Ecological Indicators 36 (2014) 242–253

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original article

Dimension reduction and data sharpening of high-dimensional vegetation data: An application to Swiss mire monitoring

Sucharita Ghosh[a,*], Ulrich Graf[b], Klaus Ecker[b], Otto Wildi[c], Helen Küchler[b], Elizabeth Feldmeyer-Christe[b], Meinrad Küchler[b]

[a] Statistics Lab, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[b] Ecosystem Dynamics Group, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[c] Biodiversity and Conservation Biology Unit, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

ARTICLE INFO

Article history:
Received 25 October 2012
Received in revised form 3 May 2013
Accepted 17 July 2013

Keywords:
Data mining
Kernel density estimation
Multidimensional scaling
Plant ecology
Species communities
Species diversity

ABSTRACT

In an era of the availability of very large databases, the problem of efficient methods to analyze such datasets remains. In large scale forest and landscape monitoring projects for instance, appropriate data mining techniques that can summarize the overall status of landscapes are necessary for planning and implementing follow-up management strategies. We consider a vegetation data set consisting of species data from more than 120 mires spread across Switzerland with a total of 20,134 plots on 2658 vascular and non-vascular plant species. Using species indicator values as proxy for site conditions, we propose some simple strategies for data mining involving multidimensional scaling and nonparametric probability density function estimation, both of which are known in the classical statistical literature. We show how commonly known techniques can be adapted in a novel and effective approach for the specific purpose of analyzing plant species occurrence in the plots to identify site conditions that influence species configurations and species diversity, thus providing important information concerning aspects of large scale vegetation structure. Our results indicate high variations among the mires with respect to site conditions that affect species assemblage and species diversity. While species indicator values continue to be popular as well as subject of much debate and research in the ecological community, our experience shows that careful methods of analysis at large landscape scales can reveal some powerful results, which can be taken up as the starting point for the next level of investigation.

1. Introduction

With the advancement of computer technology, efficient methods to analyze large datasets containing very high-dimensional observations has become a common problem. As new data come in, the urgency to analyze the existing data increases and the focus shifts towards visualization and analysis with the aim of obtaining an overview of the most important structures in the data. Our research is motivated by the problem of having an understanding of the important site characteristics that affect vegetation structure in the Swiss mires. For this we consider a vegetation data set that consists of vascular and non-vascular plant species data in more than 120 mires spread across the country. Each observation is a 2658-dimensional vector on plant species occurrence, stemming from one of 20,134 plots in these mires. Due to the lack of appropriate environmental measurements on site quality such as soil and other properties, we use species indicator values as proxy for site conditions. The use of species indicator values is well known in the literature (for a review see Diekmann, 2003), and related controversies are also documented (e.g. Zelený and Schaffers, 2012). Nevertheless, a large number of authors have continued to show interest in using species indicator values as an important tool for understanding site conditions; examples include Ecker et al. (2008), Ertsen et al. (1998), Hill and Carey (1997), Klaus et al. (2012), Payne et al. (2013), Seidling and Fischer (2008); and numerous others. However, since species indicator values are derived from the species data themselves, careful analysis is needed so as to avoid pitfalls (Zelený and Schaffers, 2012). In the current context however, each plot is assigned a species indicator value that is an average over all plots where the species occur. Details of this calculation is in Feldmeyer et al. (2007). Due to the very large number of plots involved in the entire database (ca. 20,000), the influence of any single plot becomes negligible. This validates further computations involving mean species indicator values within a given mire, as is illustrated in an analysis of the species data in Esleren, Gummenalp mire, which consists of ca. 150 plots. Moreover, the range of values of the indicator variables as well as of the other computed quantities such as resulting from multidimensional scaling,

* Corresponding author. +41 447392431.
E-mail address: rita.ghosh@wsl.ch (S. Ghosh).

