Journal of Mathematics and Statistics 1 (4): 322-325, 2005
ISSN 1549-3644

VAR-models for Predicting Biodiversity

Tonio Di Battista, S. Antonio Gattone and Mariagrazia Granturco
Department of Quantitative Methods and Economic Theory, University G. d'Annunzio
Chieti-Viale Pindaro n. 42, 65127 PESCARA, Italy

Abstract: The aim of this work is to introduce a methodology for monitoring diversity of a biological population by using the tool of diversity profiles. In particular, we develop a theory for studying the diversity profile along the temporal axis. We propose forecasting techniques for diversity profile by using the VAR model on multivariate time series of abundance vector and ARMA models both on univariate time series of abundances and on the aggregated series of diversity profiles. A Monte Carlo simulation is performed in order to test the goodness of forecasts obtained applying the three different methods proposed.

Key words: Biological community, abundance vector, diversity profiles, VAR models

INTRODUCTION

The concept of diversity arises both in ecological and non-ecological subject areas. Diversity is related to the apportionment of some quantity into a number of categories. What the actual quantity is, depends by the problem on hand. When we measure diversity we should take into account different aspects such as the number of different species and the relative abundance of different species. Most diversity measures can be classified as being heavily dependent on rare species (species richness) or on the abundance of the commonest (dominance). Consider a population of *s* species for which N_k and p_k denotes the abundance and the relative abundance of species *k*, for *k=1,2,…,s*, respectively.

Pielou[1] gave two characteristics an index of diversity should possess:

* for given *s*, the index should be a maximum when the p_k are equal;
* if the p_k are equal, the index should be an increasing function of *s*.

Patil and Taillie[2] proposed a general class of diversity index allowing all diversity measures to be encompassed into a single diversity spectrum. They started by defining diversity as the "*average rarity of species within a community*". More formally, given a community $C = \{s; p_1, p_2, ..., p_s\}$ and defining $R(p_k)$ as a measure of rarity for a species *k*, then the average rarity of species in the community is given by $\Delta(C) = \sum_{k=1}^{s} p_k R(p_k)$. A general formulation of $R(p_k)$ is $R_k = (1 - p_k^{\beta}) / \beta$ so that we get the diversity profile for community C as

$$\Delta_{\beta}(C) = \sum_{k=1}^{s} p_k \left(\frac{1 - p_k^{\beta}}{\beta}\right) \text{ when } \beta \neq 0$$

$$\Delta_{\beta}(C) = \sum_{k=1}^{s} -p_k \log(p_k) \text{ when } \beta = 0 \quad (1)$$

The value β denotes the relative importance of richness and evenness. For β =-1 we get the richness index, for β=0 the Shannon diversity index and for β=1 the Simpson index. Both Simpson and Shannon indexes are affected by the number of species and the evenness of species abundance, but they are affected differently. Thus, diversity profile can be plotted to compare communities in space and/or time over a range of different evenness emphasis. In the following we will be interested in a range of values of β belonging to the set $B = \{\beta : -1 \leq \beta \leq 1\}$ in order to have a suitable picture of the structure of the community under study[2].

Monitoring biodiversity is a growing concern of environmental agencies. While species and habitat are disappearing, it is crucial to be able to evaluate, even roughly, the biodiversity loss and predicting it. A lot of works in the literature deal with abundance and biomass prediction[3]. Nevertheless, a few works attempt to predict biodiversity[4,5]. There are no specific mathematical tools for predicting biodiversity but techniques used for predicting abundance also could be used for predicting biodiversity by using a wide range of multivariate techniques. This approach has the shortcoming to assume a linear relationship between the level of biodiversity and species abundances or any other measurable biological variable. Indeed, the literature data have failed to detect simple and linear relationship between the studied variables and diversity. For a thorough and critical review of the matter[6,7].

In this work we propose to understand, predict and manage biodiversity by viewing it as a function of species interaction or any other additional environmental variables. In particular, instead of using one single index we focus our attention to the diversity profiles of Patil and Taillie[2] in order to better describe the diversity of a community. Finally, we examine the

Corresponding Author: Tonio Di Battista, Department of Quantitative Methods and Economic Theory, University G. d'Annunzio, Chieti-Viale Pindaro n. 42, 65127 PESCARA, Italy

behaviour of forecasts of the diversity profile along the temporal axes by using multivariate VAR model and univariate ARMA model.

