# Examining indices of individual-level resource specialization

J. A. Martín-Fernández[1], M. E. Pierotti[2] and C. Barceló-Vidal[1]

[1]Dept. Informàtica i Matemàtica Aplicada, UdG, Campus Montilivi, Edifici P-IV. E-17071, Girona, Spain
josepantoni.martin@udg.edu
[2]Dept. of Biology, East Carolina University, Greenville, NC 27858, USA

## 1 Introduction

The variety of resources that a population exploits is known as the "niche width". A particular population has a narrow niche if only few kinds of the available resources are exploited by its members. When the individuals of a population exploit many different resources, then the population has a wide niche. From this point of view it seems that the niche is a property of the population as a whole. However, it is well known that many apparently generalist populations are in fact composed of individual specialists, that is, members that use only small subsets of the population's niche. This approach justifies the definition of indices to measure the individual-level resource specialization. Although this kind of analysis could be applied to any niche variation: oviposition sites, habitat, etc., we focus the discussion in terms of analysis of diet data. So as to measure species niche breadth a comparison between the frequency distribution of the species' resource use with that of all available resources is carried out. When a measure of individual specialization is considered, then one should compare the population's total diet with the individual use. In particular, the total niche width of a population should be compared with its two components: within and between-individual variation. In this sense, in the literature several indices of intrapopulation niche variation are proposed. Our goal is to describe, compare and evaluate four of the most relevant indices applied in ecology. In this work we point out how these techniques could be developed in a compositional framework, particularly when these indices are applied to discrete diet data [e.g. frequency of different prey specimen in the diet].

## 2 Background

Our description of the background is mostly based in Bolnick et al. (2002, 2003). After these articles, many other authors have dealt with this topic until nowadays, both in theoretical as practical studies. A simple search in Google Scholar—2011 March 12 (11:50am), www.scholar.google.es—about Bolnick et al. (2003) shows that this paper has been cited in 371 articles. Among the more recent ones, one could remark the papers Araújo et al. (2008) and Bolnick et al. (2010) because they update theoretical progress in the techniques related with this topic.

It is well known that total niche width of a population ($TNW$) has two components:

- the variation in resource use within individuals or within-individual component: $WIC$,
- the variation between individuals or between-individual component: $BIC$,

so that $TNW = WIC + BIC$. The $WIC$ is designed to analyze the level of concentration in the preferences of the individuals. On the other hand, $BIC$ has the goal to measure the similarity of the behaviour between the individuals.

Consider the case that the data matrix $\mathbf{X}$ contains continuous data. That is, its entry $x_{ij}, i = 1, \ldots, N; j = 1, \ldots, D$ stands for the size or the weight—or other *continuous* measure—of the $j$th prey or resource item in individual $i$'s diet. After Bolnick et al. (2002), using standard statistical analysis, the above elements for the niche variation are defined as

$$\begin{array}{lcl} TNW & = & \frac{1}{ND}\sum_{i,j}(x_{ij} - \overline{\mathbf{X}})^2, \\ WIC & = & \frac{1}{ND}\sum_{i}\sum_{j}(x_{ij} - \overline{\mathbf{x}}_i)^2, \\ BIC & = & \frac{1}{N}\sum_{i}(\overline{\mathbf{x}}_i - \overline{\mathbf{X}})^2, \end{array} \qquad (1)$$

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

where $\overline{\mathbf{x}}_i$ and $\overline{\mathbf{X}}$ are respectively the arithmetical mean of the individual $i$'s diet and the entries in the data matrix. The ratio $WIC/TNW$ is proposed (Bolnick et al., 2002) as a measure of the relative degree of individual specialization. Note that this ratio varies in a range $[0, 1]$ and for those populations where all the individuals have exactly the same behavior it holds that $WIC/TNW = 1$. On the opposite, $WIC/TNW = 0$ when the individuals have not preference for any prey but the behaviors between individuals are different.

In those studies where the data are discrete the standard definition of these three elements is different. Here the initial entry $c_{ij}$ in the data matrix $\mathbf{C}$ counts the number of diet items in individual $i$'s diet that fall in category $j$. Each row $\mathbf{c}_i$ ($i$'s diet vector) is closed to one dividing its elements by its sum $n_i = \sum_j c_{ij}$ and obtaining a *proportion* matrix $\mathbf{X}$, with entries $x_{ij}$ describing the proportion of the $j$th prey category in individual $i$'s diet. Since the data are proportions the proposed standard approaches is based on the Shannon entropy as a measure of variability in a vector of proportions. Recall that Shannon entropy $\mathbf{x} = (x_1, \ldots, x_D)$ is defined as $\mathrm{H}(\mathbf{x}) = -\sum x_i \ln x_i$ and takes values in the range $[0, \ln D]$, where $D$ is the number of categories. Note that an individual $\mathbf{e}$ that has no preference in its diet $\mathbf{e} = (1/D, \ldots, 1/D)$ takes the maximum value $\mathrm{H}(\mathbf{e}) = \ln D$. On the other hand, an individual that only likes one kind of prey, for example $\mathbf{v}_1 = (1, 0, \ldots, 0)$ has the minimum entropy value $\mathrm{H}(\mathbf{v}_1) = 0$. In this framework, Bolnick et al. (2002) introduce

