

**AUTOMATED CLASSIFICATION OF HABITATS**  
**Rasmus Ejrnæs<sup>1</sup>**

**-- Parallel Session --**

**Group 2-B. Analytical Tools for Measuring Trends in Biodiversity**

**Monday 5 November 2001**

**Paper presented to the:**

**OECD Expert Meeting on Agri-Biodiversity Indicators**  
**5-8 November 2001**  
**Zürich, Switzerland**

---

<sup>1</sup> National Environmental Research Institute, Dept. of Landscape Ecology, Denmark.

# Automated classification of habitats

*Rasmus Ejrnæs*

*National Environmental Research Institute, Dept. of Landscape Ecology  
Denmark*

## **Abstract**

Assessing the type or quality of a habitat can be done in many ways. Lately, subjective judgements by specialists, tend to be replaced by standardised judgements, typically involving a search for rare or typical species, or species and structures indicative of targeted features. A new method for optimising the extraction of conservation-relevant information from lists of vascular plant species is presented. This method involves three steps:

- 1) Collection of a reference data set spanning the observed variation in conservation interest for certain habitat categories.
- 2) Reduction of dimensionality by multivariate data analysis
- 3) Creation of a classification model that relates floristical composition to conservation interest

A model is presented, that classifies uncultivated, unshaded habitats of the Danish agricultural landscape, and it is shown that this model successfully classifies test data with respect to their value as habitats for rare and semi-rare species, as well as their overall contribution to species diversity of the landscape. The opportunities and perspectives for a wider use of automated classification models of this kind are discussed.

## **Introduction**

There is a growing understanding that habitats are the backbone of diversity. This is reflected by the European Community Habitats Directive (Anon 1992), obliging member countries to map, protect, monitor and, if necessary, restore a number of specified habitat types of recognised common importance for diversity in Europe. Also the Danish legislation for nature protection (Anon 1992b) specifies a number of characteristic semi-natural habitat types, where active land use changes are prohibited.

This emphasis on habitats raises the question how to identify and distinguish between habitats of different kinds. The first challenge is related to standardised ways of mapping habitat types. For this, a common understanding of the definitions and descriptions of habitat types is needed. Such definitions, and even identification keys, are provided for the habitat types of the Habitats Directive (Anon 1991). In practice however some of the types are identified on behalf of their species composition and often biotopes encountered during field mapping of areas do not fit perfectly into one or another class. The second challenge is related to the question of habitat quality. Some potential habitats may in fact be too intensively impacted by agricultural improvements to be included in the group of protected habitats. There are no standards for making this kind of decision, and up till now it is typically accomplished by subjective judgement involving the specific experiences of field biologists. This uncertainty is not only a problem for initial mapping of protected habitat types but also for the interpretation of repeated biological monitoring. Is the condition of an area favourable? Is it perhaps improving or vanishing?

The objective of this paper is to present methods for standardised classification of habitats with regard to recognised types or quality.

