

## Habitat modelling in GIMOLUS – webGIS-based e-learning modules using logistic regression to assess species-habitat relationships

Michael RUDNER<sup>1</sup>, Boris SCHRÖDER<sup>1</sup>, Robert BIEDERMANN<sup>1</sup>, Mark MÜLLER<sup>2</sup>

<sup>1</sup> Landscape Ecology Group, Carl von Ossietzky University of Oldenburg, P.O. Box 2503, D-26111 Oldenburg, Germany, michael.rudner@uni-oldenburg.de

<sup>2</sup> ILPÖ, University of Stuttgart, Keplerstraße 11, D-70174 Stuttgart

The GIMOLUS-project is funded by the German Ministry of Education and Research (BMBF), FKZ: 08NM143



### Abstract

Habitat modelling is a tool for regionalisation of biotic information by predicting the spatially explicit distribution of species on the basis of environmental properties. Since this recent method has a high relevance in questions of nature conservation and ecological research, there is a need to teach it in environmental studies. The multimedia GIMOLUS-environment provides all elements necessary to teach and learn the application of this method in a realistic setting. It's constituting an innovation to environmental courses as it introduces the proceeding of a complete modelling procedure on practical examples offering a high interactivity level on low technical requirements. The internet-based learning unit 'habitat modelling with logistic regression' uses webGIS applications to illustrate the sampling procedure and the spatial extrapolation of habitat models. The learning unit comprises six learning modules. Spatially explicit habitat modelling is presented starting from sampling, through model evaluation, to predictions of the probability of species occurrence. Step by step, the complete model calculation as well as the evaluation are proceeded in the GIMOLUS-environment, as shown in an exemplary case study. The data sets used in the analyses of species-habitat-relationships are sampled in a virtual landscape.

### Zusammenfassung

Es wird eine internet-basierte Lerneinheit vorgestellt, die an entscheidenden Stellen WebGIS einsetzt. Die Einführung der vollständigen Bearbeitung eines Modellierungsverfahrens an praxisnahen Beispielen bei zugleich hohem Grad an Interaktivität und geringen technischen Anforderungen stellt eine Neuerung für umweltorientierte Studiengänge dar. Diese Einheit ist in sechs Module gegliedert. Die räumlich explizite Habitatmodellierung wird von der Probenahme in der virtuellen Landschaft mit WebGIS über die Modellbildung und die räumliche Extrapolation wiederum im WebGIS bis zur Validierung vermittelt. Prognosekarten für das Auftreten von Arten werden erzeugt. Die gesamte Modellrechnung und -bewertung wird schrittweise in der GIMOLUS-Umgebung durchgeführt, wie an einer Fallstudie gezeigt wird. Die Datensätze für die Analyse von Art-Habitat-Beziehungen werden aus der virtuellen Landschaft erzeugt oder in der Geodatenbank vorgehalten.

### Keywords

webGIS, e-learning, habitat models, logistic regression, spatial extrapolation, validation, regionalisation, spatial sampling

## 1 Introduction

Habitat models serve two complementary yet related purposes: First, they serve as a tool for regionalisation of biotic information by predicting the spatially explicit distribution of species on the basis of environmental properties. Second, they improve our understanding of species-habitat relationships and provide quantitative descriptions of habitat requirements (Morrison et al. 1998). They may directly predict the effect of landscape change on species distribution (Schröder 2000), and serve as a basis for population dynamic models in changing landscapes (Akçakaya et al. 1995, Söndgerath & Schröder 2002). Thus, habitat modelling plays a substantial role in modern ecological research and conservation biology (Scott et al. 2002). To provide students with this up-to-date methodology, habitat modelling should become part of the curriculum in environmental study courses. Integrated courses that reflect the statistical background, teach the model building procedure and deal with spatial extrapolation via GIS are scarce. Teaching of habitat modelling in environmental courses lacks the training on detailed practical examples, whereas the methods sometimes are taught in statistics lessons using chiefly medicinal examples. The presented learning unit will fill this gap, interactively working through the habitat modelling procedure at examples located in a virtual landscape. Hereby we follow the advice of the German Science Council (Wissenschaftsrat 1998), to develop multimedia learning modules that promote problem-oriented and interdisciplinary learning. The GIMOLUS-project offers the possibility to impart spatially explicit habitat modelling without requiring access to statistical software or GIS. The internet-based learning modules are realised in XML with embedded interactive elements in JavaScript or Java. The modules are linked to the virtual landscape which may be accessed via the ArcIMS®-webGIS. Customised webGIS functions for sampling procedures or spatial extrapolation of habitat models are coded in PHP with embedded SQL.

