.07	–0.38	0.08	0.35	–0.41	0.28	1.00									
DC	0.04	0.61	0.08	–0.22	0.46	–0.15	–0.83	1.00								
DVI	0.02	–0.01	–0.33	–0.82	0.66	–0.78	–0.49	0.43	1.00							
NDVI	0.01	–0.03	–0.32	–0.81	0.67	–0.77	–0.49	0.41	0.99	1.00						
RVI	–0.01	–0.06	–0.32	–0.79	0.68	–0.75	–0.47	0.40	0.99	0.99	1.00					
SAVI	0.01	–0.03	–0.32	–0.81	0.67	–0.77	–0.49	0.41	0.99	0.99	0.99	1.00				
SOM	–0.01	–0.03	–0.22	–0.41	0.31	–0.38	–0.69	0.53	0.48	0.49	0.47	0.49	1.00			
SALT	–0.15	–0.46	0.09	0.39	–0.25	0.30	0.69	–0.62	–0.43	–0.41	–0.38	–0.41	–0.51	1.00		
pH	0.05	0.35	–0.01	–0.16	0.28	–0.09	–0.56	0.53	0.28	0.27	0.27	0.27	0.54	–0.68	1.00	
TN	0.04	0.18	–0.13	–0.36	0.37	–0.32	–0.72	0.61	0.48	0.48	0.45	0.48	0.83	–0.61	0.69	1.00

Notes: All correlation is significant at $p < 0.05$ level. SP, ALT, DN_{SIR}, DN_{VR}, DN_{NIR}, DN_{VG}, DR, DC, DVI, NDVI, RVI, SAVI, SOM, SALT and TN are respectively abbreviation of slope, altitude, digital number of shortwave infrared band, digital number of visible red band, digital number of visible red band, digital number of near-infrared band, digital number of visible green band, nearest distance to Huanghe River, neatest distance to coastline, difference vegetation index, normalized difference vegetation index, ratio vegetation index, soil adjusted vegetation index, soil organic matter, soil salinity and soil total nitrogen

gression Analysis and Spatial Predictions (GRASP) (Lehmann *et al*., 2002), a set of S-PLUS (v.6.0-Mathsoft Inc., Seattle, Washington) functions developed to facilitate the analysis of large numbers of species. Regressions were fitted by using generalized additive models. A logistic link and binomial error term were used for individual species models and a smoothing spline method was chosen to smooth the predictors, taking four as the degree of freedom by default. The model for each species was fitted by using a backwards stepwise procedure in which all predictors were initially fitted. The significance of dropping each predictor was tested by using the analysis of deviance (ANOVA, *F*-test) to decide adding or removing them from the model. Model fitting proceeded until no more variables could be removed. Contribution of each predictor was assessed by calculating the average change in residual deviance when dropping each predictor from the final regression model.

The goodness of fitting for each GAMs model was tested by the deviance of the model (D^2), which was obtained by using the following calculation formula:

$$D^2 = (ND - RD) / ND \quad (6)$$

where *ND* is the null deviance; *RD* is the residual deviance which can not be explained by the model; *'ND–RD'* is the explained deviance. If D^2 is 1, which means no residual deviance and the deviance can be explained completely by the model.

Final regression model was then used to predict the probability of species occurrence for each grid in the study area. Based on the probability maps of species occurrence, the species with the greatest estimated probability was assigned to the pixel.

2.3.3 Model validation

Probability map of each species based on GAMs was assessed by using cross-validation of Receiver Operating Characteristic (ROC) statistics (Fielding and Bell, 1997). ROC curve is a graphical plot which illustrates the performance of a binary classifier system. It is created by plotting the fraction of true positives out of the positives (TPR is true positive rate) vs. the fraction of false positives out of the negatives (FPR is false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity (1 – specificity) or true negative rate.

Unlike Kappa statistics, the ROC method avoids the problem of choosing a threshold value, therefore, ROC is regarded as a suitable method to validate binomial model (Lehmann *et al*., 2002). Cross-validation of ROC statistics were performed with five subsets of the entire dataset, each subset containing an equal number of randomly selected data points. Each subset was then dropped from the model, and then the model was recalculated and predictions were made for the omitted data points. Combination of the predictions from the different subsets was then plotted against the observed data. It has been suggested that models running at an Area Under Curve (AUC) > 7 are acceptable (Hosmer and Lemeshow, 2000).

