Journal of Experimental Marine Biology and Ecology 448 (2013) 72–84

Contents lists available at ScienceDirect

Journal of Experimental Marine Biology and Ecology

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

Causal modeling with multivariate species data

Warren L. Paul [a,*], Marti J. Anderson [b]

[a] Department of Environmental Management and Ecology, La Trobe University, Albury-Wodonga, Victoria 3689, Australia
[b] New Zealand Institute for Advanced Study (NZIAS), Massey University, Albany Campus, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 23 November 2012
Received in revised form 30 May 2013
Accepted 30 May 2013
Available online 12 July 2013

Keywords:
Amoco Cadiz oil spill
Causal modeling
Environmental impact studies
Linear and nonlinear ordination analysis
Multivariate species data
Structural equation modeling

ABSTRACT

Recent advances in causal modeling have made it possible to build and test structural equation models without any restriction on the functional forms or error distributions of the structural equations. We propose here a method for building and testing causal models that uses ordination axes arising from multivariate species data. This is demonstrated through the analysis of macrobenthic species abundance data observed at multiple times before and after the 1978 Amoco Cadiz oil spill (Dauvin, 1982). The available data consist of 21 quarterly observations on 257 species during the period 1977–1982. A causal model of the impact and subsequent recovery was built and tested using distance-based redundancy analysis (dbRDA). In addition, to predict the time required for recovery of the community, nonlinear models were fitted to the first two PCO axes, and the fitted nonlinear models were used to generate predictions for 20 years beyond the last observation in the data set. These predictions were found to compare favorably with the results from longer term studies carried out by Dauvin (1998). The methods described here are sufficiently well established to be used in ecological research, and will allow ecologists to move towards plausible causal models and generate stronger inferences from observational multivariate community data than has been achieved to date.

1. Introduction

Observational studies in ecology have a long and noble history (e.g., Emerson, 1836; Hutchinson, 1957; Von Humboldt, 1805; Wallace, 1855; Whittaker, 1960). Structured quantitative observational field studies of ecological systems are still considered essential for both the description of ecological patterns and the generation of ecological hypotheses concerning potential underlying processes that might give rise to those observed patterns (Underwood et al., 2000). However, causal inference regarding actual processes through the analysis of observational patterns alone is problematic (Popper, 1968; Underwood, 1990). Experimental studies that manipulate hypothesized processes and eliminate the effects of potentially confounding variables are necessary for stronger causal inference (Fisher, 1955; Hurlbert, 1984; Underwood, 1997). Nevertheless, well-designed observational studies remain a principal method of inquiry in ecology. In many situations, experiments to pinpoint causal processes are either infeasible or unethical, and ecologists are often more interested in understanding complex systems rather than individual processes. Ecologists are also increasingly interested to test theories at very broad temporal (evolutionary) and spatial (global) scales (e.g., Gotelli et al., 2009), for which experimental manipulations would be simply impossible.

Structural equation modeling (SEM) is often employed in ecology to address some of these issues by combining a qualitative description of the causal processes thought to be operating within a system, in the form of a path diagram (or causal diagram), with the statistical analysis of observational data (Grace, 2006; Pugesek, 2003; Shipley, 2000a). SEM has advanced considerably in recent years due largely to the work of two groups of researchers — one led by the computer scientist Judea Pearl at the Cognitive Systems Laboratory at UCLA, and the other by the philosophers Peter Spirtes, Clark Glymour and Richard Scheines at Carnegie Mellon University (see, for example, Pearl (1995, 2000) and Spirtes et al. (2000)). This framework, called causal modeling or structural causal modeling (Pearl, 2009), uses the developments in graphical models and the logic of intervention to clarify the causal content of SEM.