## 2 Module structure

### 2.1 GIMOLUS-project

The objective of the GIMOLUS-project ('GIS- und modellgestützte Lernmodule für umweltorientierte Studiengänge' i.e. 'learning modules based on GIS and modelling in environmental courses') is to provide multimedia modules and learning units relevant to different aspects of general or specific environmental studies. The modules deal with GIS and modelling issues which are embedded in a webGIS-based virtual landscape. The students are thought to use e-learning modules as a complement to traditional courses. A specific web-based system is built up by the GIMOLUS-team to provide the administration of users and content. All modules are defined according to an XML-structure with single pages being generated dynamically on request and sent in HTML-format.

Each module is a small unit that the user may proceed within 30 minutes to one hour time. Groups of modules serve as learning units to give a comprehensive illustration of relevant issues of a scientific topic. Finally, modules of different topics can be combined to demonstrate complex, interdisciplinary relationships.

GIMOLUS-modules are interconnected and located in a virtual landscape which is built up on a realistic map basis (Elsenz catchment in the Kraichgau region, south-west Germany).

Environmental data providing the setting for different didactic purposes are surveyed in or referred to the virtual landscape. The users access to this virtual landscape via the webGIS-system ArcIMS . This 2D-system enables the display of thematic maps as well as their storage in a geo-database. Spatially explicit results of model applications may also be displayed in the browser via the webGIS-application. In the case of interactive exercises, the users may save their specific progress and results in the databases for a limited period of time (Vennemann & Müller 2002).

## 2.2 Technical requirements

The technical requirements on the hardware and software of the users are low. Provided with a GIMOLUS-account the user needs not more than access to the internet as well as an up-to-date internet browser provided with i) a Java virtual machine, ii) JavaScript activated, and iii) a flash-plugin. Additionally, a standard text editor and a spreadsheet are needed to print results of computations or to prepare data sets.

## 2.3 Habitat modelling with logistic regression

Among other methods (e.g. canonical correspondence analysis Dullinger et al. 2001), logistic regression is a well established method to perform habitat modelling (Trexler & Travis 1993, Pearce & Ferrier 2000). A recent review of the literature reveals that logistic regression is the most frequently used statistical technique in this context. It is a simple and robust procedure, and yields comparatively high performance as well as interpretable model parameters (Manel et al. 1999). In addition, excellent documentation as well as availability in the standard statistical software packages may also explain the frequent application of logistic regression. Since this method is implemented in the GIMOLUS via JavaScript access to software packages is not necessary during the teaching/learning process.

Logistic regression is used in analyses where the dependent variable,  $y$ , has only two possible values, e.g. in our case: the presence or absence of a species. The probability of one of the two states, here: the probability of occurrence  $\pi(\bar{x}) = Prob(y = 1 | \bar{x})$ , is assumed to be a function of one or more independent explanatory variables  $\bar{x}$ , which is the vector of  $k$  environmental variables ( $x_j$  with  $j$  ranging from 1 to  $k$ ). The specific function estimated by logistic regression is given in Eq. (1).

$$\pi(\bar{x}) = Prob(y = 1 | \bar{x}) = \frac{e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}} \quad (1)$$

It can be obtained by transforming a linear regression for a logit function (see Eq. (2), Hosmer & Lemeshow 2000).  $\beta_j$  designates the regression coefficient estimated for the  $j^{\text{th}}$  habitat factor, which are usually estimated using the maximum likelihood method.