$$\begin{aligned} TNW &= \mathrm{H}(\mathbf{r}) = -\textstyle\sum_j r_j \ln r_j, \\ WIC &= \textstyle\sum_i q_i \mathrm{H}(\mathbf{x}_i) = -\sum_i q_i \left(\sum_j x_{ij} \ln x_{ij}\right), \\ BIC &= TNW - WIC, \end{aligned}$$

where the two vectors $\mathbf{q}$ and $\mathbf{r}$ could be seen as respectively two *average vector* across the columns and rows of data matrix $\mathbf{C}$ or weighted *marginal profiles* of the data matrix of proportions $\mathbf{X}$:

$$\begin{aligned} q_i &= \frac{\frac{1}{D}\sum_j c_{ij}}{\sum_i\left(\frac{1}{D}\sum_j c_{ij}\right)} = \frac{n_i \sum_j x_{ij}}{\sum_i\left(n_i \sum_j x_{ij}\right)},\ i = 1, \ldots, N, \\ r_j &= \frac{\frac{1}{N}\sum_i c_{ij}}{\sum_j\left(\frac{1}{N}\sum_i c_{ij}\right)} = \frac{\sum_i (n_i x_{ij})}{\sum_i\left(n_i \sum_j x_{ij}\right)},\ j = 1, \ldots, D, \end{aligned}$$

where $n_i = \sum_j c_{ij}$. In addition to the ratio $WIC/TNW$, which informs about the general degree of individual specialization, Bolnick et al. (2002) introduce two indices to measure the diet overlap between an individual $i$ and the population: $PS_i$ and $W_i$. The former, $PS_i$ is based on the Manhattan (or City-Block) distance between the $i$'s diet vector $\mathbf{x}_i$ and the *average vector* $\mathbf{r}$:

$$PS_i = 1 - \frac{\sum_j |x_{ij} - r_j|}{2}. \tag{2}$$

Note that when an individual has exactly the same preferences than the average vector then $PS_i = 1$. When the individual only eats one kind of preys—namely $j$—then $PS_i = r_j$. The latter, $W_i$, is based on the likelihood ratio assuming a multinomial probability distribution:

$$W_i = \prod_j \left(\frac{r_j}{x_{ij}}\right)^{x_{ij}}. \tag{3}$$

Both measures inform about the level of specialization of one particular individual. When the average across all the individuals is computed then one obtains two measures of relative degree of individual specialization in the population: $\overline{PS}$ and $\overline{W}$.

From another completely different approach, Araújo et al. (2008) propose the index $E$. In the framework of network analysis, the index $E$ informs about the mean pairwise diet dissimilarity between individuals and is defined as $E = 1 - \overline{ps_{ik}}$. Here, $ps_{ik}$ is again based on the Manhattan distance between two diet vectors: $ps_{ik} = 1 - \frac{\sum_j |x_{ij} - x_{kj}|}{2}$. In fact the average $\overline{ps_{ik}}$ takes into account that the similarity $ps_{ik}$ is symmetric and is computed as $\overline{ps_{ik}} = \frac{\sum_{i,k} ps_{ik}}{N(N-1)/2}$.

Numerous studies have discussed in the past the properties and weaknesses of the above four measures—$WIC/TNW$, $\overline{PS}$, $\overline{W}$, $E$—but is not the aim of this work to summarize these previous analysis. We want to emphasize the fact that the four measures are based in statistical concepts applied

to vectors of diet distribution: arithmetical mean, variance, Manhattan distance and multinomial probability distribution. This fact encourages our goal to examine these indices from a compositional point of view.

## 3 A lot of questions, only few answers.

After Aitchison (1986) compositional data analysis (CODA) has progressed in both theoretical basis and applied techniques. Nevertheless in any study a main initial decision remains: *Are my data compositional?* Sometimes it is obvious that CODA is the more appropriate way for dealing with a problem, but in other situations the analyst have to decide if its data are compositional or not. In our case, when the diet vector contains discrete data the standard approach (Bolnick et al., 2002) transforms the vector diet in a vector of proportions and assumes that the total sum of the individual's diet is not a crucial information. On the other hand, when the matrix of resource data is based on information from continuous variables the standard approach (Bolnick et al., 2002) applies classical statistical tools for univariate data in the real space. Note that the elements—$TNW, WIC, BIC$—in equation (1) are based on arithmetical means and variances, both descriptive measures of a distribution. In other words, although they apply tools to deal with the distributions of the individual's diet they use measures that are affected by the sum of the diet vectors.

With above approach in mind, we review the most important aspects of the topic *individual-level resource specialization* focusing our discussion in those basic concepts related to CODA. We introduce our questions through simple examples:

- *Scale invariance*: let $\mathbf{x} = (45, 15, 21)$ and $\mathbf{x}^* = (15, 5, 7)$ be the diet vectors of two individuals. Do have they the same individual-level resource specialization? Note that the values in the vectors could be both continuous and discrete.