Causal diagrams are now established as a mathematical language, and as such their role in causal modeling extends far beyond being a convenient tool for communicating a (composite) causal hypothesis or deriving algebraic equations, as path diagrams have traditionally been used in SEM. Specifically, the structure of a causal diagram entails certain probabilistic constraints, in the form of conditional independencies, which can be read directly from the diagram using a graphical criterion called *d*-separation. These constraints underpin the causal content of SEM: they are the basis from which to determine the identifiability of causal effects (i.e., determine whether an effect of interest can be isolated from the effects of potentially confounding variables), they constitute the testable part of the causal model, and they help define the nontestable

* Corresponding author at: La Trobe University (Albury-Wodonga campus), P.O. Box 821, Wodonga 3689, Victoria, Australia. Tel.: +61 2 6024 9871; fax: +61 2 6024 9888.
E-mail address: w.paul@latrobe.edu.au (W.L. Paul).

0022-0981/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jembe.2013.05.028

part (i.e., the class of equivalent models) and thereby gauge the strength of causal inferences drawn from the model.

The connection between graphical models and statistical models has also led to a new approach for testing structural equation models, one which enables local tests of causal structures, as well as global tests in some cases, and does not require the usual assumption of multivariate normality (Pearl, 2000; Shipley, 2000a, 2000b). This approach allows model building and testing to be done with conventional statistical methods and general-purpose statistics packages (Shipley, 2000b), including – we suggest here – canonical correspondence analysis (CCA, ter Braak, 1986), redundancy analysis (RDA, Gittins, 1985; Rao, 1964), and distance-based redundancy analysis (dbRDA), (Legendre and Anderson, 1999; McArdle and Anderson, 2001) for the analysis of multivariate species data.

Causal modeling of multivariate species data, per se, is not new. Legendre and Troussellier (1988) developed a method for choosing among competing causal structures that describe the possible relationships among a set of species, a set of environmental variables, and a set of geospatial coordinates. They compared the values of Mantel and partial Mantel statistics, computed from resemblance matrices for the three sets of variables, with the values expected for a given causal structure. Leduc et al. (1992) used Mantel statistics in conjunction with a path analysis procedure to calculate path coefficients, i.e., standardized partial regression coefficients. Borcard and Legendre (1994) proposed using partial canonical analysis (i.e., constrained ordination analysis) as an alternative to partial Mantel tests in causal analyses. This approach uses RDA or CCA to partition the variation in a community into environmental and spatial components, which are then interpreted with respect to a set of competing causal models suggested by the authors. These methods are described in Legendre and Legendre (2012, Chapters 10 and 11).

The partitioning of multivariate variation according to a suite of environmental, spatial, temporal or other predictor variables, followed by a scientist's own interpretation, is certainly useful and appealing for analyzing ecological patterns in community structure. It does not, of itself, however, achieve the goal of allowing rigorous inferences regarding actual underlying causal processes (e.g., such as so-called "niche" or "neutral" dynamics) that might shape those communities (Anderson et al., 2011).

We propose that ordination axes can be modeled directly within the structural equation approach of a causal modeling framework, which allows ecologists to move towards plausible causal models and to generate stronger inferences from observational multivariate community data than has been achieved to date. This approach also yields working statistical models with a causal (mechanistic) basis that can be used for making predictions.

Here, we demonstrate the use of dbRDA and nonlinear models of PCO axes in causal modeling, with a special emphasis on environmental impact studies involving multivariate community data. An example is shown in an analysis of the response of an assemblage of macrobenthic species to the 1978 Amoco Cadiz oil spill, using data collected by Dauvin (1982). We begin by providing the background to the Amoco Cadiz data set in Section 2. In Section 3 we describe the causal diagrams that might apply in cases where data on the sediment oil concentration and water temperature may or may not be available. In Section 4 we explain the *d*-separation criterion and its application in determining that the effect of the oil spill is identifiable with the available data. In Section 5 we derive the structural equation model from the causal diagram, and in Section 6 we suggest functional forms for the structural equations, based on theory. Section 7 describes the exploratory analyses that were undertaken prior to model building and testing, which are described in Sections 8 and 9. Section 10 describes a complementary procedure to causal modeling with dbRDA that uses nonlinear models of PCO axes for the purpose of predicting the recovery time of the macrobenthic community following the oil spill. We conclude by summarizing the main elements of the causal modeling process and highlighting areas for future research. All of the analyses were done using the R computer package (R Development Core Team, 2010), and the entirety of the R code and data used are provided as supplementary online material.