3 Results and Analyses

3.1 Fitted models based on GAMs

The stepwise selection of statistically significant predictors for *T. chinensis*, *P. australis* and *S. salsa* were performed by using the models in Table 3. Predictors in the fitted GAMs model for each species are identical, however, the explanatory capacity of the fitted models for different species are different. For *S. salsa* and *P. australis*, the fitted model explains 57% and 58% of the null deviance, respectively, but for *T. chinensis*, the fitted model explains 52% of the null deviance.

Table 4, Table 5 and Table 6 are ANOVA analyses for the fitted GAMs model which are constructed by testing the significance of removing in turn each predictor from the selected model. From those ANOVA tables, we could see that predictors selected into the fitted models were all confirmed by the statistical significance test. In addition, we could also get the information about the contribution of each predictor in the fitted GAMs, which were also obtained by removing each predictor from the fitted model and calculating the associated change in deviance.

For *T. chinensis* (Table 4), SALT, ALT and SOM are the most important predictors among environmental predictors, pH is the next most important, while DR is the least important; DN_{NIR} and DN_{SIR} are the most important predictors among the spectral variables.

In the fitted model for *P. australis* (Table 5), contributions of environmental predictors rank in the following order: ALT > SP > pH > DR > SOM > SALT. Among spectral predictors, DN_{VR} and DN_{NIR} make greater contribution than DN_{SIR}.

In the fitted model for *S. salsa* (Table 6), the contri-

Table 3 Selected GAMs models for *T. chinensis*, *P. australis* and *S. salsa*

Species	Number of quadrat	Model	Null deviance	Explained deviance	D^2
T. chinensis	152	s(SP, 4) + s(ALT, 4) + s(DN_{SIR}, 4) + s(DN_{VR}, 4) + s(DN_{NIR}, 4) + s(DR, 4) + s(SOM, 4) + s(SALT, 4) + s(pH, 4)	1833.25	956.11	0.52
P. australis	183	s(SP, 4) + s(ALT, 4) + s(DN_{SIR}, 4) + s(DN_{VR}, 4) + s(DN_{NIR}, 4) + s(DR, 4) + s(SOM, 4) + s(SALT, 4) + s(pH, 4)	2080.74	1204.75	0.58
S. salsa	148	s(SP, 4) + s(ALT, 4) + s(DN_{SIR}, 4) + s(DN_{VR}, 4) + s(DN_{NIR}, 4) + s(DR, 4) + s(SOM, 4) + s(SALT, 4) + s(pH, 4)	2068.68	1186.22	0.57

Notes: s in fitted model is symbol of spline smoothing curve; 4 in bracket is degree of freedom; D^2 is result of divide null deviance by explained deviance

Table 4 ANOVA analyses of selected GAMs model for *T. chinensis*

Test	*Df*	Deviance	*F* value	*p* value
–s(SP, 4)	–3.87	–24.12	15.77	< 0.001
–s(ALT, 4)	–3.73	–80.01	54.24	0
–s(DN_{SIR}, 4)	–3.76	–51.07	34.39	0
–s(DN_{VR}, 4)	–3.68	–13.13	9.03	< 0.001
–s(DN_{NIR}, 4)	–3.55	–222.01	158.11	0
–s(DR, 4)	–3.92	–11.23	7.25	< 0.001
–s(SOM, 4)	–3.85	–87.70	57.48	0
–s(SALT, 4)	–3.76	–108.99	73.25	0
–s(pH, 4)	–3.78	–57.47	38.42	0

Table 5 ANOVA analyses of selected GAMs model for *P. australis*

Test	*Df*	Deviance	*F* value	*p* value
–s(SP, 4)	–4.10	–81.83	34.32	0
–s(ALT, 4)	–4.17	–130.59	53.89	0
–s(DN_{SIR}, 4)	–4.02	–26.48	11.34	< 0.001
–s(DN_{VR}, 4)	–3.96	–52.62	22.87	0
–s(DN_{NIR}, 4)	–4.08	–48.58	20.49	< 0.001
–s(DR, 4)	–4.07	–73.38	31.00	0
–s(SOM, 4)	–3.91	–60.97	26.87	0
–s(SALT, 4)	–3.88	–23.06	10.23	< 0.001
–s(pH, 4)	–4.05	–78.98	33.57	0

Table 6 ANOVA analyses of selected GAMs model for *S. salsa*

Test	*Df*	Deviance	*F* value	*p* value
–*s*(SP, 4)	–3.93	–40.48	16.16	< 0.001
–*s*(ALT, 4)	–3.74	–88.72	37.24	0
–*s*(DN_{SIR}, 4)	–3.85	–16.22	6.62	< 0.001
–*s*(DN_{VR}, 4)	–3.91	–91.34	36.63	0
–*s*(DN_{NIR}, 4)	–3.78	–62.49	25.98	0
–*s*(DR, 4)	–3.95	–85.44	33.96	0
–*s*(SOM, 4)	–3.80	–33.15	13.71	< 0.001
–*s*(SALT, 4)	–3.99	–117.60	46.20	0
–*s*(pH, 4)	–3.93	–62.59	25.02	0

bution of SALT, ALT, DR, pH is greater than that of SP and SOM. Among spectral predictors, DN_{VR} and DN_{NIR} contribute much more than DN_{SIR}.