$$\text{logit}(\pi(x)) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (2)$$

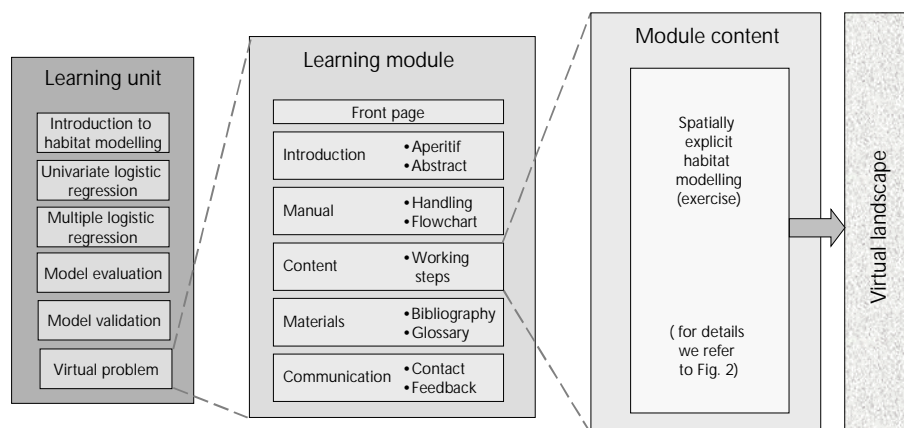
## 2.4 Workflow of spatially explicit habitat modelling

The workflow of habitat modelling with spatial extrapolation (see box 'module content' in Fig. 2) starts with defining the scientific problem. The user selects a target object (e.g. a plant

species) for which he wants either to quantify the species-habitat relationships, or to analyse and predict its spatial distribution, or to design measures to optimise its habitat. Then the user has to collect data following a specified sampling design. Each user samples her/his own data set. Model building starts with estimation and visualisation of univariate models regarding single environmental variables hypothesised to affect the species' distribution. The next step will be the selection of explanatory variables in a multivariate model, where the user may follow a stepwise selection method. The next step will be the calculation and assessment of the model's goodness-of-fit regarding calibration and discrimination. The final model needs to be interpreted ecologically, since statistically significant correlations do not necessarily reflect causal relations. If regionalisation was formulated as an aim, the final model has to be extrapolated from point samples to the study area by applying the model on maps of the independent variables. The resulting map of predicted probabilities of occurrence of the target species may be classified to yield a map of predicted incidences using a cut-off value depending on the model evaluation. Based on these maps the model should be validated comparing the predicted probabilities of occurrence with observed incidences.

## 2.5 Learning unit 'Habitat modelling with logistic regression'

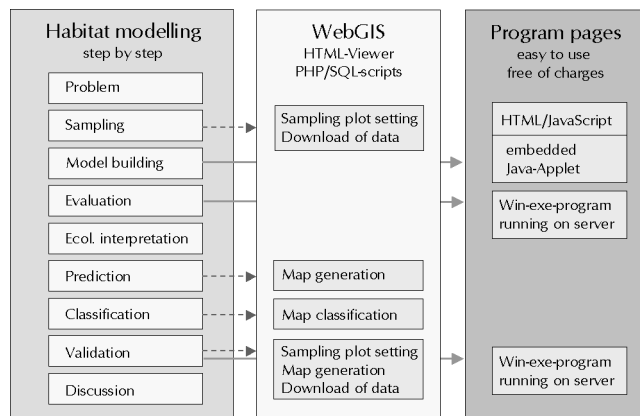
Our learning unit 'Habitat modelling with logistic regression' consists of six learning modules (Fig. 1). An introductory module is followed by modules dealing with single steps of the logistic regression method and modelling procedures as described above. The last element of the learning unit is the so-called 'virtual problem', i.e. an exercise that comprises all steps of the modelling procedure from data sampling in the virtual landscape to validation and discussion of specific modelling results.



