Diversity Measures and Coarse-graining in Data Analysis with an Application Involving Plant Species on the Galápagos Islands

Radu Cornel GUIASU
Department of Multidisciplinary Studies
Glendon College, York University, Toronto, Ontario M4N 3M6, Canada
rguiasu@glendon.yorku.ca

and

Silviu GUIASU
Department of Mathematics and Statistics
York University, Toronto, Ontario M3J 1P3, Canada
guiasus@pascal.math.yorku.ca

ABSTRACT

In a numerical entity-characteristic incidence matrix we can use simple or multiple regression and calculate correlations between pairs of characteristics. However, in order to detect similarities/dissimilarities, interdependence, and multiple probabilistic causality among the characteristics we have to group the entities in classes. The number of uniform classes obtained by coding the given values of these characteristics depends on the balance between the class uncertainty and class ambiguity. The similarity, interdependence, and multiple probabilistic causality among characteristics are analyzed. When a set of entities and the abundance of their components are given, the average within-entity diversity and the average between-entity diversity are studied. The results are applied to the number of endemic and immigrant plant species in the Galápagos Islands.

Keywords: Uniform Coding, Class Uncertainty and Ambiguity, Similarity, Interdependence, Probabilistic Causality, Diversity measure, Endemic and Immigrant Plant Species, Galápagos Archipelago.

1. INTRODUCTION

In analyzing the relationship between a set of entities and a set of characteristics, more often than not the values of these characteristics are different real numbers with different ranges of possible values. This allows us to do simple and multiple regression analysis and to calculate correlations between pairs of characteristics. Such data, however, do not allow a significant analysis of the similarity/dissimilarity, interdependence, and probabilistic causality among the given characteristics or entities. In order to detect them we have to uniformly coarse-grain the specific range of each characteristic, grouping the entities in classes determined by the coding of these ranges. This grouping of entities in classes reduces the initial uncertainty of identifying the given entities but increases the amount of ambiguity induced by the fact that we make no distinction between the entities that are grouped in the same class. The number of possible classes depends on the balance between class uncertainty and ambiguity. Once the coarse-graining is performed, we can measure the dissimilarity between characteristics (using, for instance, the Hamming distance), the global interdependence or connection between characteristics (using either the well-known Watanabe's logarithmic measure or the new quadratic measure proposed in this paper), and the probabilistic causality between characteristics, (using both the indicator applied by Tuldava [1] for the one-cause model and its generalization given here for the multiple-cause model). Also, if we have a set of entities, like the distinct local habitats of a larger region, and a probability distribution for each entity, reflecting the abundance of its components (species found in these habitats, for example), a weighted measure of diversity may be used for calculating the average within-entity component diversity and the average between-entity component diversity in the given set of entities.

The methodology presented here is applied to the data set provided by the measurement of seven characteristics in the study of the plant species of 29 islands from the Galápagos Archipelago ([2], [3]). Special attention is given to the relevance of the area versus elevation in measuring the local and global plant species diversity.

2. METHODOLOGY

Let $X = \{x_1, \dots, x_N\}$ be a set of entities and $C = \{c_1, \dots, c_m\}$ a set of characteristics. An incidence $N \times m$ matrix $A = [a_{ij}]$ is given, assigning a real number a_{ij} to each pair (x_i, c_j) representing the value of the characteristic c_j for the entity x_i. If the characteristics are independent, the main problem in classification theory is to group the entities in disjoint classes such that the interdependence inside these classes is large and the interdependence among classes is small, ideally zero. In many problems, however, the characteristics are not all independent and the analysis is now focused on them in order to see which characteristics are similar/dissimilar and which characteristics influence the behaviour of other characteristics in a kind of probabilistic causality.

Using the given incidence matrix, regression analysis allows us to approximate the unknown true probabilistic dependence between characteristics by rough strictly deterministic, mainly linear, functions, and to calculate the correlation between the pairs of characteristics. The difficulty of such an approach is that similarity and causality between characteristics are not strictly deterministic and, on the other hand, more often than not, the global interdependence between more than two characteristics is essential, which cannot be detected by the standard correlations. The difficulties are even greater when different characteristics have very different scales of possible values. A possible way of solving such an impasse is by coarse-graining the range of possible values of the characteristics from the incidence matrix.