3.2 Predicted vegetation type based on fitted GAMs

The fitted GAMs were exported to lookup tables and the occurrence probability maps of *T. chinensis*, *P. australis* and *S. salsa* were produced in Arcview 3.2 (Fig. 3 to Fig. 5).

From Fig. 3 to Fig. 5, we could see that the occurrence probability of *S. salsa* is very high in the area near the coastline. With the increase of the distance to the coastline, the occurrence probability of *T. chinensis* and *P. australis* become higher and higher. This means that from coastline to land, the dominant plant species change gradually from *S. salsa* to *T. chinensis*, and then *P. australis.* This kind of vegetation distribution pattern is especially evident in the northern part of the study area. This maybe due to the long distance from the northern part to the Huanghe River, and the vegetation distribution is little affected by the Huanghe River. In the southern part of the study area, the vegetation distribution pattern is greatly affected by the Huanghe River and the occurrence probability of *T. chinensis* and *P. australis* is very high.

The predicted vegetation distribution pattern is in agreement with the characteristics of the actual vegetation distribution in the Huanghe River Delta. The Huanghe River Delta is formed by the deposition of a large amount of sand and mud transported by the Huanghe River. The downstream movement of freshwater combined with the inland movement of saline water from the ocean generates a salinity gradient in the estuarine systems (Yue *et al*., 2003). The salt content in the soil of the newly deposited land is more than 3%, on which *S. salsa* is partially distributed. The *S. salsa* increases organic matter in the soil which makes the area