**Fig. 1:** Structure of the learning unit 'Habitat modelling with logistic regression' and a general learning module.

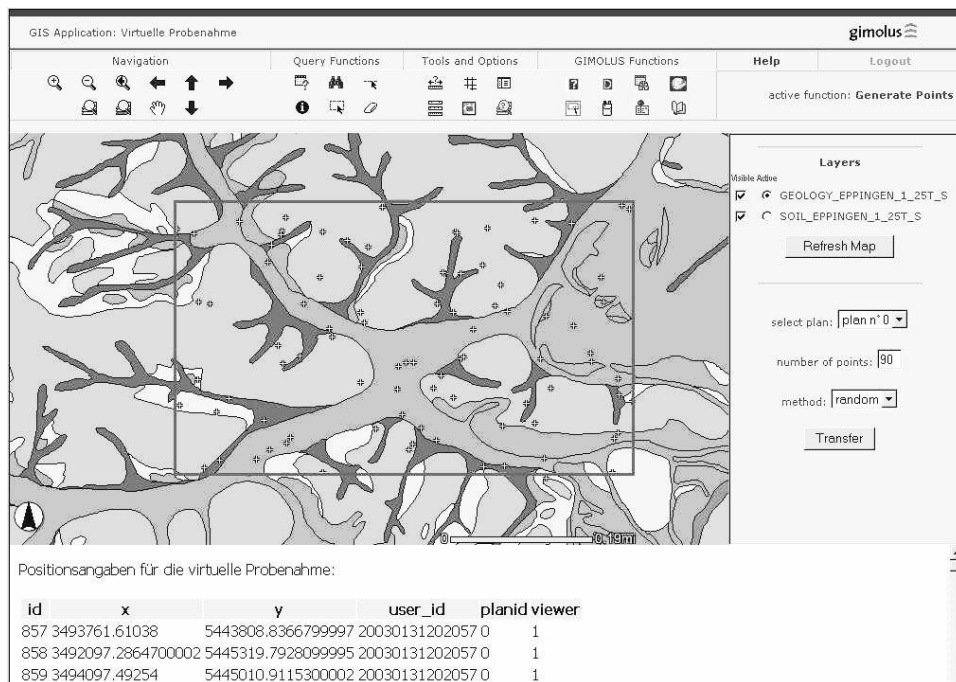
## 2.6 Workflow of the module 'virtual problem'

The structure of the virtual problem follows the procedure of habitat modelling. Step by step, the user will proceed in the module and is taught to develop a habitat model. Fig. 2 illustrates the flowchart of procedures involved with its corresponding technical realisation.



**Fig. 2:** Technical realisation of the working steps specified for the module content 'habitat modelling'.

The user is asked to analyse the species-habitat relationship of *Salvia pratensis*, a perennial plant species, in one part of the virtual landscape. The module starts with the selection of environmental factors, that are supposed to play a key-role for the presence of *S. pratensis*. The following module page describes the sampling procedure in the virtual landscape. From this page the user has direct access to the webGIS that will be displayed in a new browser window (Fig. 3).



**Fig. 3:** WebGIS sampling page.

The user will select the relevant factors in the theme overview. With the number of sampling points being specified, three sampling strategies can be selected: random sampling, regular grid sampling, or sampling along a linear transect. The sampling procedure is coded in PHP with embedded SQL. The buttons in the GIMOLUS-functions-group offer the sampling functionality step by step. The user has to activate the sampling, to delimit the study area and to evaluate the data on the sampling points. The sampled data are transferred to a table and the frequency distribution may be charted.

The user might enlarge the downloaded data set for columns that will contain the squared values of the sampled factors, if unimodal models have to be estimated. To get an overview about the data, the user shall determine univariate models (sigmoidal and unimodal relationships) for the whole set of sampled variables in a first step. The module provides a program page to realise the logistic regression method (Fig. 4).

