The 2014 International Conference on Marine and Freshwater Environments

Joint Conferences:
The 2014 Annual Conference of the International Society for Environmental Information Sciences (ISEIS)
The 2014 Atlantic Symposium of the Canadian Association on Water Quality (CAWQ)
The 2014 Annual General Meeting and 30th Anniversary Celebration of the Canadian Society for Civil Engineering Newfoundland and Labrador Section (CSCE-NL)
The 2nd International Conference of Coastal Biotechnology (ICCB) of the Chinese Society of Marine Biotechnology and Chinese Academy of Sciences (CAS)

iMFE 2014

# Parametric functional analysis of variance for fish biodiversity

Tonio Di Battista[1], Francesca Fortuna[2] and Fabrizio Maturo[3]

*1 Department of Quantitative-Economic and Philosophical-Educational Science, University of Chieti-Pescara, Italy*

*email: dibattis@unich.it*

*2 Department of Quantitative-Economic and Philosophical-Educational Science, University of Chieti-Pescara, Italy*

*email :francesca.fortuna@unich.it*

*3 Department of Quantitative-Economic and Philosophical-Educational Science , University of Chieti-Pescara, Italy*

*email:fabmatu@gmail.com*

## Abstract

The conservation and restoration of biodiversity in the marine environment is a crucial aspect of fishing and related activities. Human activities cause changes in fish population and deep transformation in the type and quality of the water. Fishing, restocking and pollution often bring to reduction and distribution changes of indigenous fish species to the benefit of the diffusion of exotic species. In this context protection and management of water environments become a primary objective. Therefore it is necessary to implement initiatives for protecting and restoring the quality and integrity of native species. Any decision-making process must be based on a careful analysis of the collected data. In this paper we propose a parametric functional approach to study the biodiversity in marine environment.

**Keywords:** Parametric Functional data analysis, fANOVA models, Water quality, Fish biodiversity

## Introduction

Public water policy requires assessment of ecosystems species diversity, and monitoring to determine changes that can be used to predict population declines and loss of environmental resources (Burger et al., 2013). Animals, plants, micro-organisms and their complex interactions, in fact, react to human impacts in different ways, with some organisms responding more quickly and definitively than others (Paoletti, 1999). In an ecological framework, biodiversity relies on the variety of living organisms in a delineated study area (Heywood and Watson, 1995; Pavoine and Doledec, 2005). However, it is difficult to quantify this broad and complex concept; in fact, nowadays there is not yet a universally accepted biodiversity measure.

In this paper we aim to assess water quality through biodiversity by proposing the combined use of parametric biodiversity indices and the functional data analysis approach. This allows us to consider the multidimensional aspect of biodiversity and use statistical techniques, such as functional linear models, for studying the relationships between biodiversity and environmental characteristics.

## Materials and Methods

Biodiversity is a multidimensional concept accounting for both species richness (the number of different species represented in an ecological community) and species evenness (a measure of the relative abundance of each species in an area). Since it represents a good indicator of ecosystem quality, it should be analyzed and quantified to ensure its protection. In order to evaluate biodiversity we refer to parametric families of diversity indices (Hill, 1973; Patil and Taillie, 1982), which are usually referred to as diversity profiles. They consist of a sequence of measurements allowing different aspects of community structure to be encompassed in a single diversity spectrum. Diversity profiles present considerable advantage respect to the classical biodiversity indices. It is well known, in fact, that the use of a single index greatly reduces the complexity of the ecological systems (Gattone and Di Battista, 2009; Gove et al., 1994; Patil and Taillie, 1979).

In particular the β diversity profile (Patil and Taillie, 1979, 1982) has been applied:

$$\Delta_\beta = \sum_{j=1}^{s} \frac{(1-p_j{}^{\beta})}{\beta} p_j = \sum_{j=1}^{s} \frac{1-\sum_{j=1}^{s} p_j{}^{\beta+1}}{\beta} p_j \qquad \beta \geq -1 \qquad (1)$$

where $p_j$ is the relative abundance vector with $p_j = P_j / \sum_{j=1}^{s} P_j$ such that $0 \leq p_j \leq 1$ and $\sum_{j=1}^{s} p_j = 1$, and $P_j$ is the abundance of the j-th species (the number of individuals belonging to the species j). The value of $\beta$ denotes the relative importance of richness and evenness. The restriction that $\beta \geq -1$ assures that $\beta$ profile in equation (1) has certain desirable properties (Patil and Taillie, 1979, 1982). The plot of equation (1) versus $\beta$ provides the diversity profile which is a decreasing and convex curve. Some of the most frequently used indices of diversity are special cases of equation (1); in fact for $\beta = -1$ we get the richness index, for $\lim_{\beta \to 0}$ we have the Shannon diversity index (Shannon, 1948) and for $\beta = 1$ we obtain the Simpson index (Simpson, 1949).