2.1 Grouping the entities in distinct classes

Coarse-graining of the characteristic values from the given incidence matrix is based on a coding method which will be explained in the next subsection. The coding results in a partition of the entities in disjoint classes. Grouping the entities in classes decreases the amount of initial uncertainty on the set of entities, because it is easier to identify a class than an individual entity, but increases the amount of ambiguity, because when a class is given, no distinction among its components is made ([4]). In other words, if we are too close to a forest each tree seems to be different whereas if we are too far, all trees seem to be alike; the secret is to be somewhere in-between, which means to group the similar trees in relevant classes.

Let $X = \{x_1, \dots, x_N\}$ be a set of entities, and let:

$$p(x) > 0, \qquad \sum_{x \in X} p(x) = 1,$$

be a probability distribution on X. Let $P_n(X) = \{X_1, \dots, X_n\}$ be a partition of X in n disjoint subsets. Define:

$$p(X_i) = \sum_{x \in X_i} p(x), \quad (i = 1, \dots, n).$$

The amount of uncertainty on X measured by Shannon's entropy ([5]):

$$H(X) = -\sum_{x \in X} p(x) \log p(x).$$

Similarly, the amount of uncertainty on the partition $P_n(X)$ is:

$$H(P_n(X)) = -\sum_{i=1}^{n} p(X_i) \log p(X_i).$$

In the information balance:

$$H(X) + H(P_n(X) \mid X) = H(P_n(X)) + H(X \mid P_n(X)),$$

where $H(P_n(X) \mid X)$ is the conditional entropy of $P_n(X)$ given X, and also $H(X \mid P_n(X))$ is the conditional entropy of X given $P_n(X)$, we have:

$$H(P_n(X) \mid X) =$$

$$-\sum_{x \in X} p(x) \sum_{i=1}^{n} p(X_i \mid x) \log p(X_i \mid x) = 0,$$

because the probability of the class X_i conditioned by the entity x is equal to:

$$p(X_i \mid x) = \begin{cases} 1 & \text{if } x \in X_i, \\ 0 & \text{if } x \notin X_i, \end{cases}$$

which implies:

$$p(X_i \mid x) \log p(X_i \mid x) = 0,$$

for every $i = 1, \dots, n$, extending by continuity the function $-t \log t$ to be equal to zero at $t = 0$. Therefore:

$$H(X) = H(P_n(X)) + H(X \mid P_n(X)),$$

which shows that the uncertainty $H(P_n(X))$ decreases from $H(X)$ to 0, whereas, at the same time, the amount of ambiguity:

$$H(X \mid P_n(X)) =$$

$$= -\sum_{i=1}^{n} p(X_i) \sum_{x \in X_i} p(x \mid X_i) \log p(x \mid X_i)$$

increases from 0 to $H(X)$ in the extreme cases when $n = N$ and $n = 1$, corresponding to $P_N(X) = \{\{x_1\}, \dots, \{x_N\}\}$ and $P_1(X) = \{X\}$, respectively, where $p(x \mid X_i)$ is the probability of the entity x

conditioned by the class X_i. If we want to balance these two opposite components of $H(X)$, we choose the number n of classes such that:

$$H(P_n(X)) \approx H(X \mid P_n(X)) \approx \frac{1}{2} H(X).$$

Quite often the numerical values from the incidence matrix make the entities from X to be all different. In such a case, $p(x) = 1/N$, for each $x \in X$, and if all classes of the partition $P_n(X)$ contain the same number of entities, namely N/n, then we have:

$$p(X_i) = \frac{N}{n} \times \frac{1}{N} = \frac{1}{n}, \quad (i = 1, \ldots, n),$$

$$H(X) = \log N, \qquad H(P_n(X)) = \log n.$$

In such a case, if we want to balance the uncertainty and ambiguity, we choose the number of equal classes to satisfy:

$$\log n \approx \frac{1}{2} \log N \quad \text{which implies:} \quad n \approx \sqrt{N}.$$