### **The concept**

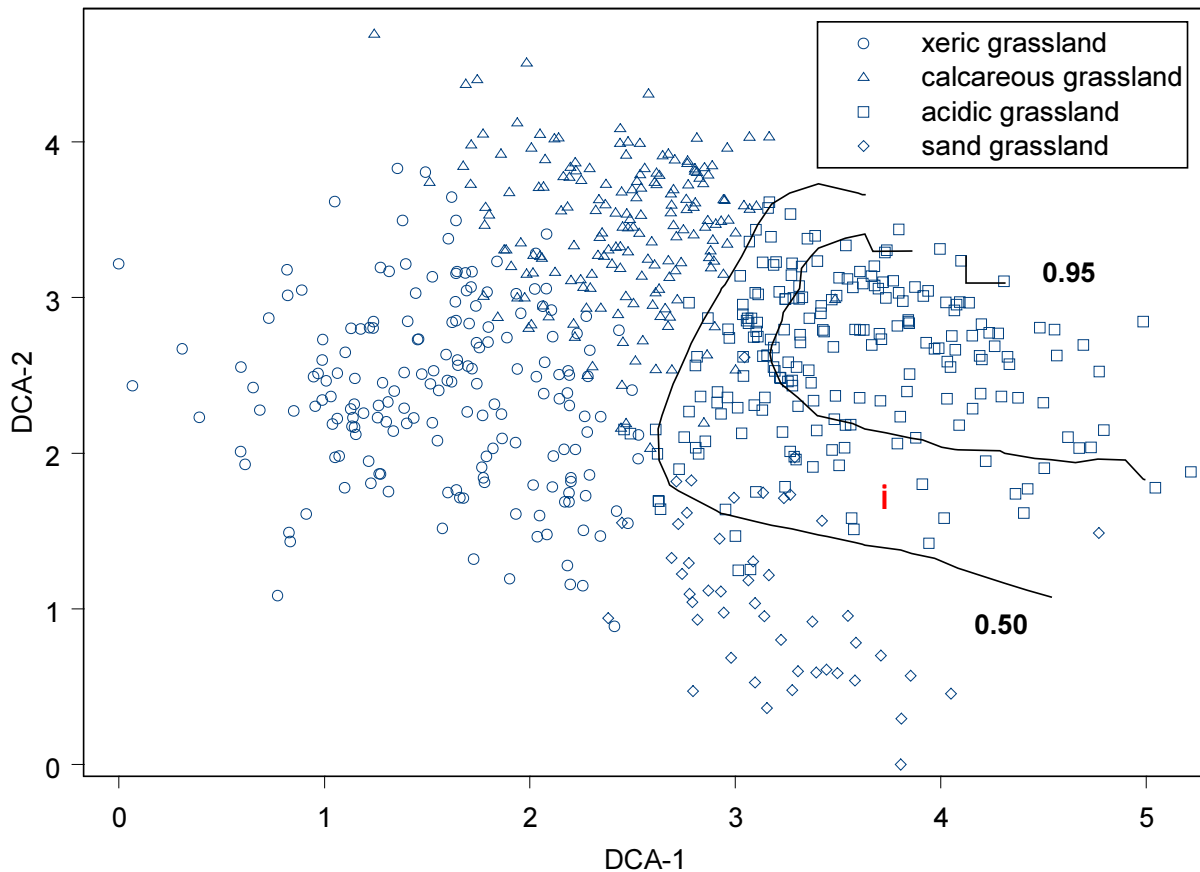
The procedure for standardised classification can be summarised to three steps:

- 1) Collection of a large reference or learning data set to teach the classifier. This data set should cover the full range of habitats that is the target of the classification model. This means that if a classifier is intended for prediction of habitat types for grassland habitat types within a certain geographical area, say Denmark, the reference data should cover the variation in grassland habitats within the area. The information collected from habitats should have high indicator potential for the targetted feature, whether this is type of vegetation or quality of habitat with respect to impacts of agricultural pressures. A prerequisite for the consecutive steps is an a priori classification or ranking of plots in this learning data set with regard to the target classification. If no a priori classification exists, clustering may be used to reach a reasonable classification of data.
- 2) Projection of the high-dimensional species\*plot matrix into a low-dimensional space by ordination. This space may eventually be interpreted in ecological terms by correlation of axes with environmental data (Ejrnæs & Bruun 2000).
- 3) Production of a classification model that successfully classifies plots according to the a priori classification. For a model to be considered reliable, its robustness and generality should be validated.

### **Prediction of habitat types**

My first example is from Danish grasslands on well-drained soils. A classification of Danish grasslands in four main types comprising 12 plant communities was published in 2000 (Bruun & Ejrnæs 2000). This classification is based on a TWINSPAN (Hill 1979) clustering of 620 vegetation samples from semi-natural and natural grasslands across Denmark. Figure 1 projects the four main grassland types in two dimensions achieved by ordination (Detrended Correspondence Analysis, DCA, Hill 1979b). A classification model was produced to predict grassland type from ordination axes, and this principle is shown in figure 1 as contour lines indicating 50% and 95% probability for membership of a plot belonging to acidic grasslands.

