Ottermanns R. (2008): Mustererkennung, Dimensionsreduktion und statistische Modellierung in der Ökologie - dargestellt am Beispiel der Lebensgemeinschaften grasiger Feldraine in deutschen Agrarlandschaften. Dissertation RWTH Aachen. Studien zur Mustererkennung, Band 27. Logos Verlag, Berlin. ISBN 978-3-8325-2032-8.

Abstract:

The aim of this study was to investigate analogies between ecological, statistical and pattern-based analysis of ecological communities, to use a synthesis of these different approaches to adopt the general pattern recognition process and, based on this, to define an ecological pattern recognition process. To achieve this, different data-based, empirical models were developed utilizing explorative data analysis, pattern recognition methods and synecological assessment approaches.

In detail the study focused on bringing together the different approaches of ecological and statistical analysis in an integrative analysis concept. It adopted up-to-date approaches from pattern recognition (statistical and stochastic models) and informatics (machine learning and artificial intelligence) to display data structures and processes in complex community data sets. Furthermore, these approaches targeted at pointing out the characteristics in ecological data with respect to the applied methods, illustrating the numerical requirements and comparing them to the implicit assumptions made in the mathematical methods, and initiating the development of suitable interfaces to integrate the manifold ecological background knowledge into the statistical pattern recognition process.

The analysed data set was composed from invertebrate and vegetation samples in grassy field margins in the German agricultural landscapes of the Jülicher Börde, the north-eastern border of the Leipziger lowland and the area Mainfranken near Würzburg. In each area plots with and without adjacent land utilization (farming activities) were investigated. A large number of environmental variables, plants, carabid beetles, spiders, springtails, hymenopterans, hover flies and lady birds were sampled in each plot and integrated into the analyses.

By the use of explorative data analysis important characteristics like correlation, distribution, outliers and species accumulation curves, were visualised for the primary variables (sites and species) and the environmental variables. The results confirmed the great importance of explorative data analysis for basic orientation within the data set as well as for data pre-processing prior to more complex subsequent analyses. A predetermined biocoenological classification reflected the available ecological background knowledge. In contrast, the null classification represented a null hypothesis with respect to the site classification. Divisive as well as hierarchical classification were suitable to reproduce these a priori class assignments, but both methods showed considerable shortcomings when applied to these ecological data. Despite this, both methods proved appropriate to give an overview of data set structure, for grouping of similar species distribution patterns, and to derive required advance information for classifier design. By

the extraction of class-specific indicators, the site clusters obtained from the classification were characterized in an ecological way. Additionally, the bioindicator potential of selected species were assessed with respect to the classification, and a feature selection in terms of pattern recognition was performed (reduction of dimensionality). Ecological knowledge was generated and used in following ordination analyses and supervised learning.

The results from classification analyses were compared using a newly developed index. The site clusters as well as the related indicator assemblages were compared between the different classification methods with respect to separation, unambiguousness and homogeneity. To display the variability of the site clusters and for feature extraction, distance-based ordination methods were used. In a subsequent gradient analysis, based on these results, the dependence of species density on environmental variables was analysed. Discrete system states were defined and the variability was visualised via suitable confidence intervals. To assign new sites to the classes a presence-based classifier was designed. Using a probabilistic model a Bayes error rate could be calculated, which allowed assessment of the classification quality, including ecological background knowledge. From a series of stochastic simulations recommendations concerning the required sample number and indicator quality were derived. The indirect quantification of a postulated pesticide effect on the communities could be performed by partial ordination analysis. For species which showed a particular sensitivity to this effect, shifts in dominance were depicted and a comparison of literature data for toxicological test species was conducted. The sensitivities of all observed species were combined in a newly developed type of species sensitivity distribution, which was used to set up a method for detection of indirect toxicological effects in the field. The derivation of a retrospective, statistical minimal model allowed the identification of key factors which determined the population density of sensitive species. Niche modelling analysis was then performed to verify how these species react to the key factors within a multivariate context. For the species which showed dominance formation caused by the pesticide effect, predictive models based on artificial neural networks allowed to make predictions about the population dynamics with respect to land utilisation and derive hypotheses about the most important species interactions within the observed communities. These models also demonstrated the suitability of the ecological pattern recognition process as a data pre-processing tool prior to the use of artificial neural networks.

In addition to the study's initial hypotheses, a large number of secondary hypotheses and issues concerning the data, as well as the theoretical aspects of the applied methods were derived in the course of the analysis and modelling process. The outlook carries promising and necessary refinements of the methods used in ecological pattern recognition. Hence the need for further research was documented. In an intense discussion the suitability of the applied methods of dimension reduction and statistical modelling was elucidated in more detail and the important term of information content was defined with respect to ecological data.

In conclusion, a reasonable ecological pattern recognition process, which combined the different approaches in statistics, ecology and machine learning in a profitable way, was expressed and placed in context with the concept of general ecological modelling.