**Exercise on multivariate modelling (stepwise forward)**

Problem (overview)

- You have recorded (or chosen) a data set, that comprises incidences of your target-object (dependent variable) and values for the independent variables (metric scale).
- Generate a multivariate model using the logistic regression method. Please use the javascript-application on this page. Select the variables for the model following a stepwise forward procedure.
  - 1. step: univariate models with linear or additionally squared consideration of the independent variable.
  - 2. step: starting with the best univariate model other variables are added stepwise. Continue if the variable is contributing significantly to the model, otherwise remove the variable.
- Select the best model, formulate the model and note the formula for subsequent visualisation.
- Use the check-boxes for variable selection.

Variable selection      Data      dependent variable:

	N	C_N	C_N_sq	WI	S_exp	bomae	Basens	KAKpot	KAK_sq	pH
1	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2		<input checked="" type="checkbox"/>								
3			<input checked="" type="checkbox"/>							
4				<input checked="" type="checkbox"/>						
5					<input checked="" type="checkbox"/>					
6						<input checked="" type="checkbox"/>				
7							<input checked="" type="checkbox"/>			
8								<input checked="" type="checkbox"/>		
9									<input checked="" type="checkbox"/>	
10										<input checked="" type="checkbox"/>

Number of selected independent variables:

Enter the number of data points:

1. row contains columns' headers

Results of the parameter estimation

Click the solve-button. The results will appear below.

Descriptive statistics

dependent variable:   
 23 cases have Y=0; 20 cases have Y=1.

Variable	Avg	SD
N	72.4419	28.0534

Model parameters (significance, coefficients...)

predictors: N pH  
 dependent variable:

-2 Log Likelihood...  
 59.4012 Null Model  
 37.4203 Current Model

Likelihood Ratio Test...  
 Chi<sup>2</sup> df p  
 21.9809 2 0.0000 Overall model fit

Data series (input data, conditional probability)

N	pH	salv_prat	Calc Prob
120.0000	6.8000	0	0.1314
128.0000	6.6000	0	0.0589
111.0000	6.9000	0	0.2382
115.0000	6.3000	0	0.0649

-2LL-value of the model - will be used for the rating of model improvement while adding a variable:  
 (Use a dot as decimal separator!)

Hints on input

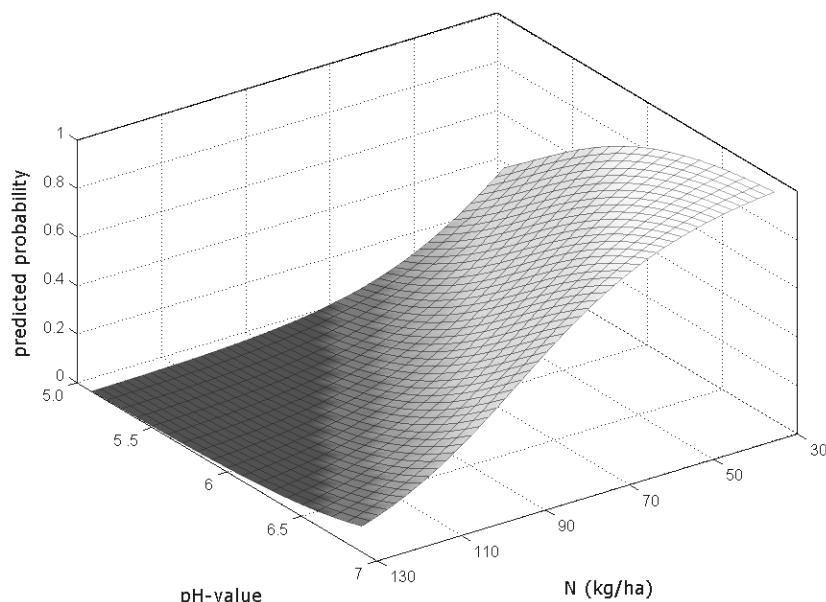
Hints on printing and saving the results

Sources:  
 The javascript-code is based on an interactive web-page by John C. Pezzullo, Georgetown University, Washington DC USA, with the permit of the author.  
 Logistic regression Web Page by John C. Pezzullo

Fig. 4: Logistic regression modelling page.