**Figure 1.** A projection of plots belonging to 4 main grassland types in Denmark in two dimensions deriving from a DCA-ordination of plots. Class membership is shown with different symbols, and the probability of membership to acidic grasslands is indicated by contour lines. A red star represents a new plot passively positioned according to its species list.



An attractive feature of DCA-ordination is the opportunity for passive ordination and classification of new plots according to the learning data set. This is illustrated on figure 1, where the red star represents a new plot, passively positioned according to its species list, and it is clear that this plot, given that it represents a semi-natural or natural grassland, is to be considered an acidic grassland (with some affinity to sand grassland). Such species based classification models for Danish plant communities are presently being implemented in an interactive database made accessible for the public through the internet (Nygaard et al. unpublished).

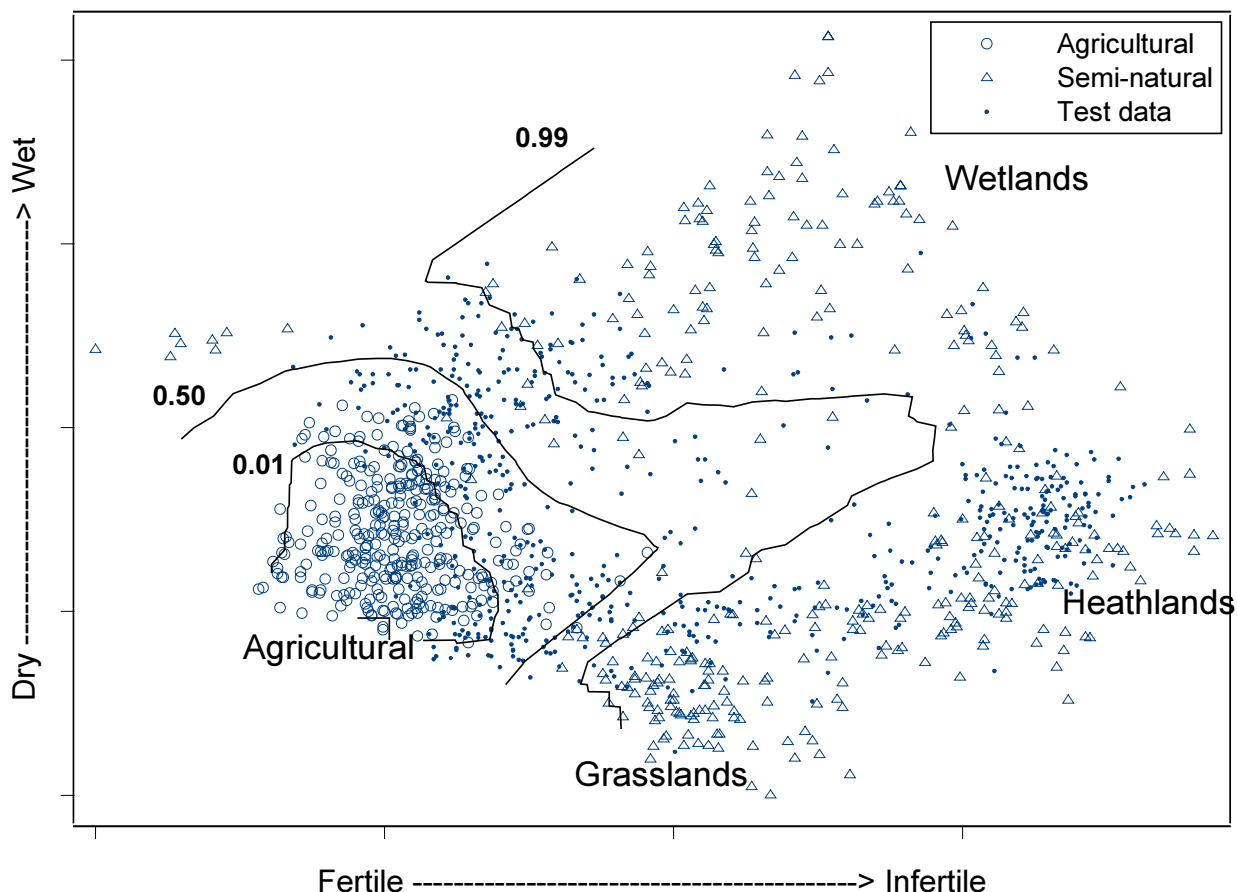
The presented classification deals with grassland types identified by unsupervised clustering. There is however no barrier for producing a similar model for explicit prediction of EEC Habitats Directive types. The prerequisite for this is a supervised classification of existing data in order to reproduce the EEC-types in the specific data set from Danish vegetation.