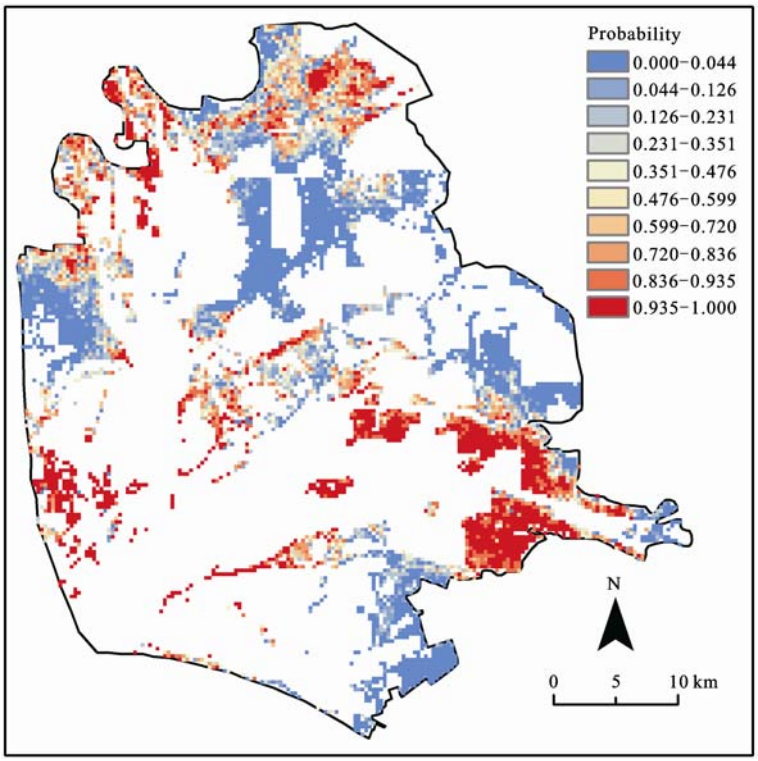

Fig. 3 Predicted occurrence probability of *T. chinensis* based on GAMs

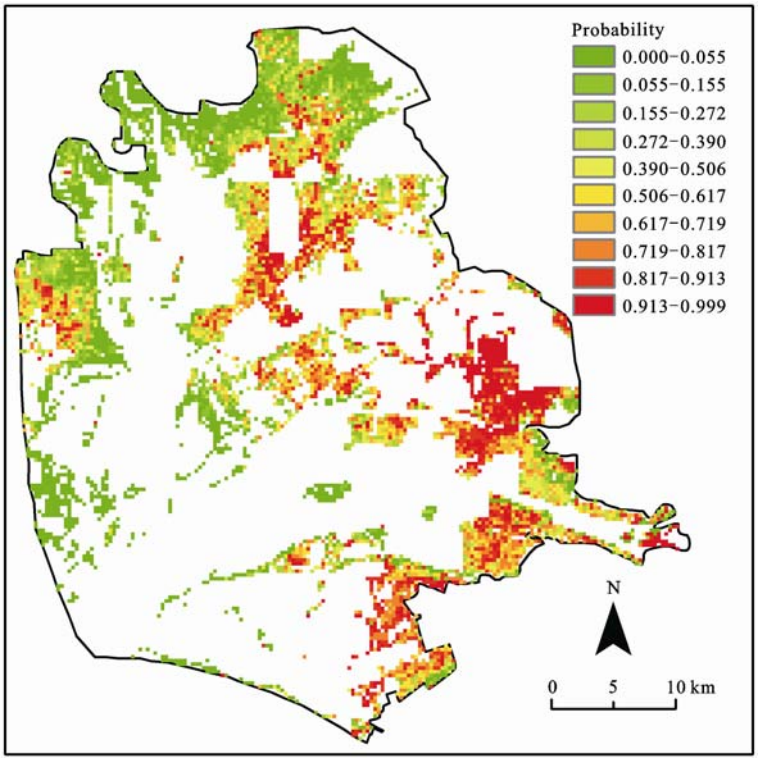

Fig. 4 Predicted occurrence probability of *P. australis* based on GAMs

become the *T. chinensis* land. The secreting salt effects of *T. chinensis* and accumulation of their dead branches and leaves result in the reduction of soil salt content and the increase in soil fertility, which makes the area

evolve to the *P. australis* land.

At last, according to the occurrence probability maps of *T. chinensis*, *S. salsa* and *P. australis*, the vegetation type of each grid was determined (Fig. 6).