The procedure estimating parameters of the logistic regression model is coded in JavaScript and embedded in a HTML-page. The code is based on a JavaScript program written and published by John Pezzullo (2002). Data sets up to ten independent variables may be processed. After selecting the variables in question the program may be started. The results are presented in three groups: descriptive statistics, model specifications, and data. The univariate models may be visualised in an embedded Java-Applet (Orr 2002), that has been presented to the user already in the module on univariate logistic regression.

The next module page gives the instructions for stepwise variable selection. One by one the variables of the best univariate model shall be integrated to the multivariate model depending on significant model improvement. The visualisation of multivariate models is limited to bivariate plots, so-called 'response surfaces' (Fig. 5).



**Fig. 5:** Exemplary response surface depicting the probability of occurrence of *Salvia pratensis* depending on two habitat factors (N-content, pH-value).

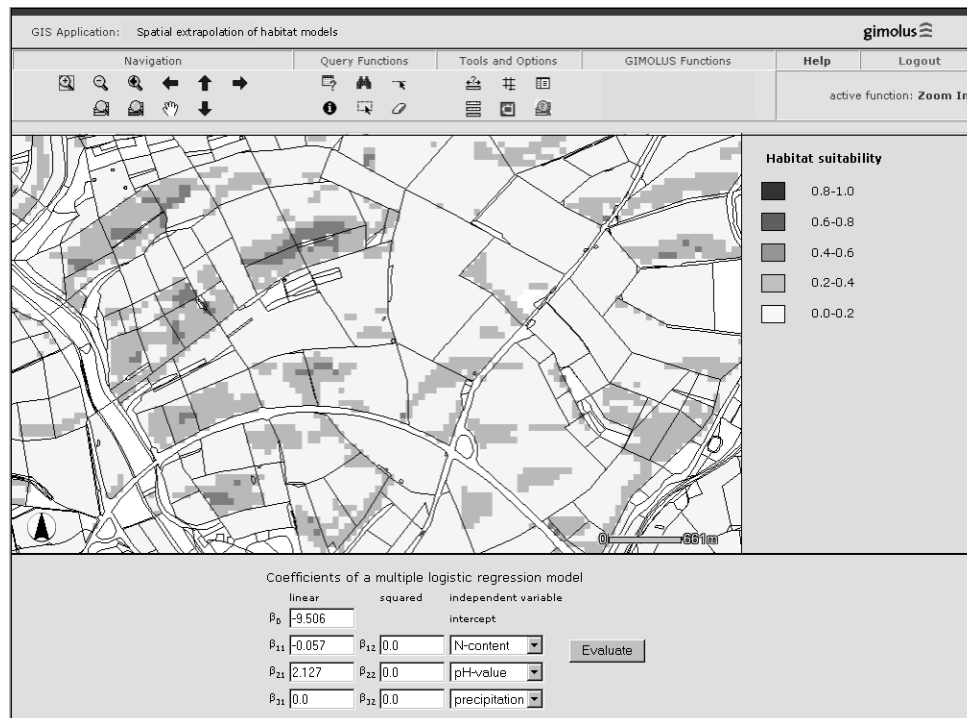
Observed incidences and calculated probabilities will be compared to evaluate the model. The corresponding data are provided in the output window of the logistic regression page. Assessment measures are calculated in the ROC-AUC-program (Schröder 2002), that can be run on the server as executable windows-file. The display of the graphical user interface is routed to the client via a Citrix-Metaframe XP™ connection. The ROC-AUC-program yields an assessment of model discrimination, i.e. the model's ability to correctly predict the species' distribution (AUC-value, etc. cf. Tab. 1).