DIVERSITY INDEX ESTIMATION

Let us suppose that the ecological population is made up of *N* units and is partitioned into *s* species. Moreover, let N_k be the abundance of the *k-th* species (*k=1,2,…,s*). Hence, $\mathbf{N} = (N_1, N_2, .., N_s)^T$ denotes the abundance vector, while $\mathbf{p} = (p_1, p_2, .., p_s)^T$ represents the relative abundance vector with $p_k = N_k \Big/ \sum_{k=1}^{s} N_k$. As seen in the previous section, the diversity may be expressed as a function, say $\Delta(\mathbf{p})$, of the relative abundance vector. We consider the problem of estimating **N** and **p** and accordingly $\Delta(\mathbf{p})$ on the basis of a sample of biological units under SRS with replacement of size *n*. Let

$$\mathbf{n} = (n_1, n_2, .., n_s)^T \qquad (2)$$

be the estimated abundance vector, and

$$\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, .., \hat{p}_s)^T \qquad (3)$$

the estimated relative abundance vector, where $\hat{p}_k = \frac{n_k}{n}$ (*k=1,2,…,s*). Therefore, under SRS, $\mathbf{n}$ is a realisation of the multinomial random vector $\mathbf{N} = (N_1, N_2, .., N_s)^T$ with parameters *n* and **p**. Accordingly, since $\hat{\mathbf{p}} = n^{-1}\mathbf{n}$, then $\hat{\mathbf{p}}$ turns out to be an unbiased and consistent estimator for $\mathbf{p}$. The variance-covariance matrix of $\hat{\mathbf{p}}$ is given by $Var(\hat{\mathbf{p}}) = n^{-1}\sum$ where $\sum = diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$. Moreover, a straightforward consistent estimator for $\sum$ is given by $\hat{\sum} = diag(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T$. Finally, the use of the Central Limit Theorem provides that

$$n^{1/2}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} N_s(\mathbf{0}, \sum), \quad \text{as } n \to \infty.$$

Being $\Delta_\beta(\mathbf{p})$ function of a random variable, we may consider it also as a random variable. Moreover, by using the Delta method Tong[8] proves that as $n \to \infty$, if

$$\Phi_{k,\beta}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}_k}\Delta_\beta(\mathbf{x}),\ k=1,2,\ldots,s,\ \beta \in B$$

are defined in a neighbourhood of **p** and non-null at **p**, then the variance-covariance matrix of $\{\Delta_\beta(\mathbf{p}) : \beta \in B\}$ is given by

$$\mathbf{T} = \Phi(\mathbf{p})\sum\Phi(\mathbf{p})^T.$$

Forecasts for diversity profiles: As seen in the previous section, vector of abundances $\mathbf{n} = (n_1, n_2, ..., n_s)^T$ can be viewed as a realisation of the multinomial random vector $\mathbf{N} = (N_1, N_2, .., N_s)^T$. Suppose that for each of *s* species of a given community *C* the time series of absolute abundances is available. It is plausible to think that information about abundance $\hat{\mathbf{n}}_{kt}$ observed at time *t*=1,2,…,T for the species *k*=1,2,…,s is contained in the past values observed for the abundance vector.

In other words, we suppose that time series of absolute abundance of the specie *k* is a realisation of the discrete stochastic processes $\{\mathbf{N}_k(t) \quad t = 1, 2, ..., T\}$ and the set of series observed in the whole community is a realisation of a vector stochastic process $\{\mathbf{N}(t) = [\mathbf{N}_1(t), ..., \mathbf{N}_s(t)]^T\}$. In particular, we assume that absolute abundances are realisation of a stable Vector Autoregressive Process of order *l* (VAR(*l*))[9], such that:

$$\mathbf{N}(t) = \mathbf{v} + \mathbf{A}_1\mathbf{N}(t-1) + ... + \mathbf{A}_l\mathbf{N}(t-l) + \mathbf{u}(t) \qquad (4)$$

where $\mathbf{N}(t) = [\mathbf{N}_1(t), \mathbf{N}_2(t), ..., \mathbf{N}_s(t)]^T$ is a *s*-dimensional random vector, $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_l]$ is a *(s×sl)* matrix of coefficients, V is a *(s×1)* vector of intercept terms allowing for the possibility of a non zero mean $\mathrm{E}[\mathbf{N}(t)]$ and $\mathbf{u}(t) = [\mathbf{u}_1(t), ...\mathbf{u}_s(t)]^T$ is a *s*-dimensional white noise process. The matrix of coefficients A can be estimated by ordinary least square.

In particular, we focus our attention on the abundance vector of (2), so that we say that $\mathbf{n}_t$ is the abundance vector observed at time *t*. Consequently, once $\mathbf{n}_t$ has been observed, we can straightforwardly derive the relative abundance vector observed at time *t*, say $\hat{\mathbf{p}}_t$ and the diversity profile observed at time *t*, say $\Delta_{\beta,t}(\hat{\mathbf{p}}_t)$.