### Prediction of habitat quality

It is quite obvious that this concept has potentials for other applications. In 1998, Ejrnæs et al. (in press) collected a data set of more than 900 vegetation analyses representing all open (non-forest) uncultivated habitats of the Danish agricultural landscape. Data was collected in 10 randomly placed transects of 1 x 5 km selected to cross water streams and stratified from NW-Funen to W-Jutland. Of these plots, 334 were found to differ significantly from semi-natural and natural communities and were therefore classified as

“agricultural habitats”. The residual plots were kept as test data for the classification model. A comparable sample (333 plots) of semi-natural and natural habitats was selected from a vegetation database to represent the variation found in semi-natural grassland, heathland and wetlands.

**Figure 2.** Data and classification model for prediction of habitat quality. Source: Ejrnæs et al. In press.



An ordination of all data revealed that there were clear floristical differences between semi-natural plots and agricultural plots (Figure 2). Also, environmental interpretation revealed that the first axis of the ordination represented floristical turnover caused by a gradient in soil fertility, whereas the second axis represented a gradient in wetness. In this gradient plane agricultural habitats clustered together, while semi-natural habitats dispersed according to type of habitat, i.e. grasslands, heathlands, oligotrophic wetlands and mesotrophic wetlands. With this in mind, a classification model was produced by neural network to predict the probability of a sample of belonging to Danish semi-natural vegetation (Ejrnæs et al. in press). The model is illustrated (Figure 2) by contour lines indicating the probability of a sample of belonging to semi-natural habitats.

### Validation

Two strategies were adopted for validation of the model. The first validation test implied a classification of the unclassified test data of the 1998 inventory (616 plots) by the neural network classifier. 410 plots were classified as semi-natural and 206 plots were classified as agricultural. These plots were compared with respect to their richness of native species,  $\beta$ -diversity, percentage samples with semi-rare species and mean percentage of native species. The comparison revealed that although no difference could be detected with regard to richness of native species per plot, semi-natural plots differed significantly from agricultural plots

having a higher percentage of native species, higher  $\beta$ -diversity and more plots contained at least one semi-rare species.

The second validation test implied a classification of new data that were not used for the ordination of the learning and test data. This new data set was selected from existing Danish data to represent plots with one or more red-listed species, that in the current Danish Red List (Stolze & Pihl 1998) are classified as either “presumed extinct” or “endangered”. We were able to find 78 plots with such species. Luckily, all 78 samples were predicted to have more than 99 % probability of belonging to semi-natural habitats. This result confirms that if valuable habitats are protected, a simultaneous protection of species of conservation interest will be achieved. It is worth noticing that the rare species of the samples did not have any influence on the actual prediction (because they were not present in the reference data used for passive ordination). This means that even if red-listed species are not found during an inventory, the quality of a habitat may be reliably predicted.