Since diversity profile, regardless of how it is calculated, expresses diversity as a function of the relative abundance vector in a functional domain, it can be analyzed in a functional context (Gattone and Di Battista, 2009). Functional data analysis (FDA) addresses problems in which the observations are described by functions rather than finite dimensional vectors (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). It is important to emphasize that the functional datum should be regarded as a single entity instead of a sequence of observation. However, the empirical observation must necessary refer to the discretization of the domain; thus, in real applications, functional data are often observed as a sequence of point data. In this context, FDA approach is able to convert discrete observations to functional form by means of appropriate techniques such as the use of basis functions. Moreover, in a FDA framework, it is possible to use the functional tools to obtain more information on the data such as the analysis of the slopes of functions, reflected in their derivatives, and so on, by highlighting the characteristics of the curves.

In an ecological framework, since diversity profile is not simply a sequence of observations, but a function in a fixed domain, it is possible to analyze the intrinsic structure of the data through FDA approach. With reference to water management, this framework has been used to explore

differences in water quality trends between site (Henderson, 2006). Indeed, we focus on a particular aspect of functional data analysis, called parametric FDA (De Sanctis and Di Battista, 2012; Di Battista and Fortuna, 2013). In this case, the functional datum is expressed by a specific function known in advance. The observations, in fact, belong to a parametric family of functions, called S, with s real parameters, that is:

$$S = \{ f(\boldsymbol{\theta}, x) \} \tag{2}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots\ldots, \theta_s)'$ represents a set of unknown parameters taking values in a parameter space Θ while x is the functional domain. In this framework, functional data constitute a subset S of some $L^p$ space, with $0 < p < \infty$ and with the usual $L^p$-norm, $\|f\|_p$ (Rudin, 2006):

$$\|f\|_P = \left\{ \int_X |f|^P d\mu \right\}^{\frac{1}{P}} < \infty \tag{3}$$

where X is an arbitrary measure space with a positive measure μ. In particular, we consider every $L^p$ space with $p > 0$ (Banach spaces).

In an ecological setting, S could be the family of diversity profiles, such as $\beta$ profile in equation (1), and for each i-th sites, i=1, 2, …, N, every relative abundance vector can be assumed as a single parameter, $\mathbf{p_i} = (p_{i1}, \ldots\ldots, p_{is}) = \boldsymbol{\theta}_\mathbf{i}$, so that, $p = \theta$.

The advantage of parametric FDA approach is that the approximation by means of basis functions is not suitable because the underlying data process is known in advance and it is important to preserve its parametric form.

**Parametric fANOVA model**

In order to quantify the effects exerted on a functional observation by some factors, each at multiple levels, a parametric functional analysis of variance (fANOVA) model has been applied (Ramsay and Silverman, 2005).

We assume that there is a single factor with K different levels or groups $(k = 1, 2, \ldots, K)$ and $n_k$ observations within each group; so the model for the i-th observation $(i = 1, 2, \ldots, N)$ in the k-th group can be expressed as follows:

$$f_{ik}(x) = \mu(x) + \alpha_k(x) + \varepsilon_{ik} \tag{4}$$

where $f_{ik}(x)$ is a functional response in the k-th group, $\mu(x)$ is the grand mean function (i.e. the average function across all treatments), $\alpha_k(x)$ represents the specific effects of being in a specific treatment and the residual function, $\varepsilon_{ik}(x)$ is the unexplained variation specific to the i-th observation within the k-th group. The model in equation (4) can be written in matrix notation as:

$$\mathbf{f}(x)=\mathbf{Z}\boldsymbol{\gamma}(x)+\boldsymbol{\varepsilon}(x) \quad (5)$$