Then a forecast for period *T+h*, may have the form $\mathbf{n}_{T+h} = f(\mathbf{n}_T, \mathbf{n}_{T-1}, ...)$ where $f(\cdot)$ denotes some suitable function of the past observations.

We consider forecasts which are linear function of past observations. Assuming that only a finite number l, say, of past n_k values are used in the prediction formula we get, for $k = (1, 2, ..., s)$

$$\hat{n}_{k,T+h} = v_k + \alpha_{k,1,1}n_{1,(T+h-1)} + \alpha_{k,2,1}n_{2,(T+h-1)} + ... + \alpha_{k,s,1}n_{s,(T+h-1)} + ...$$

$$+ \alpha_{k,1,l}n_{1,(T+h-l)} + \alpha_{k,2,l}n_{1,(T+h-l)} + ... + \alpha_{k,s,l}n_{1,(T+h-l)} \qquad (5)$$

To simplify the notation let $\mathbf{n}_t = (n_{1,t}, n_{2,t}, ..., n_{s,t})^T$, $\mathbf{v} = (v_1, v_2, .., v_s)$ and for $r = 1, 2, ..., l$

$$A_r = \begin{bmatrix} \alpha_{11,r} & \cdot & \cdot & \cdot & \alpha_{1s,r} \\ \cdot & & & & \cdot \\ \alpha_{s1,r} & \cdot & \cdot & \cdot & \alpha_{ss,r} \end{bmatrix}.$$

Then (2) can be written compactly as

$$\hat{\mathbf{n}}_{T+h} = \mathbf{v} + A_1\mathbf{n}_{T+h-1} + ... + A_l\mathbf{n}_{T+h-l} \quad (6)$$

which is the optimal forecast obtained from a vector autoregressive model of the form (4).

Naturally, we might focus our attention to diversity profiles. Under the assumption that absolute abundances are realisations of a VAR(*l*) process, the β-diversity profile is a non linear combination of abundance vector observed at time *t*, so it is also a random variable. In particular, for a fixed value $\beta \in B$ the diversity profile $\Delta_\beta(\mathbf{p})$, is a real function of the components $\mathbf{p} = (p_1, p_2, .., p_s)^T$ defined in $\Re^s$, then $\{\Delta_\beta(\mathbf{p}) : \beta \in B\}$ is a *m*-dimensional random variable, where *m* is the cardinality of the set B. The function $F : \Re^m \to [0,1]$,

$$F\left(\Delta_{\beta 1}, \Delta_{\beta 2}, ..., \Delta_{\beta m}\right) =$$

$$P\left(\Delta_{\beta 1} < \delta_{\beta 1}, \Delta_{\beta 2} < \delta_{\beta 2}, ..., \Delta_{\beta m} < \delta_{\beta q}\right)$$

is the joint distribution function of $\{\Delta_\beta(\mathbf{p}) : \beta \in B\}$. Obviously, for a any fixed value $\beta \in B$ $\Delta_{\beta,t}(\mathbf{p})$, *t=1,2,…,T*, is a stochastic temporal process. For simplicity we assume that time series of β-profile is a realisation of a linear ARMA(*p,d,q*) process[10].

For a fixed $\beta \in B$, using an ARMA(*p,d,q*) model , forecast of diversity for period *(T+h)* might be expressed as:

$$\hat{\Delta}_{\beta,T+h}(\mathbf{p}) = w + \phi_1\Delta_{\beta,T+h-1} + ...\phi_p\Delta_{\beta,(T+h-p)} - ... - \theta_1\varepsilon_{T+h-1} - ... - \theta_q\varepsilon_{(T+h-q)} \quad (7)$$

under the assumption that order of integration *d* is estimated outside the model for non stationary mean time series.

Suppose that the goal is to forecast the diversity profile (1) in order to analyse the dynamic structure of a given biological community. When temporal observations for absolute abundances are available, it is theoretically possible to obtain a forecast for the diversity profile in three different ways.

First of all, using the multivariate forecasting model in (5) and (6) it is possible to obtain forecasts for absolute abundances, which aggregated by using (1) lead to forecasts for β-diversity profile. Obviously, the process under study is unknown and in practice the coefficients of assumed VAR(*l*) process must be estimated from a given multiple time series. Criteria for determining the order *l* of the model and for checking the assumptions underlying a VAR analysis have to be followed[9].

