COENOSES 4 (2): 91-94, 1989
© 1989 by Edizioni LINT Trieste

A METHOD TO COPE WITH COLLINEARITY OF ECOLOGICAL DATA SETS IN COMMUNITY STUDIES

K.I. Stergiou, National Centre for Marine Research, Agios Kosmas, Hellenikon, 16604 Athens, Greece

Keywords: Factor analysis, Multiple regression, Collinearity, Anglerfish, *Lophius budegassa*, Greek waters, Eastern Mediterranean

Abstract. The use of factor scores in multiple regression models which suffer greatly from collinearity between the variables in the right hand of the equation is discussed. Data on the abundance of the anglerfish, *Lophius budegassa*, (dependent variable) and temperature, salinity, depth and prey availability (independent variables) in the Mediterranean Sea were used. Temperature, depth and prey abundance were significantly correlated with each other. Multiple regression techniques performed on the original, logarithmically transformed data revealed temperature as the only variable determining the abundance of *L. budegassa*. However, factor analysis (performed on the independent variables) and multiple regression techniques (performed on factor scores and the dependent variable) revealed that the abundance of *L. budegassa* is largely determined by temperature, depth and prey availability and, to a lesser extent, salinity. Hence, collinearity may bias greatly the fit of multiple regression models by masking and/or underestimating the importance of some other factors.

Introduction

In a recent paper Koslow et al. (1987) showed that regression analysis failed to a certain extent to define a group of environmental variables for the description of fishery recruitment for cod and haddock in the Northwest Atlantic during 1950-1980 due to the intercorrelations between the independent variables. Stimulated by this work, a way to cope with collinearity, between ecological data sets used in the development of multiple regression models in community studies, is presented. Such correlations between the right hand variables of the equation result from their sharing of common factors. They can substantially affect the fit of the model (Neter and Wasserman 1974) when a statistical identification of the factors affecting the abundance and distribution of the biota is required. When prediction is the object multicollinearity is not important (Fewster 1987). The present essay is based on data of temperature, salinity, depth and prey availability as related to the abundance of the anglerfish, *Lophius budegassa*, in the Eastern Mediterranean Sea (Patraikos, Korinthiakos Gulfs and the Ionian Sea, Greece) (Fig. 1). *L. budegassa*, is a demersal fish dwelling at depths ranging from shallow, coastal waters to 500 m. It is distributed in the Mediterranean and the Eastern North Atlantic from British Isles to Senegal (Caruso 1986).

Material and methods

Trawl samples were collected in June 1985 by means of a 425 HP motorboat equipped with a 14 mm mesh-size net (from knot-to-knot) at the cod-end. Sampling took place over a grid of 27 stations in the Patraikos, Korinthiakos Gulfs and the Ionian Sea (Fig. 1). Each haul lasted 45 min. At the end of each haul bottom temperatures and salinities were measured by means of a CTD. Measurements were taken at 17 out of the 27 sampling stations (Table 1) because of unfavorable weather conditions. The total number of individuals of all species caught was also determined. No replicate samples were taken at each station. Bottom type was the same at each sampling station (muddy sand). The abundance of *L. budegassa* and of its potential prey is expressed in number of individuals (per 45 min). The abundance of the potential prey of *L. budegassa* was

Table 1. Temperature in °C (T), salinity in o/oo (S), depth in m (D), abundance of prey (P) and abundance of *L. budegassa* (L) in number of individuals at sampling stations in the Patraikos, Korinthiakos Gulf and the Ionian Sea in June 1985.

STATION	T	S	D	P	L
1	14.5	39.5	76	567	10
4	15.2	39.6	47	71	0
7	14.0	39.3	115	1291	21
9	14.5	39.3	68	446	10
10	14.4	39.5	95	689	16
11	15.4	39.3	30	59	4
16	15.1	39.1	41	97	1
14	13.9	38.9	99	15	6
20	14.1	39.5	112	193	2
23	14.1	39.2	115	413	9
25	15.1	39.0	86	156	2
29	15.1	39.2	103	464	6
30	15.4	38.8	43	3	0
32	15.0	39.2	115	159	3
33	15.0	39.4	72	379	0
34	15.2	39.4	50	117	0
35	15.5	39.6	40	39	2