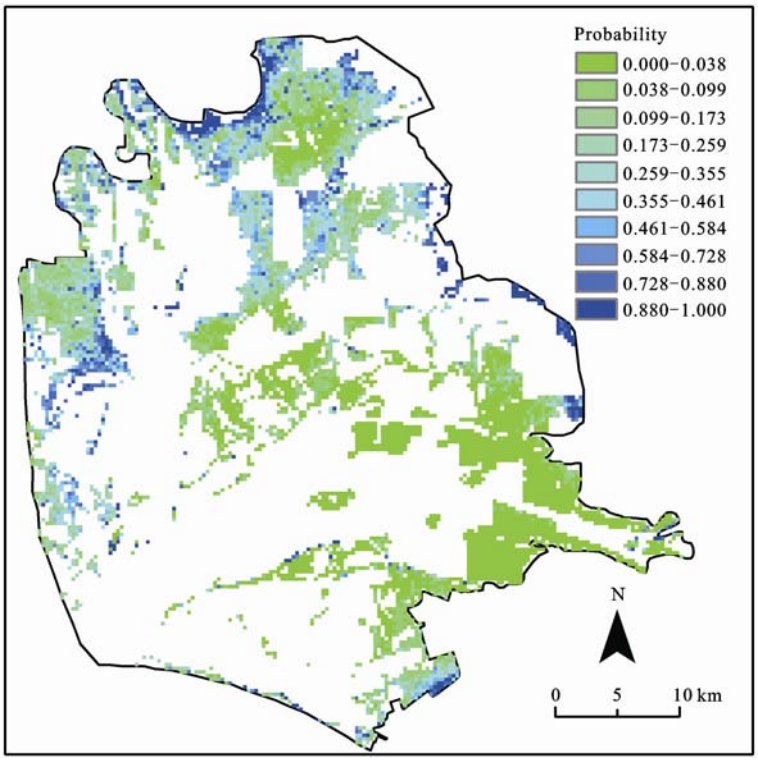

Fig. 5 Predicted occurrence probability of *S. salsa* based on GAMs

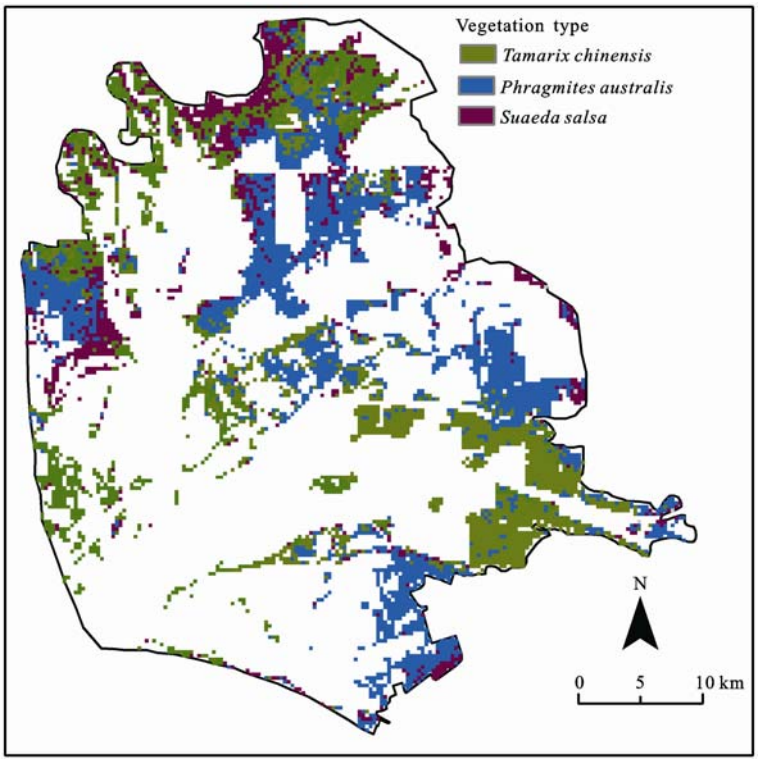

Fig. 6 Predicted vegetation type map based on species occurrence probability

3.3 Model validation based on ROC curve

The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a negative one (Fawcett, 2006). The area under the ROC curve (AUC) results were considered excellent when AUC values between 0.9 and 1.0, good when between 0.8 and 0.9, fair when between 0.7 and 0.8, poor when between 0.6 and 0.7 and failed when between 0.5 and 0.6 (Obuchowski, 2003). According to this, the accuracy of the fitted GAMs models in this study could be all classified into the excellent class (Fig. 7).

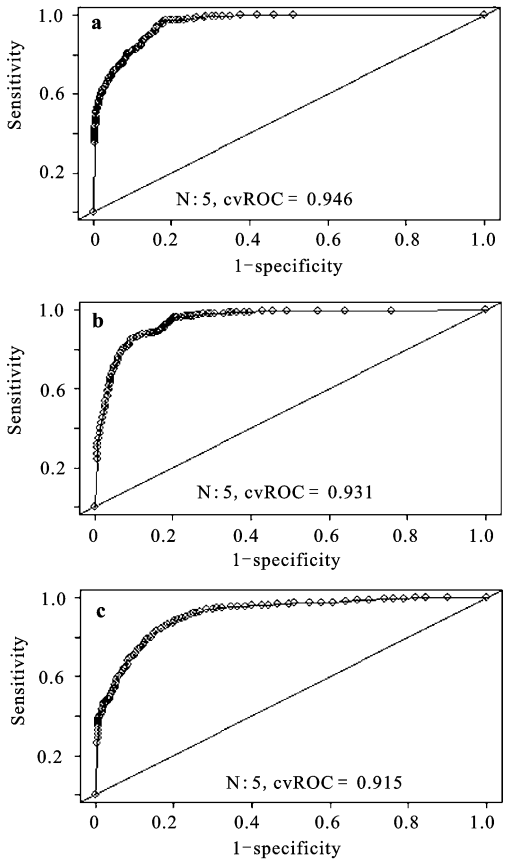

Fig. 7 Cross validation of fitted GAMs models based on Receiver Operating Characteristic (ROC) curve for *P. australis* (a), *S. salsa* (b) and *T. chinensis* (c)

4 Discussion

4.1 Performance of fitted GAMs model

Classical vegetation mapping methods, such as supervised classification and non-supervised classification, only utilize spectral features to classify vegetation and ignore the relation between vegetation and environment. In this research, the corresponding relations between vegetation and environmental variables as well as spectral features, were used to map the vegetation. Hence, theoretically, the approach adopted in this research is robust for vegetation mapping, and the results also approve the practicability and effectiveness of this method. The D^2 values are all higher than 0.5 (Table 4), which indicates that more than half of the deviance can be explained by the fitted GAMs model, the percent of the explained deviance is higher than the similar researches (Shen and Zhao, 2007; Wehrden *et al*., 2009). The AUC values are all higher than 0.9 (Fig. 7) and this means that the prediction results are high accuracy and acceptable (Hosmer and Lemeshow, 2000). Therefore, we can say that integrating vegetation survey data, spectral data and environmental data by GAMs is a feasible approach to map vegetation.