### **Future improvements**

It is important to understand that any model of this kind relies heavily on the availability of adequate reference data, and its area of application will be restricted by the variation described by reference data. In the case of the presented classifier for habitat quality, the classification basically uses vascular plants to indicate the combination of wetness and fertility of the habitat, and because this has a predictable and strong relationship to the degree of agricultural impact, the model can be used for predicting habitat quality.

It is however possible to imagine situation where the model would predict a plot, that we would not consider as semi-natural, as semi-natural. This would be the case for recently disturbed, but naturally infertile areas (e.g. abandoned fields on poor sandy soils or abandoned gravel excavation areas). An acceptable assessment for such areas would require a new data set to be collected representing successional gradients from infertile, recently disturbed areas to semi-natural habitats.

The presented models both used species lists of vascular plants for prediction. This make sense for open areas, where vascular plant vegetation is abundant and respond strongly to the gradients in question. For forests, we might realise that other species groups are needed to produce reliable predictions (e.g. wood-inhabiting beetles or fungi, epiphytic lichens or bryophytes etc.).

### **Applicability of methods with respect to habitat-indicators for agriculture**

In relation to the development of indicators for agriculture, the presented methods may help in the calculation of proposed OECD indicator defined as:

*The share of agricultural area covered by semi-natural agricultural habitats*

It is acknowledged (OECD 2001) that the actual definition of semi-natural habitats is complicated, and it is suggested that a broad definition is used, and that data are aggregated from national agricultural censuses or national land inventories. Currently no inventories are carried out in Denmark that may be used to calculate this area, the closest being an existing mapping of protected uncultivated areas dating back to the early nineties. This mapping was carried out by the councils under time pressure, and can not be considered very reliable (see also Ejrnæs In press). In the Danish landscape, unlike countries with larger areas inaccessible for cultivation, semi-natural areas are extremely fragmented and distributed all over the landscape, and a reliable mapping can only be carried out by field work. The presented methods may help making decisions in cases where it is not clear whether an area can be classified as semi-natural.

Attention should also be made to the power of the presented method in detecting changes in the quality of habitats over time, and to interpret these changes in environmental terms. It should be acknowledged that in Denmark active habitat destruction of semi-natural habitats is prohibited, and therefore the major threat to semi-natural habitats come from deposition of ammonia released from neighbouring farms and fields, and succession due to ceased grazing and cutting of hay.

A quantitative indicator, such as the one proposed by OECD, may be useful for indicating the present situation in Denmark, and for making comparison to other countries, but will be of limited use for indication of future trends in Danish semi-natural habitats.

## References

- Anon. 1991. CORINE Biotopes manual, Habitats of the European Community. EUR 12587/3, Office for Official Publications of the European Communities, 1991
- Anon. 1992. Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora, O.J. L206, 22.07.92
- Anon. 1992 b. Bekendtgørelse af lov om naturbeskyttelse. LBK nr 835 af 01/11/1997.
- Bruun, H.H. & Ejrnæs, R. 2000. Classification of Dry Grassland Vegetation in Denmark. *Journal of Vegetation Science* 11: 585-596.
- Ejrnæs, In press. A Perspective on Indicators for Species Diversity in Denmark. Manuscript for OECD Expert Meeting 5.-8. November 2001.
- Ejrnæs, R. & Bruun, H.H. 1995. Prediction of Grassland Quality for Environmental Management. *Journal of Environmental Management* 43: 171-183.
- Ejrnæs, R., Aude, E., Nygaard, B., Munier, B. In press. Prediction of habitat quality using ordination and neural networks. *Ecological Applications*.
- OECD 2001. Environmental indicators for agriculture vol 3. Methods and results. OECD Publication Service, Paris.
- Stoltze, M. & Pihl, S. 1998. Rødliste 1997 over planter og dyr i Danmark. Miljø- og Energiministeriet, Danmarks Miljøundersøgelser, Skov- og Naturstyrelsen. [current Danish Red List].