2. Background to the Amoco Cadiz data set

The Amoco Cadiz oil tanker ran aground off the coast of France in 1978, spilling 1.6 million barrels of oil. The slick spread to the Bay of Morlaix within about a week of the event. For approximately one and a quarter years prior to the spill, Dauvin (1982) had been monitoring the macrobenthic community in the soft sediments of the Bay of Morlaix at approximately quarterly intervals, and monitoring continued for a number of years afterwards. The data used here are an example dataset included in version 6 of PRIMER (Clarke, 1993; Clarke and Gorley, 2006). These data are also provided as supplementary online material (see the Supplement for details). The data set consists of 21 quarterly samples taken from the sandy sediments of the Bay of Morlaix between April 1977 and February 1982, with counts recorded for each of 257 macrobenthic species. The Amoco Cadiz ran aground on March 16, 1978, between the fifth and sixth sampling occasions.

3. Constructing the causal diagram

Designing environmental impact studies in a way that deals with spatial and temporal confounding has been a subject of much consideration and debate in the literature (Ellis and Schneider, 1997; Green, 1979; Hurlbert, 1984; Smith et al., 1993; Stewart-Oaten and Bence, 2001; Stewart-Oaten et al., 1986; Underwood, 1991, 1992). Using the causal modeling framework, Paul (2011) indicated that spatial and temporal confounding in environmental impact studies may be controlled by conditioning on (adjusting for) the actual spatial or temporal positions of sample units. This was articulated in the context of a causal model for the response of a single species to an environmental impact. However, the idea is easily extended to a multivariate set of response variables (Pearl, 2000, d-separation, p. 16), such as multivariate species data.

Based on the work of Paul (2011), a possible causal diagram for the Amoco Cadiz study is shown in Fig. 1A. The variables are defined as follows: "temporal position" (Z_1) is the time of sampling in a rank-ordered sequence of integers (1, 2,..., 21), "temperature" (Z_2) is the water temperature, "time from exposure" (Z_1) is the time of sampling minus the first time of sampling after the spill ($-5, -4,...,15$), "spill" (X) is a binary variable contrasting the periods before vs. after the occurrence of the oil spill, "oil" (Z_4) is the sediment oil concentration, and "species abundance" (Y) is the multivariate species abundance data.

When Dauvin (1982) began monitoring the Bay of Morlaix he would not have anticipated this oil spill, so there are no oil data. Furthermore, there were no temperature data recorded with this data set. Accordingly, the first causal model to be built will correspond to the causal diagram shown in Fig. 1B, i.e., the causal diagram without either of the mediating variables of temperature or oil. (Note that a mediating variable, also known as an intervening variable, is a variable that lies on the causal path between two other variables.) However, to demonstrate model building with mediating variables the sea surface temperature data for a reasonably proximate site at Plymouth, UK, will be substituted for the temperature within the Bay of Morlaix. The monthly mean sea surface temperature data at Plymouth were obtained from the Centre for Environment, Fisheries & Aquaculture Science (CEFAS, 2011). The augmented causal diagram is shown in Fig. 1C. It must be stressed here that this is being done for demonstration purposes only; it is not appropriate in an actual model-building situation to simply substitute variables in a causal diagram with variables that were measured at some other site.

The causal diagram is a graphical description of the hypothesized data-generating process. Imagine, for example, that this study was done as an experiment in a laboratory where the experiment had three factors: spill (treated with oil or not), temperature (with various levels that