where $\boldsymbol{\gamma}(x) = \left(\gamma_0 = \mu(x),\ \gamma_1 = \alpha_1(x),\ \ldots\ldots, \gamma_K = \alpha_K(x)\right)'$ is the (K+1) vector of parameter functions, $\mathbf{f}(x)$ is a N vector of functional observations, $\boldsymbol{\varepsilon}(x)$ is a vector of N residual functions and $\mathbf{Z}$ is a design matrix of dimension (N, K + 1), coding group membership. In particular each row of the matrix $\mathbf{Z}$ corresponds to a single observation; the first column consists entirely of ones to represent the overall mean and the subsequent K columns correspond to different groups with value one if the observation belongs to the k-th group, zero otherwise.

In order to ensure the identifiability of treatments functions $\alpha_k$, the sum to zero constrained is imposed:

$$\sum_{k=1}^{K} \boldsymbol{\gamma}_k(x) = 0 \qquad \forall x \quad (6)$$

The model is equivalent to standard ANOVA, with the difference that the parameter $\boldsymbol{\gamma}(x)$, and hence the predicted observations $\hat{\mathbf{f}}(x) = \mathbf{Z}\boldsymbol{\gamma}(x)$, are vectors of functions rather than vectors of numbers.

The parameter vector $\boldsymbol{\gamma}(x)$ can be estimated using the standard least squares criterion; thus, minimizing the residual sum of squares:

$$\mathrm{LMSSE}(\gamma) = \int \left[\mathbf{f}(x) - \mathbf{Z}\boldsymbol{\gamma}(x)\right]' \left[\mathbf{f}(x) - \mathbf{Z}\boldsymbol{\gamma}(x)\right] dx \quad (7)$$

Minimizing equation (7) subject to the constraint in equation (6), gives the least squares estimates of the functional parameters:

$$\boldsymbol{\gamma}(x) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{f}(x) \quad (8)$$

If there are no particular restrictions on the way in which $\boldsymbol{\gamma}(x)$ varies as a function of x, it is possible to minimize the discrete version of equation (7), individually for each x:

$$\left\|\mathbf{f}(x) - \mathbf{Z}\boldsymbol{\gamma}(x)\right\|^2 \quad (9)$$

In order to determine if there is any statistically significant differences between group functions, a pointwise F statistic can be used (Ramsay and Silverman, 2005):

$$F(x) = \frac{Var[\hat{\mathbf{f}}(x)]}{\frac{1}{N}\sum_{i=1}^{N}[\mathbf{f}_i(x) - \hat{\mathbf{f}}(x)]^2} \quad (10)$$

where $\hat{\mathbf{f}}(x)$ are the predicted values from a fitted fANOVA model as in equation (4). The equation (10) gives a function built from the series of point estimates at each point of the domain. It is the dependence of this quantity from x that makes the procedure different from the standard univariate or multivariate case.

**Application: fish biodiversity in the province of Arezzo**

In order to provide an example of the advantages of the parametric FDA approach in the analysis of water quality monitoring network, the framework described has been applied to a real data set concerning ichthyic biodiversity in the province of Arezzo, Italy (further details on the data may be found in http://www.ittiofauna.org/provinciarezzo/carta ittica/index.htm).

In 2006 fish abundance data have been collected for a total of 32 species and 104 streams which belong to the basin of four important rivers of Central Italy: Arno, Tevere, Marecchia and Foglia.

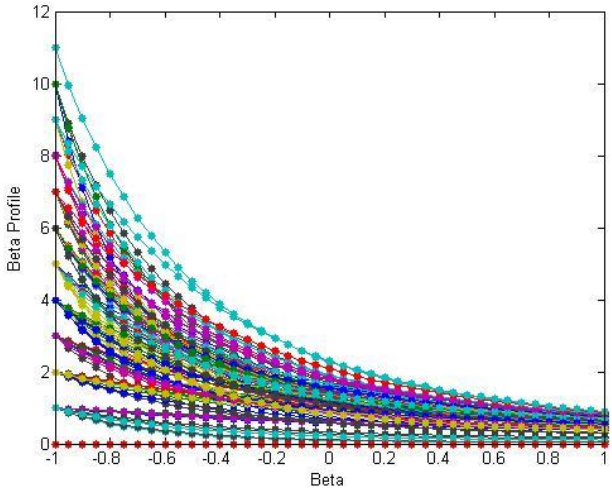


**Figure 1**. Functional effects of fish zonation in the province of Arezzo.

Ichthyic biodiversity in the province of Arezzo has been evaluated through the β-profiles in equation (1).