**Tab. 1:** Model specifications (regression coefficients, goodness-of-fit, ROC-curve).

variable	regression coefficients	standard error	p-Value (Wald-test)
-	$\beta_0 = -9.506$	-	-
N (kg/ha)	$\beta_1 = -0.057$	0.017	0.001
pH-value	$\beta_3 = 2.127$	0.963	0.027
cut-off value $p_{crit}$	0.42	Cohen's $\kappa$	0.72
% correct	86.05	$R^2$ Nagelkerke	0.534

The ROC curve shows a high performance model with an AUC of 0.874. The y-axis is 'sensitivity' (0 to 1) and the x-axis is '1-specificity' (0 to 1). The curve is significantly above the diagonal line, indicating good discrimination power.

After the statistical evaluation, the ecological meaningfulness of all parameters of the final model needs to be discussed and the spatial extrapolation of the model can be performed. The according module page giving the instructions for extrapolation, is directly linked to the webGIS-application. The webGIS page contains input-fields for model parameters, i.e. regression coefficients related to explanatory variables. Based on the values of the referred variables the model will be calculated for all polygons in the study area. The results will be written to a new column in the geo-database. The calculated probabilities of occurrence are presented as a habitat suitability map (Fig. 6). This application is again coded in PHP with embedded SQL. All calculations are performed in the database.



**Fig. 6:** Map of probabilities of occurrence of *Salvia pratensis* in the virtual landscape estimated with the model specified in Tab. 1 (webGIS application).

In a further step the map shall be classified to predict presence (1) or absence (0) of the species. Based on the output of the ROC-AUC program (Tab. 1) the user has needs to select an appropriate cut-off value to distinguish between the two states presence and absence. This value has to be typed to an input field on the webGIS page to classify the map.

The validation of the results is carried out by sampling the study area for a second time. A minimum distance between sampling points of the first and the second data set has to be respected. This data of second sampling provide both the observed species incidences and the calculated probabilities of occurrence. This data set is analysed in the ROC-AUC program. The obtained AUC-value is a measure for the transferability of the model to the whole study area (Schröder 2000).



### 3 Advantages and limitations

The GIMOLUS-learning unit 'habitat modelling with logistic regression' gives a detailed overview about the sequence of different steps necessary to estimate habitat models and provides an introduction to the methods involved. Beginning with the description of meaningful applications to the spatial extrapolation of the modelling results and its validation, all essential working steps are prepared in a multimedia environment. The modules can assist in teaching this procedure to students of environmental studies, but they are also designed for self-study purposes. The low requirements of software and technical equipment of the client computer are essential for a broad acceptance of the internet-based learning unit. The user only needs access to the internet, a browser and a GIMOLUS-account. Neither the GIS-software embedded in the system nor the statistical programs require temporal (training) or financial investments (licence fees). This is essential, since the GIS-application is fundamental to visualise the model's spatial extrapolation.

We do not know any other online or offline course providing a comparable comprehensive ability to teach / learn habitat modelling considering not only statistical issues but also a real-world application with an interactive GIS-background. We take the advantages of e-learning systems as the possibility of asynchronous and spatially independent learning. The design of the practical exercises in the learning module 'virtual problem on habitat modelling' shows a procedural character, that is well suited to promote the increase of the users' skills in habitat modelling.

The simplicity of operation is related to limitations of functionality on the other hand. The only procedure for building multivariate models provided in the learning unit is the stepwise variable selection. The user has to walk through this procedure step by step. As a part of a learning environment, this is not a disadvantage, because the user is forced to assess each step. Finally the spatial reference of the geo-database is limited to the virtual landscape. For teaching purposes this is also advantageous, as the virtual landscape is structured in a simpler and clearer way than real landscapes (Wesner et al. 2002). The didactical value regarding the presentation of the content in e-learning modules is based on the important part of interactive exercises and the formulation of practical problems in the modules (cf. virtual problem). The user is asked to improve and integrate his knowledge working on larger exercises. The understanding of single issues is promoted by interactive graphs and by the proceeding of exercises. The integration of the knowledge of the whole procedure and the interrelation of the working steps will be taught by means of the interactive virtual problem at the end of the learning unit. The didactical orientation of the module is corresponding to the demands of media didactics to use the potential of electronic media in a skilful way following the sequence model for e-learning units by Gagné (Kerres 1999) that postulates the transfer of learnt issues on new situations as an essential part. The learning process should notably be based on the independent activity of the learning person that will be supported by the learning environment attending their learning interest (Kerres et al. 2002).