1470-160X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2013.07.021

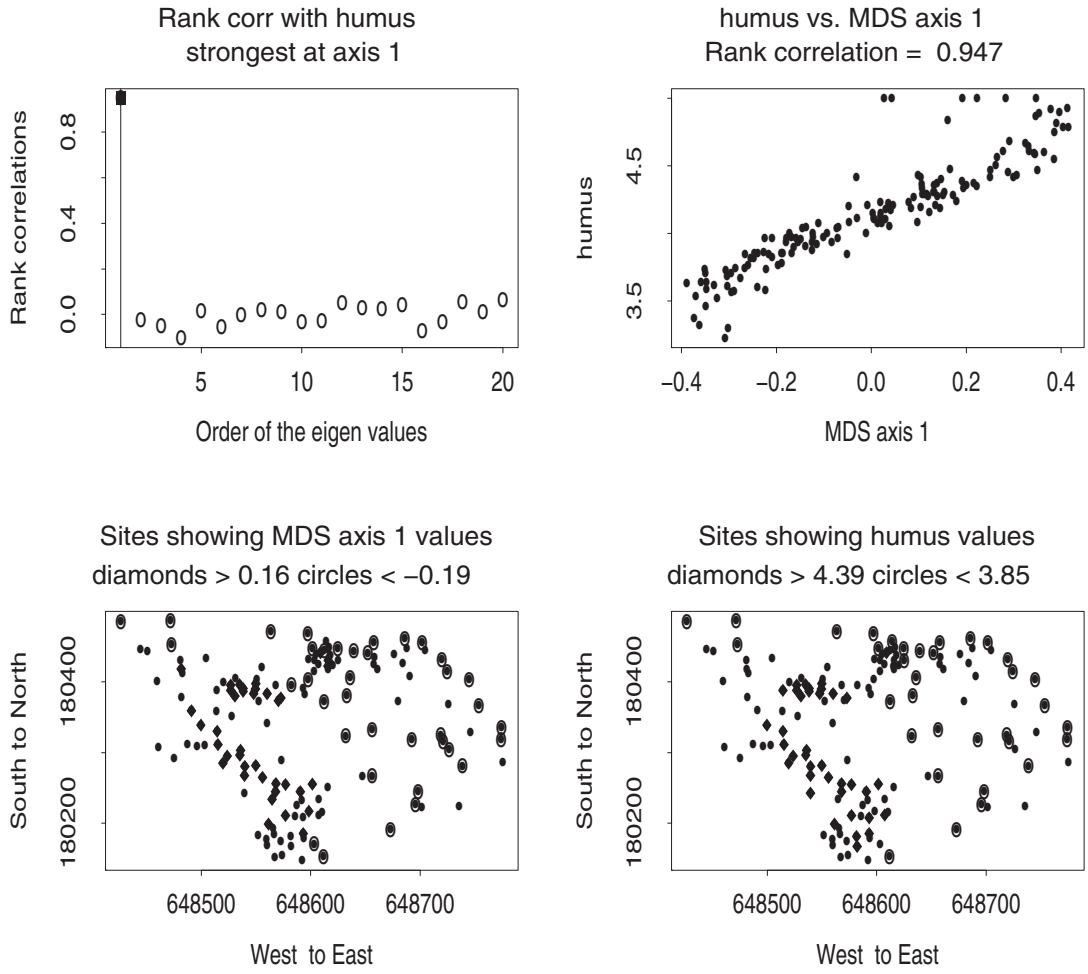

Fig. 3.1. Multidimensional scaling, Esleren, Gummenalp mire: principal (MDS) axis – 1 and Humuszahl (humus).

indicate that the presented correlations are not due to the presence of any dominant species across all plots in the mire. If this had been the case, then the computed quantities which take species presence in the plots into account, would not have had the appreciable variation that can be seen in the results of the analyses presented in the following sections. In contrast, a plot level mean indicator value based solely on the plant species present in the same plot may have resulted in spurious correlations. Such indicators were not used here. Generally speaking (as a referee has also pointed out) it is always necessary to consider a careful analysis of the species indicator values.

Two aspects of large scale vegetation structure are of interest here, namely species configuration and species diversity. These are some of the important aspects of landscape vegetation structure. Our results indicate high variations among the mires, with respect to site conditions affecting species assemblages and diversity. We illustrate our findings in detail for one mire and draw overall conclusions for all mires in Switzerland. Details of sampling methods are in Ecker et al. (2008) and Klaus (2007).

For analyzing species configuration, the high-dimensional vector valued observations are reconfigured using multidimensional scaling based on a binary distance metric, followed by identification of explanatory variables that have high correlations with the first two principal axes. The first step leads to dimension-reduction which reconfigures the high-dimensional presence–absence data using new coordinates and the second step provides an interpretation of first step results in the light of additional explanatory variables. For the reconfiguration step, we apply multidimensional scaling (MDS), also known as principal coordinates analysis, to the data vectors of binary-valued elements corresponding to presence or absence of each plant species. The output of the MDS is a new configuration of the data, where sites with similar species assemblages are placed nearer. The resulting map and additional explanatory variables are then used to find possible explanations for the heterogeneity in the species communities. Specifically, we consider the sub-space spanned by the first two principal axes, corresponding to the two largest eigenvalues and interpret these in the light of additional explanatory variables. For a technical description of the MDS, see Mardia et al. (1979), whereas application to phytosociology can be found in Kenkel and Orlóci (1986); also see Dufrêne and Legendre (1997) and references therein. It is to be noted that since the results of the MDS are indeterminate with respect to translation, rotation and reflection, interpretation of the MDS principal axes in terms of other known interpretable variables is important for practical reasons. For the Swiss mire data, we show that the principal axes are interpretable in terms of mean species indicator values reflecting site conditions. In other words, the two step procedure leads to identifying site characteristics that explain a large percentage of the total configuration.

The second aspect of vegetation structure that is of interest is the probability density function of a variable. For instance, we may consider a diversity index, and its relationship with site characteristics. Using a pdf, statistical inference can be based on relative frequencies of events of interest. For instance, one can identify sites that are either species rich or have low species richness. More importantly, the estimated pdf can be used for data sharpening, a method to