As shown in figure (1), it is not possible to identify a clear order between river streams because the profiles cross each others. However it seems that the discrimination among river streams is mainly explained by species richness (for $\beta = -1$). It is possible to distinguish clearly an extreme situation of full dominance, with a site (in the basin of Chiana) with only one species.

Naturally, several variables affect biodiversity and the analysis of their interaction is complex. As an example, we consider the classification of European rivers according to fish fauna zonation which represents the fish fauna found in them. This variable shows variations in taxonomic composition which are related to physical and chemical changes. Thus, it describes the spatial distribution of fish by identifying different habitat. In order to provide information on the effect of habitat on fish biodiversity in the province of Arezzo, we resort to this qualitative factor, named zonation of fish fauna. In particular, we refer to the classification proposed by Huet (1949) which distinguishes four zones. In the province of Arezzo we have only two zones: Salmonids and Cyprinus zones. The first zone is usually characterized by slope and cold (max 15°C) rivers, high and well oxygenated water, fast water stream, uneven substrate (with rock, stones, pebbles and gravel) and absence of aquatic vegetation. In the second case, instead, the rivers present slight slope, warm waters in summer, very slow water stream and substrate prevalently muddy. In order to quantify how much of the pattern of biodiversity variation is

explainable by the level of fish zonation, the fANOVA model in equation (4) has been applied under the constraint in equation (6). Figure (2) displays the two functional effects of being in a specific zone. It is evident as Cyprinus zone exerts a positive effect on fish biodiversity. This effect is present throughout the whole domain, and especially for $\beta > 0$; therefore there is a low effect for species richness $(\beta = -1)$.

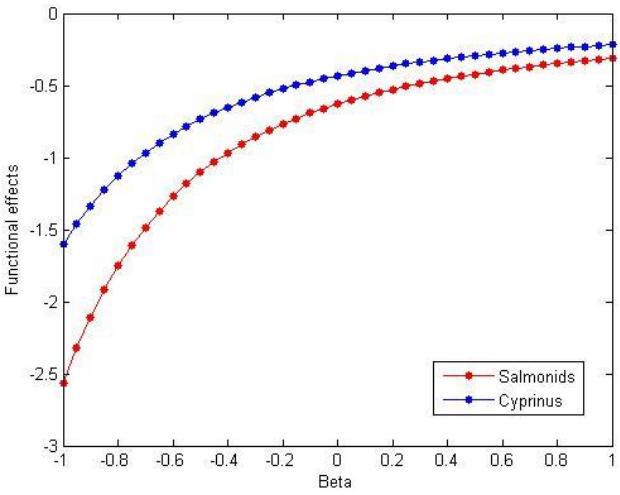


**Figure 2**. Functional effects of fish zonation in the province of Arezzo.

Figure (3) shows the predicted β-profile for each fish zonation group. Obviously, the lower diversity is present in the group of Salmonids. Since the two profiles no intersect each other, we can say that in the Cyprinus zones there is greater biodiversity; thus, in the river streams of the province of Arezzo, there is a prevalence of these fish species in terms of richness and evenness.

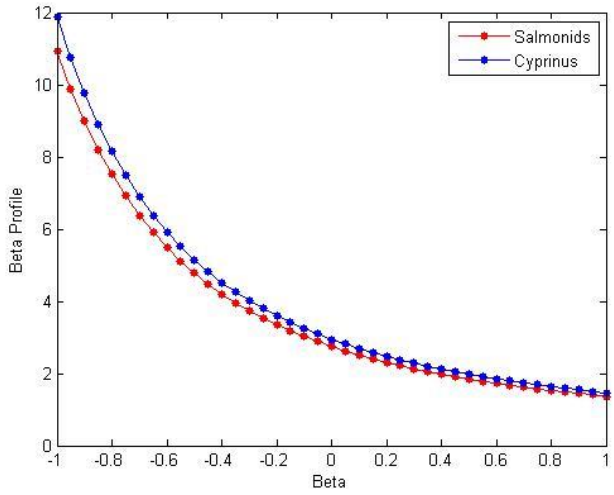


**Figure 3**. Estimated β profiles for each of the two fish zonation in the province of Arezzo.

Figure (4) shows the pointwise F test function in order to formally test the null hypothesis that there is no statistically significant differences among group functions. In particular the blue curve represents the observed F statistic calculated as in equation (10); while the grey line indicates the

5% significance level for the F distribution with (K-1) and (N-K) degree of freedom. In this case we have 1 and 102 degree of freedom respectively, so the value of the 5% significance level of F is equal to 2.75. The observed F statistic is everywhere above the significance level, so we can conclude that there are clear differences between the zonation groups in terms of their mean function.