- *Subcompositional coherence*: let $\mathbf{x} = (40, 2, 8)$ and $\mathbf{x}^* = (30, 4, 16)$ be the diet vectors of two individuals, and $\mathbf{s} = (40, 2)$ and $\mathbf{s}^* = (30, 4)$ be its corresponding subvectors in the two first resources. The difference in the individual-level resource specialization between $\mathbf{x}$ and $\mathbf{x}^*$ should be larger than between $\mathbf{s}$ and $\mathbf{s}^*$?

- *Perturbation invariance*: let $\mathbf{x} = (10, 12, 8)$ and $\mathbf{x}^* = (20, 4, 2)$ be the diet vectors of two individuals. The comparison between its level of specialization should be based on the vector $\mathbf{x} - \mathbf{x}^* = (-10, 8, 6)$ or on the vector $\mathbf{x}/\mathbf{x}^* = (1/2, 3, 4)$?

- *Rounded, structural and count zeros*: let $\mathbf{x} = (23, 0, 9)$ be the vector of the resources expended by one individual. The zero value in the second component is it a rounded zero because the individual ate an amount—e.g. weight—very small of such kind of prey? If the data are discrete—counts—the zero value derives from something that, given enough time, we would have otherwise observed?

In CODA the three first questions are crucial. Scale invariance and subcompositional coherence are the two main principles (Aitchison, 1986, postscript: p. 2) and perturbation is the basic operation in the vector space structure of the simplex (e.g. Pawlowsky-Glahn and Buccianti, 2011, chapter 2). The latter question, about the nature of the zeros, requires the collaboration of the expert in the sampling experiment or the analyst who designed the sampling process. Once the nature of the zeros is fixed then one of the techniques recommended (e.g. Pawlowsky-Glahn and Buccianti, 2011, chapter 4) could be applied. In any case, it is very important to state that the answer of above questions determines the decision about which measures are appropriate to evaluate the individual-level of specialization. The standard approach (Bolnick et al., 2002) not satisfies the two principles and is based on classical operations—sum and subtraction—between vectors.

Beyond the possibility to apply CODA to this topic, note that standard approaches not fully use the information given in the covariance structure of the data set. These approaches are mostly based in univariate statistics—mean and variance—and distances. Despite of a distance is a multivariate tool, $\overline{PS}$ measure—equation (2)—is based on the distance between one individual and the *average*

*vector* $\mathbf{r}$ but it not takes into account the spread of the data set. Araújo et al. (2008), in a context of clusters produced by a network analysis, try to capture the multivariate structure through the pairwise distances between individuals included in the measure $\overline{ps_{ik}}$.

The lack of multivariate information in the standard measures suggest two different questions:

- *Mahalanobis distance*: let $\mathbf{x} = (30, 12, 20)$ be the diet vector of one individual. The level of specialization of this individual should be relative to the other individuals in the population. Therefore it seems a reasonable idea to incorporate the covariance structure to the distances to the *average vector* and to the pairwise distances. Could it be useful?

- *Multivariate probability distributions*: the multinomial distribution is a multivariate probability distribution that satisfies $Cov(\mathbf{X}_j, \mathbf{X}_k) = -np_jp_k,\ (j \neq k)$, where $n$ is the sample size and $p_j, p_k$ respectively the probability that a item falls in categories $\mathbf{X}_j$ and $\mathbf{X}_k$. Note that fixed a category $\mathbf{X}_j$ the correlation between such category and any other category $\mathbf{X}_k$ is negative. This fact implies that when an individual varies its preference for a particular kind of prey then all the rest of preferences should to vary in the opposite direction. The question is why two different preferences can't be positively correlated. Is there any multivariate probability distribution that will be able to give such model?

## Acknowledgments

This research has been supported by the Spanish Ministry of Science and Innovation under the project "CODA-RSS" Ref. MTM2009-13272; by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424.

## References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Araújo, M. S., P. R. J. Guimaraes, R. Svanbäck, A. Pinheiro, S. F. dos Reis and D. I. Bolnick (2008). Network analysis reveals contrasting effects of intraspecific competition on individual versus population diets. *Ecology 89*(7), 1981–1993.

Bolnick, D. I., L. H. Yang, J. A. Fordyce, J. M. Davis and R. Svanbäck (2002). Measuring Individual-Level Resource Specialization. *Ecology 83*(10), 2936–2941.

Bolnick, D. I., R. Svanbäck, J. A. Fordyce, L. H. Yang, J. M. Davis, C. D. Hulsey and M. L. Forister (2003). The ecology of individuals: incidence and implications of individual specialization. *Am. Nat. 161*, 1–28.

Bolnick, D. I., T. Ingram, W. E. Stutz, L. K. Snowberg, O. L. Lau, J. S. Paull (2010). Ecological release from interspecific competition leads to decoupled changes in population and individual niche width. *Proceedings of the Royal Society of London, Ser. B. 277*(1689), 1789–1797.

Pawlowsky-Glahn, V., A. Buccianti, (eds) (2011) *Compositional data analysis: Theory and applications.* Wiley, New York (USA). 356 p.