In this research, prediction accuracy varies with vegetation type. *P. australis* and *S. salsa* are better predicted than *T. chinensis*. One cause is the difference in the amount of training samples, since existing researches have proved that the map quality might be improved when quadrats are collected at higher density (Guisan and Zimmerman, 2000; Pfeffer *et al*., 2003; Miller *et al*., 2007). Another cause might be the difference in the plant community structure (Dirnböck *et al*., 2003). Sparse plant communities with low coverage yield the phenomenon of diffuse background spectral reflectance and bring out the effects of mixed spectrum. Therefore, the discriminating power of spectral variables becomes relatively weak for open vegetation types (Goward *et al*., 1994). In the presented study, the habitat of *T. chinensis* lacks a dense vegetation cover and the effect of mixed spectrum is very strong, which lead to the weak correlation between vegetation and spectral variables. On the contrary, *P. australis* meadow and *S. salsa* community have dense vegetation cover, the effects of mixed spectrum are weak and spectral features have stronger ability to delineate vegetation patches.

In addition, we found that predictors in the GAMs model of different species are the same, but the contribution of the same predictor in different GAMs model is different, this might indicate that there are differences in mesophyll structure and environment requirements among those vegetation types.

4.2 Prediction of vegetation type

Most researches in plant ecology employed GAMs to

predict the species occurrence probability instead of vegetation type (Shen and Zhao, 2007; He *et al*., 2008; Wen *et al*., 2008). However, this is not appropriate for all application of conventional vegetation maps, for example, in conservation, it is necessary for us to know whether a place does or does not belong to a certain vegetation type. In such cases, vegetation map with clear class may be appropriate (Schmidtlein *et al*., 2007). Therefore, in this research, we defined the vegetation type of each prediction cell according to the probability of species occurrence, and the species with the greatest estimated probability was assigned to the pixel. However, we should admit that it is insufficient for us to determine the vegetation type based on the probability of species occurrence, inter-specific relationships should also be involved in the prediction of vegetation type.

4.3 Effects of response variables on prediction

Response variable required in this research is binomial data (presence/absence), so the generation of response variables has great effects on the performance of the fitted GAMs model.

In order to generate the required binomial data for the GAMs model (take *T. chinensis* for instance), we specified that if the importance value (IV) of *T. chinensis* is higher than other species in the quadrat, then this quadrat was assigned with 1, otherwise, the quadrat was assigned with 0. However, if there are two or more dominant species in the same quadrat, for example, *T. chinensis* and *P. australis* are both dominant species in a quadrat, but the importance value of *P. australis* is slightly lower than *T. chinensis*, so this quadrat was still assigned with 1. While the spectral features derived by this quadrat represent the spectral characteristics of *P. australis* to a great extent, therefore, the correlation between spectral features and *T. chinensis* was relaxed and the discernment capacity of the spectral variables decreased (Treitz *et al*., 1992). Futuremore, information loss due to the transition from species-plot matrix to binomial data (presence/absence) weakened the relation between plant species and spectral features, this also decreased the discernment of the spectral variables.

4.4 Effects of predictor variables on prediction

The multi-spectral image used in this research has only four wavebands with 10 m spatial resolution. The low spectral and spatial resolutions limit the discriminating power of the spectral variables. The spatial resolution of the DEM (with 50 m horizontal resolution) used in this study is not fine enough to describe the micro-topography and part of the micro-scale variation in topography missed. For soil variables, the number of soil samples limits the prediction accuracy of the interpolation by ordinary Kriging. Hence, inadequate resolution of predictors is partially responsible for the unexplained variation.

In this research, the quadrat size, the spatial resolutions of DEM and remote sensing image are different. The optimal condition is that resolutions of all the data are the same, but such coherence is always impossible. Even the quadrat size, the spatial resolution of DEM and the spatial resolution of remote sensing image are the same, spatial error of the remote sensing image or imprecision in the location of the quadrat can also relax the field-to-image correlation (Peleg and Anderson, 2002; Weber, 2006). Another problem is the temporal difference between the field plot measurement and remote sensing image. Ideally, to strengthen the field-to-imagery correlation, the field data and remote sensed data should be collected at the same time. However, it is impossible for us to collect all the required field data at the same time when the remote sensing image is acquired, especially for large area. This makes the spectral properties derived from the satellite image can not represent the actual spectral features of the vegetation and relax the field-to-imagery correlation exactly (Ohmann and Gregory, 2002; Karl, 2010).