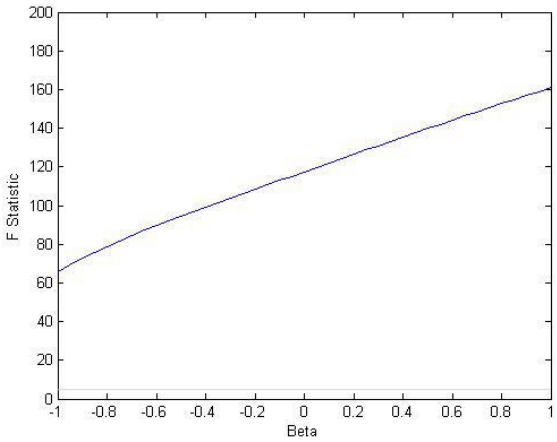


**Figure 4**. Functional F test for the fANOVA model in the province of Arezzo.

## Conclusions

In this paper we have presented some advantages provided by the study of biodiversity through the functional approach. In particular, the joint use of diversity profiles and functional parameters allows us to consider diversity in its multidimensional aspect, evaluating it in relation to richness and evenness. Furthermore, statistical analysis techniques such as the analysis of variance can be implemented in order to understand the relationships between functional data and other variables of particular interest.

## References

Burger, J., Gochfeld, M., Powers, C., Clarke, J., Brown, K., Kosson, D., Niles, L., Dey, A., Jeitner, C., Pittfield, T. (2013). Determining environmental impacts for sensitive species: Using iconic species as bioindicators for management and policy. *Journal of Environmental Protection* 4, 87–95.

De Sanctis, A., Di Battista, T. (2012). Functional analysis for parametric families of functional data. *International Journal of Bifurcation and Chaos* 22 (9), 1250226–1–1250226–6.

Di Battista, T., Fortuna, F. (2013). Assessing biodiversity profile through fda. *Statistica* 1, 69–85.

Ferraty, F., Vieu, P. (2006). Nonparametric functional data analysis. Springer, New York.

Gattone, S., Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society* 58, 267–284.

Gove, J., Patil, G., Swindel, D., Taillie, C. (1994). Ecological diversity and forest management. In: Patil, G., Rao, C. (Eds.), Handbook of Statistics, vol.12, *Environmental Statistics*. Elsevier, Amsterdam, pp. 409–462.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics* 17, 65–80.

Heywood, V., Watson, R. (1995). Global biodiversity assessment. Cambridge University Press, Cambridge, UK.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432.

Huet, M. (1949). Apercu des relations entre la pente et les populations piscicoles des eaux courantes. *Schweitz. Zeitschr. Hydrol.* 11, 333–351.

Paoletti, M. (1999). Using bioindicators based on biodiversity to assess landscape sustainability. *Agriculture, Ecosystems and Environment* 74, 1–18.

Patil, G., Taillie, C. (1979). An overview of diversity. In: Grassle, J., Patil, G., Smith, W., Taillie, C. (Eds.), Ecological Diversity in Theory and Practice. International Co-operative Publishing House, Fairland, MD, pp. 23–48.

Patil, G., Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association* 77, 548–567.

Pavoine, S., Doledec, S. (2005). The apportionment of quadratic entropy: a useful alternative for partitioning diversity in ecological data. *Environmental and Ecological Statistics* 12, 125–138.

Ramsay, J., Silverman, B. (2005). Functional Data Analysis, 2nd edn. Springer, New York.

Rudin, W. (2006). Real and complex analysis. McGraw-Hill, New York.

Shannon, C. (1948). A mathematical theory of communication. Bell Syst. tech. J. 27, 379–423.

Simpson, E. (1949). Measurement of diversity. *Nature* 163, 688.