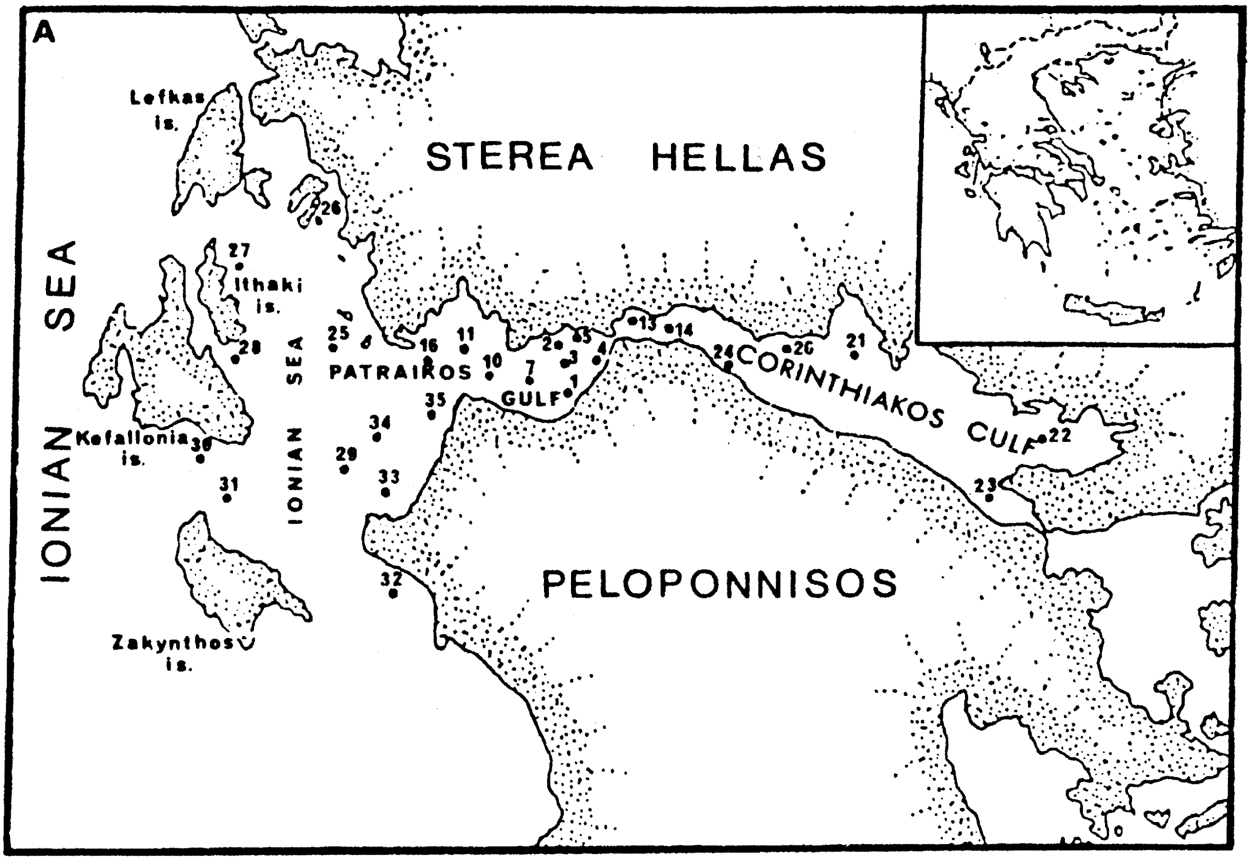

Fig. 1. Location of sampling stations.

estimated by summing up the number of individuals per station of the following species: *Merluccius merluccius*, *Argentina sphyraena*, *Trisopterus minutus capelanus*, *Trachurus trachurus*, *T. mediterraneus mediterraneus*, *Arnoglossus laterna*, *Cepola macrophthalma*, *Gobius sp.*, and *Callionymus maculatus*, which represent more than 90% by weight of its diet in Greek waters (Tsimenidis 1979).