Furthermore, even if resolution mismatch does not exist and the resolution of DEM is fine enough to reflect micro-scale variations controlling the distribution of plant community, it can not account for all of them. The reason is that biological factors, land use and other disturbances have great effects on the distribution of plant community (Dirnbök *et al*., 2003; Dullinger *et al*., 2003). Therefore, the influences of land use and other variables which are not closely related to environmental conditions also limit the explanatory power of environmental and spectral variables (Dirnböck *et al*., 2002).

5 Conclusions

In this research, GAMs were used as an analysis tool to integrate environmental factors and spectral factors to predict the vegetation distribution in the Huanghe River

Delta. The results proved that: 1) Integrating vegetation survey data, environmental data and spectral data based on GAMs is a practical way to map the vegetation. 2) Prediction accuracy varies with community type, and community structure has great effects on the performance of GAMs model. 3) Inaccuracies could not be accounted for by environmental descriptors and spectral variables, confounding effects of additional controls like land use and disturbance, also have certain contribution to the imprecision. This study provides a promising illustration of the power of combining spectra data and environmental data on vegetation mapping with GAMs. This approach will be helpful for the researches of vegetation ecology and remote sensing classification.

References

Austin M P, 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modeling. *Ecological Modeling*, 157(2–3): 101–118. doi: org/10.1016/S0304-3800(02)00205-3

Braun-Blanquet J, 1933. Phytosociological nomenclature. *Ecology*, 14(3): 315–317. doi: org/10.2307/1932802

Dirnböck T, Dullinger S, Gottfried M *et al*., 2003. Mapping alpine vegetation based on image analysis, topographic variables and Canonocal Correspondence Analysis. *Applied Vegetation Sciences*, 6(1): 85–96. doi: 10.1111/j.1654-109X.2003.tb00567.x

Dirnböck T, Hobbs R J, Lambeck R J *et al*., 2002. Vegetation distribution in relation to topographically driven processes in south-western Australia. *Applied Vegetation Science*, 5(1): 147–158. doi: 10.1111/j.1654-109X.2002.tb00544.x

Dirnbök T, Dullinger S，Grabherr G, 2001. A new grassland community in the Eastern Alps (Austria): Evidence of environmental distribution limits of endemic plant communities. *Phytocoenologia*, 31(4): 521–536.

Dullinger S, Dirnbök T, Grabherr G, 2003. A resampling approach for evaluating effects of pasture abandonment on subalpine plant species diversity. *Journal of Vegetation Science*, 14(2): 243–252. doi: 10.1111/j.1654-1103.2003.tb02149.x

Fawcett T, 2006. An introduction to ROC analysis. *Pattern Recognition Letter*, 27(8): 861–874. doi: 10.1016/j.patrec.2005.10.010

Ferrier S, Guisan A, 2006. Spatial modeling of biodiversity at the community level. *Journal of Applied Ecology*, 43(3): 393–404. doi: 10.1111/j.1365-2664.2006.01149.x

Ferrier S, Watson G, Pearce J *et al*., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: The north-east New South Wales experience. I. Species-level modelling. *Biodiversity Conservation*, 11(12): 2275–2307. doi: 10.1023/A:1021302930424

Fielding A H, Bell J F, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environment Conservation*, 24(1): 38–49. doi: 10.1017/S0376892997000088

Goward S N, Huemmerich K F, Waring R H, 1994. Visible-near infrared spectral reflectance of landscape components in West Oregon. *Remote Sensing of Environment*, 47(2): 190–203. doi: 10.1016/0034-4257(94)90155-4

Guisan A, Zimmermann N E, 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2–3): 147–186. doi: 10.1016/S0304-3800(00)00354-9

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

He Xiaohui, Wen Zhongming, Wang Jinxin, 2008. Spatial distribution of major grassland species and its relations to environment in Yanhe River catchment based on generalized additive model. *Chinese Journal of Ecology*, 27(10): 1718–1724. (in Chinese)

Hosmer D W, Lemeshow S, 2000. *Applied Logistic Regression*. New York: Wiley & Sons.

Joseph C G, Gary D R, 2004. *Expert System*: *Principles and Programming*. Stamford: Thomson Learning.