Second, using an univariate ARIMA(*p,d,q*) model, forecasts for each time series of abundances can be estimated and aggregated to obtain forecasts for diversity profile (1). Finally, forecasts could be directly obtained for time series of β-diversity profiles by using an univariate ARIMA(*p,d,q*) model.

When univariate model is applied the order is obviously unknown and have to be estimated using one of the automatic selection criterion proposed in the literature[10]. In the next section, by means of a Monte Carlo simulation we evaluate the performance of the three different methods for forecasting diversity.

MONTE CARLO EXPERIMENT

The sample behaviour of forecasts for diversity profile is investigated by a Monte Carlo experiment. The process used in the simulation is a stable VAR process of order *l*=1, where the disturbance are *i.i.d.* $N(\mathbf{0}, \mathbf{I}_s)$ whit *s*=2,3,4,5,6,7,8. For every parameterisation 300 replications are considered. In order to test the performance of estimators, length time series is fixed at 105 simulated observations, since a stabilisation of mean square error of estimates is obtained when length is not less then 100 simulated observations.

Forecasts for different hypothesis described earlier are computed as follows. Using multivariate technique a VAR(*l*) model is estimated for the simulated data of absolute abundances by ordinary least square; order of model is specified using AIC criterion for multiple time series whit a maximum number of parameters equal to 4[9].

When univariate technique is applied, an ARIMA(*p,d,q*) model is estimated respectively for each univariate series of simulated absolute abundances and for the related time series of diversity profiles; order of the model is specified using AIC criterion for univariate time series whit a maximum number of parameters equal to 6 and estimation is performed by ordinary least square.

For each of the above methods, estimation is performed using the first T=100 observations; the estimated model is used to produce a sequence of 5 forecasted values for β-profile and Root MSE of forecasting is computed. The results for increasing number of species are shown in Fig. 1.

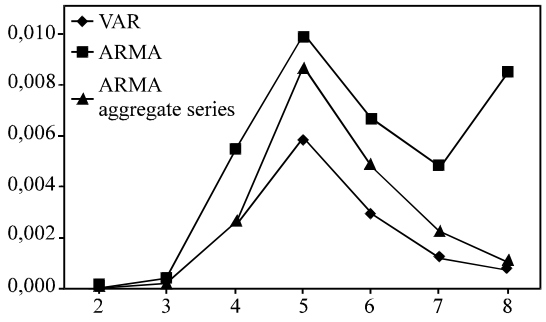

Fig. 1: Average root MSE of forecasting over the Monte Carlo experiment for increasing number of species

The Fig. 1 reveals that forecasts obtained using a VAR(*l*) model show always lowest average Root MSE, particularly when the number of simulated species is large.

Results highlight that VAR models are suitable for predicting diversity with respect to ARMA models as the performance of multivariate temporal models is higher when the variables are highly dependent on each other. In fact, it is known that species of biological community are highly correlated.

Finally, we point out that, recently different case studies have implemented monitoring systems for the construction of panel-data for abundance vector of biological populations. Since our methodology suitably should fit these new research fields, our next goal is to apply the methodology proposed to a real data set.

REFERENCES

1. Pielou, E.C., 1975. Ecological Diversity. London, Wiley.
2. Patil, G.P. and C. Taillie, 1982. Diversity as a concept and its measurement. J. Am. Stat. Assoc., 77: 548-567.
3. Verner, J., M.L. Morrison and C.J. Ralph, 1986. Wildlife2000: Modelling Habitat Relationships of Terrestrial Vertebrates. Wisconsin University. Madison WI.
4. Mc Arthur, R.H., H. Reicher and M.L. Cody, 1996. On the relation between habitat selection and bird species diversity. Am. Nat., 100: 319-332.
5. Lek-Ang, S., L. Deharveng and S. Lek, 1999. Predictive models of collembolan diversity and abundance in a riparian habitat. Ecological Modelling, 120: 247-260.
6. James, F.C. and C.E. Mc Culloch, 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Ann. Rev. Ecol. Syst., 21: 129-166.
7. Lek, S., M. Delacoste, P. Baran, I. Dimopaulos, J. Lange and S. Aulonier, 1996. Application of neural networks to modelling non-linear relationships in ecology. Ecol. Mod., 90: 39-52.
8. Tong, Y.L., 1983. Some distribution properties of the sampling species diversity indices and their applications. Biometrics, 39: 999-1008.
9. Lutkepohl, H., 1993. Introduction to Multiple Time Series Analysis. Springer-Verlag, 2nd Edn.
10. Priestley, M.B., 1981. Spectral Analysis and Time Series. Academinc Press.