All variables were transformed to natural logarithms in order to stabilize the variance and achieve normality (Zar 1984). Factor analysis (principal component, varimax rotation) was performed on temperature, salinity, depth and adundance of prey in order to cope with highly significant correlations observed between these independent variables. Such correlations result from their sharing of common factors and can substantially affect the fit of the model (Neter and Wasserman 1974) in the case of a statistical approach as opposed to a predictive objective (Fewster 1987). When intercorrelations are present standard errors of partial regression coefficients are large and, hence, partial regression coefficients are inaccurate estimates of the relationships in the population (Zar 1984). As a consequence, partial regression coefficients are not found to be statistically significant, even when the independent and dependent variables are related in the population (Zar 1984). In addition, intercorrelation may lead to increased roundoff error in the computation of regression statistics (Zar 1984). Two factors, explaining the greater proportion of the total variance, were retained. Subsequently, factor scores were computed based on the regression method (Norusis 1986b). Factor scores, which are not correlated with each other (Norusis 1986b), were used as independent variables for the development of a stepwise linear regression model. Backward elimination and forward selection models (Norusis 1986a) were also evaluated. All of these procedures identify subsets of variables that are good predictors of the dependent variable. Variable selection is based on certain entry and/or removal criteria (Norusis 1986a). Residual analysis was conducted in order to test the final model for violations of the major assumptions of the regression technique. The model was consequently compared with that resulted from multiple regression performed on the original, logarithmically transformed data. Statistical analysis was done using SPSS/PC+ Statistical package with the default criteria. Although there is no measure of how chance effects of the semiquantitative bottom trawling sampling procedure may influence variability, the data can be used to discuss and compare these two methods of

Table 2. Correlation matrix of the logarithmically transformed data: A = temperature, B = salinity, C = depth, D = prey and E = abundance of *L. budegassa*.

	A	B	C	D
B	.0210			
C	−.7452**	−.0776		
D	−.3887	.4416	.5614*	
E	−.6719*	.0749	.5518	.4866

Number of cases: 17 one-tailed significance: * — .01 ** — .001

statistical analysis.

Results and discussion

The results are presented in Tables 1 through 5. Temperature, depth and prey abundance were significantly correlated with each other (Table 2). Although salinity was not found to be significantly correlated with the remaining independent variables such correlations are to be expected in nature at least as far as salinity, depth and temperature are concerned. The results of the fit of a stepwise multiple regression model on the logarithmically transformed data are shown in Table 3. Forward selection and backward elimination models resulted in the same equation produced by the stepwise selection model. Temperature was the only variable that met entry and removal criteria, explaining 45% of the variance in the abundance of *L. budegassa*.

Factor analysis (principal component, varimax rotation) was performed on the independent variables (Table 4). Two factors were extracted, explaining 85.2% of the total variance of the variables (Table 4). The estimated factor scores were used as the independent variables in the development of a stepwise multiple regression model (Table 5). Factor 1 was the only variable that met entry and removal criteria. This factor is negatively related to temperature and salinity and positively to depth and prey abundance (Table 4). Backward elimination and forward selection models also resulted in the same equation. Residual analysis did not indicate any violations of the major assumptions of linear regression. In addition, the examination of the leverages for values greater than $\frac{2p}{N} = 0.12$ (where p = number of estimated parametres including constant, N = number of cases; Villeman and Welsch 1981) did not reveal the existence of an observation having an unusual influence on the results of the regression. However, the examination of the Mahalanobis' distance showed that cases 6 and 17 (stations 11 and 35) had unusual values for the independent variables which may have a substantial impact on the results of the regression. When cases 6 and 17 were omitted, the values of r^2 and adjusted r^2 rose to 0.57 and 0.53 respectively.

It is clear that multiple regression performed on the factor scores and the dependent variable revealed that

Table 3. Results of the stepwise multiple regression model performed on the logarithmically transformed variables. Dependent variable = abundance of *L. budegassa*.

Multiple R	.67187
R. Square	.45141
Adjusted R. Square	.41484
Standard Error	.81157

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	8.12957	8.12957
Residual	15	9.87968	.65865
F= 12.34286	Signif F= .0031		

Variables in the Equation

Variable	coefficient	SE	Beta	T	Sig T
Temperature	−19.41106	5.52511	−.67187	−3.513	.0031
Constant	53.50721	14.88374		3.595	.0027

Table 4. Results of factor analysis (principal component, varimax rotation) performed on temperature (A), salinity (B), depth (C) and abundance of prey (D).