### Bibliography

- Akçakaya, H.R., McCarthy & M.A., Pearce, J.L. (1995): Linking landscape data with population viability analysis: management options for the helmeted honeyeater *Lichenostomus melanops cassidix*. *Biological Conservation* 73: 169-176.

- Dullinger, S., Dirnböck, T., Gottfried, M., Ginzler, C. & Grabherr G. (2001): Kombination von statistischer Habitatanalyse und Luftbildauswertung zur Kartierung alpiner Rasengesellschaften. Proceedings AGIT2001, Salzburg, pp. 114-123.
- Hosmer, D.W. & Lemeshow, S. (2000): Applied logistic regression. 2nd edn. Wiley, New York.
- Kerres, M. (1999): Didaktische Konzeption multimedialer und telemedialer Lernumgebungen. HMD – Praxis der Wirtschaftsinformatik 205.
- Kerres, M., de Witt, C. & Stratmann, J. (2002): E-Learning. Didaktische Konzepte für erfolgreiches Lernen. In: Schwuchow, K. & Guttman, J., eds. : Jahrbuch Personalentwicklung & Weiterbildung, Luchterhand Verlag.
- Manel, S., Dias, J.-M. & Ormerod, S.J. (1999): Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* 120: 337-348.
- Morrison, M.L., Marcot, B.G. & Mannan, R.W. (1998): Wildlife-habitat relationships - concepts and applications. 2nd edn. The University of Wisconsin Press, Madison.
- Pearce, J. & Ferrier, S. (2000): Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133: 225–245.
- Schröder, B. (2000): Zwischen Naturschutz und Theoretischer Ökologie: Modelle zur Habitateignung und räumlichen Populationsdynamik für Heuschrecken im Niedermoor. PhD - Thesis, TU Braunschweig, Braunschweig.
- Söndgerath, D. & Schröder, B. (2002): Population dynamics and habitat connectivity affecting spatial spread of populations - a simulation study. *Landscape Ecology* 17: 57-70.
- Scott, J.M., Heglund, P.J., Morrison, M., Haufler, J.B., & Wall, W.A., eds. (2002): Predicting species occurrences: issues of accuracy and scale, pp 868. Island Press.
- Trexler, J.C. & Travis, J. (1993): Nontraditional regression analyses. *Ecology* 74: 1629-1637.
- Vennemann, K. & Müller, M. (2002): Konzeption einer Internetplattform für GIS- und Modellbasierte Lernmodule. In: Strobl, J.; Blaschke, T. & Griesebner, G. (Hrsg.): Angewandte Geographische Informationsverarbeitung XIV. Beiträge zum AGIT-Symposium Salzburg 2002. Heidelberg. pp. 567-572
- Wesner, S., Wulf, K. & Müller, M. (2002): How GRID could improve E-Learning in the environmental science domain. Proceedings of the 1st LeGE-WG Workshop 15.09.2002, Lausanne.
- Wissenschaftsrat (1998): Empfehlungen zur Hochschulentwicklung durch Multimedia in Studium und Lehre. Drs. 3536/98. <http://www.wissenschaftsrat.de/drucksachen/drs3536-98/drs3536-98.htm>

## Other sources

- Orr, J.L. (1996): Formula Graphing Applets – Zoomgrapher.  
<http://www.math.unl.edu/~jorr/java/html/ZoomGrapher.html>
- Pezzullo, J. (2001): Logistic regression. <http://members.aol.com/johnp71/logistic.html>
- Rudner, M. (2003): Logistic regression page. Landscape ecology group, University of Oldenburg, gimolus-project, unpublished.
- Schröder, B. (2002): ROC-AUC-program. Landscape ecology group, University of Oldenburg, <http://www.uni-oldenburg.de/landeco/Download/Software/Roc/Roc.htm>.