Karl J W, Maurer B A, 2010. Spatial dependence of predictions from image segmentation: A variogram-based method to determine appropriate scales for producing land-management information. *Ecological Informatics*, 5(3): 194–202. doi: 10.1016/j.ecoinf.2010.02.004

Lehmann A, Overton J M, Leathwick J R, 2002. GRASP: Generalized regression analysis and spatial prediction. *Ecological modelling*, 157(2–3): 189–207. doi: 10.1016/S0304-3800(02)00354-X

Liu Guangsong, 1996. *Soil Physical and Chemical Analysis & Description of Soil Profiles*. Beijing: China Standards Press. (in Chinese)

Miller J, Franklin J, Aspinall R, 2007. Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, 202(3–4): 225–242. doi: 10.1016/j.ecolmodel.2006.12.012

Obuchowski N A, 2003. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1): 3–8. doi: 10.1148/radiol.2291010898

Ohmann J L, Gregory M J, 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, USA. *Canadian Journal of Forest Research*, 32(4): 725–741. doi: 10.1139/x02-011

Peleg K, Anderson G L, 2002. FFT regression and cross-noise reduction for comparing images in remote sensing. *International Journal of Remote sensing*, 23(10): 2097–2124. doi: 10.1080/01431160110075910

Pfeffer K, Pebesma E J, Burrough P A, 2003. Mapping alpine vegetation using vegetation observation and topographic attributes. *Landscape Ecology*, 18(8): 759–776. doi: 10.1023/B:LAND.0000014471.78787.d0

Sanders M E, Dirkse G M, Slim P A, 2004. Objecting thematic, spatial and temporal aspects of vegetation mapping for monitoring. *Community Ecology*, 5(1): 81–91. doi: 10.1556/ComEc.

5.2004.1.8

Schmidtlein S, Zimmermann P, Schupferling R *et al*., 2007. Mapping the floristic continuum: Ordination space position estimated from imaging spectroscopy. *Journal of Vegetation Science*, 18(1): 131–140. doi: 10.1111/j.1654-1103.2007.tb02523.x

Shen Zehao, Zhao Jun, 2007. Prediction of the spatial patterns of species richness based on the plant topography relationship: An application of GAMs approach. *Acta Ecologica Sinica*, 27(3): 953–962. (in Chinese)

Song C H, Woodcock C E, Seto K *et al*., 2000. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects? *Remote Sensing of Environment*, 75(2): 230–244. doi: 10.1016/S0034-4257(00)00169-3

Song C Y, Liu G H, Liu Q S, 2009. Spatial and environmental effects on plant communities in the Yellow River Delta, Eastern China. *Journal of Forest Research*, 20(2): 117–122. doi: 10.1007/s11676-009-0021-3

Song Yongchang, 2001. *Vegetation Ecology*. Shanghai: East China Normal University Press. (in Chinese)

Treitz P M, Howarth P J, Suffling R C, 1992. Application of detailed ground information to vegetation mapping with high resolution digital imagery. *Remote Sensing of Environment*, 42(1): 65–82. doi: 10.1016/0034-4257(92)90068-U

Weber K T, 2006. Challenges of integrating geospatial technologies into rangeland research and management. *Rangeland Ecology and Management*, 59(1): 38–43. doi: 10.2111/05-010R.1

Wehrden H V, Zimmermann H, Hanspach J *et al*., 2009. Predictive mapping of plant species and communities using GIS and Landsat data in a southern Mongolian mountain range. *Folia Geobotanica*, 44(3): 211–225. doi: 10.1007/s12224-009-9042-0

Wen Zhongming, Jiao Feng, Jiao Juying, 2008. Prediction and mapping of potential vegetation distribution in Yanhe River catchment in hilly area of Loess Plateau. *Chinese Journal of Applied Ecology*, 19(9): 1897–1904. (in Chinese)

White P S, Wilds S P, Stratton DA, 2001. The distribution of heath balds in the Great Smoky Mountains, North Carolina and Tennessee. *Journal of Vegetation Science*, 12(4): 453–466. doi: 10.2307/3236997

Yang Cunjian, Zhou Chenghu, 2001. Investigation on classification of remote sensing image on basis of knowledge. *Geography and Territorial Research*, 17(1): 72–77. (in Chinese)

Yue T X, Liu J Y, Jorgensen S E *et al*., 2003. Landscape change detection of the newly created wetland in the Yellow River Delta. *Ecological Modeling*, 164(1): 21–31. doi: 10.1016/S0304-3800(02)00391-5

Zang Qiyun, 1996. *Near Shore Sediment along the Yellow River Delta*. Beijing: Ocean Press. (in Chinese)

Zhang Jintun, 1995. *Quantitative Method of Vegetation Ecology.* Beijing: China Science and Technology Press. (in Chinese)

Zhang R Q, Zhu D L, 2011. Study of land cover classification based on knowledge rules using high-resolution remote sensing images. *Expert System with Application*, 38(4): 3647–3652. doi: 10.1016/j.eswa.2010.09.019

Zhao Yingshi, 2003. *The Principle and Method of Analysis of Remote Sensing Application*. Beijing: Science Press. (in Chinese)