Var	Communality	Factor	Eigenvalue	% Var	Cum %
A	.79940	1	2.16818	54.2	54.2
B	.89895	2	1.24177	31.0	85.2
C	.89637				
D	.81523				

Rotated Factor Matrix

	FACTOR 1	FACTOR 2
A	−.89409	.00121
B	−.13117	.93901
C	.94573	.04427
D	.57517	.69599

Table 5. Results of the stepwise multiple regression model performed on factor scores (FAC1, FAC2, see Table 4) and the abundance of *L. budegassa* (dependent).

Multiple R	.65655
R. Square	.43105
Adjusted R. Square	.39312
Standard Error	.82649

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	1	7.76294	7.76294
Residual	15	10.24630	.68309
F= 11.36451	Signif F= .0042		

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
FAC1	.69655	.20662	.65655	3.371	.0042
Constant	1.22161	.20045		6.094	.0000

the abundance of *L. budegassa* is greatly controlled with both biotic and abiotic factors: temperature (negatively), depth and prey availability (positively) and, to a lesser extent, salinity (negatively) as it is widely held in the literature (e.g. MacArthur 1972; Cody and Diamond 1975; Thiery 1982; Strong et al. 1984). On the contrary, multiple regression techniques performed on the original, logarithmically transformed, data indicated that the abundance of *L. budegassa* is merely associated with bottom temperature. This points out the importance of using factor scores in multiple regression models which suffer greatly from collinearity between the independent variables. The latter bias greatly the fit of multiple regression models by masking and/or underestimating the importance of other factors. It must be pointed out that in multiple regression models the assumption of non-intercorrelated variables in the right hand of the equation is an arbitrary requirement which arises from the limitations in the use of the standard distributions for significance tests and hence alternative tests suitable for a given sample may also be found (Fewster 1987).

REFERENCES

CARUSO, J.H. 1986. Lophiidae. In: P.J.P. Whitehead, M.-L. Bauchot, J.-C. Hureau, J. Nielsen and E. Tortonese (eds.). Fishes of the North-eastern Atlantic and the Mediterranean. Vol. III: 1362-1363.

CODY, M.L. and J.M., DIAMOND (Eds.). 1975. Ecology and evolution of communities. Belknap Press, Cambridge, Massachusetts. 545 p.

FEWSTER, P.H. 1987. Regression modelling of perturbation in some vegetation types. COENOSES 2: 67-74.

KOSLOW, A.J., K.R. THOMPSON and W. SILVERT. 1987. Recruitment to Northwest Atlantic cod (Gadus morhua) and haddock (Melanogrammus aegelefinus) stocks: influence of stock size and climate. Can. J. Fish. Aquat. Sci. 44: 26-39.

MACARTHUR, R.H. 1972. Geographical ecology. Harper & Row, New York. 269 p.

NETER J. and W. WASSERMAN. 1974. Applied linear statistical models. R.D. Irwin, Homewood, III. 842 p.

NORUSIS, M.J. 1986a. SPSS/PC+ for the IBM PC/XT/AT. SPSS Inc., Chicago, IL.

NORUSIS, M.J. 1986b. Advanced statistics SPSS/PC+ for the IBM/PC/XT/AT. SPSS Inc., Chicago, IL.

STRONG, D.R.Jr., D. SIMBERLOFF, L.G. ABELE and A.B. THISTLE (Eds.). 1984. Ecological communities. Princeton University Press, Princeton, New Jersey. 613 p.

THIERY, R.G. 1982. Environmental instability and community diversity. Biol. Rev. 57: 691-710.

TSIMENIDIS, N. 1979. Contribution to the study of the population dynamics of the genus Lophius L., 1758 in the Saronicos Gulf and structure of the populations in the Pagasitikos and Thermaikos Gulfs and the Thracian Pelagos. I.O.K.A.E. Spec. Publ., 4: 188 p.

VELLEMAN, P.F. and R.E. WELSCH. 1981. Efficient computing of regression diagnostics. The Amer. Statist. 35: 234-242.

ZAR, J.H. 1984. Biostatistical analysis, 2nd edition. Prentice Hall, Englenwood Cliffs, New Jersey. 718 p.

Manuscript received: September 1988