2.2 Coarse-graining the range of the values of the characteristics

The coarse-graining method has been applied long ago in statistical mechanics, where the macroscopic space is viewed as a partition of the microscopic space, and, more recently, in different fields like the study of proteins ([6]), the spectral analysis ([7]), and the theory of price fluctuations ([8]). Here we are interested in coarse-graining the possible values of the characteristics in an entity-characteristic incidence matrix. In such an incidence matrix, to make the different characteristics comparable, we order in increasing way the values of each characteristic and we split the corresponding range into the same number, say n, of equal parts; each such part gets a distinct symbol. For instance, if we use three symbols, namely 1,2,3, for the three equal parts of the increasing values of each characteristic, this means that the initial numerical values of the respective characteristic are coded into the vaguer values: 'small', 'medium', and 'large', respectively. The initial uncertainty about identifying an entity is larger than the uncertainty about identifying a class induced by coding, but this decrease in uncertainty also leads to the creation of an inherent ambiguity about distinguishing among the entities making up the same class. If we use uniform coding for the values of a characteristic y_j using four symbols, say 1,2,3,4, then we arrange in increasing order the values of this characteristic for the entities from the given incidence matrix and we assign the symbol 1 to all values up to the first quartile, the symbol 2 to all values between the first quartile and the median, the symbol 3 to the values between the median and the third quartile, and the symbol 4 to the values larger than the third quartile. This kind of uniform coding allows a fair comparison among the characteristics in spite of their very different ranges of initial numerical values. At the same time, we start with a uniform distribution of the symbols for each characteristic, which amounts to a maximum entropy distribution of the symbols assigned to the values taken by each characteristic allowing the largest interdependence among the characteristics.

Another way of coding is by ignoring the outliers first and dividing the resulting range into n equal subintervals before assigning a distinct code symbol to all the values that belong to the same subinterval of the range. In the binary case, for instance, we assign the symbol 0 to all values which are smaller or equal to the trimmed mean and the symbol 1 to all the other values.

After the coding is done for each characteristic, using the same set of n symbols, the columns of the new incidence matrix are the vectors $c_1, \ldots, c_m$, whose N components are the symbols used in the respective coding. Now, it is easy to compare the characteristics and a very convenient measure of dissimilarity between two characteristics c_j and c_k is the Hamming distance $d_H(c_j, c_k)$, which is equal to the number of different components of the vectors corresponding to the columns c_j and c_k in the coded incidence matrix ([9]). The corresponding measure of similarity is $N - d_H(c_j, c_k)$.

The entropic measures of global interdependence ([10]–[12]) may be easily applied to the columns of the coded incidence matrix. Thus, Watanabe's measure of connection among all m characteristics is:

$$W(c_1, \ldots, c_m) = H(c_1) + \ldots + H(c_m) - H(c_1, \ldots, c_m),$$

where $H(c_1, \ldots, c_m)$ is the joint Shannon's entropy of the vectors (columns) $(c_1, \ldots, c_m)$. Similarly, for any subset of characteristics, in particular, the interdependence between two characteristics, say c_1 and c_2, is:

$$W(c_1, c_2) = H(c_1) + H(c_2) - H(c_1, c_2).$$

Instead of looking for a deterministic function which shows how a characteristic c_1 is determined by the characteristic c_2, a rigid causality which may not exist in fact, we can measure the probabilistic causality from c_2 to c_1, given by the decrease in the uncertainty on c_1 if the coded value of c_2 is known, divided by the uncertainty on c_1, namely ([1]):

$$R(c_1 \mid c_2) =$$

$$= [H(c_1) - H(c_1 \mid c_2)]/H(c_1) = W(c_1, c_2)/H(c_1).$$

This entropic measure of the probabilistic causality of c_1 given c_2 may be extended to the case when there are several causes, say c_2 and c_3, involved. Thus:

$$R(c_1 \mid c_2, c_3) = [H(c_1) - H(c_1 \mid c_2, c_3)]/H(c_1) =$$

$$= [W(c_1, c_2, c_3) - W(c_2, c_3)]/H(c_1),$$

$$W(c_1, c_2, c_3) = H(c_1) + H(c_2) + H(c_3) - H(c_1, c_2, c_3).$$

Assume now that in the coded incidence matrix: $p_j(k)$ is the relative frequency of the symbol k in the N rows of the column c_j; $p_{1,2}(k, \ell)$ is the relative frequency of the pair of symbols (k, ℓ) in the N rows of the pair of columns (c_1, c_2); $p_{1|2}(k \mid \ell)$ is the relative frequency of the symbol k in those rows of column c_1 in which the column c_2 has the symbol ℓ. Then, a new quadratic measure of the interdependence (or connection) between the columns c_1 and c_2 is:

$$V(c_1, c_2) = \sum_k \sum_\ell \frac{p_{1,2}^2(k, \ell)}{p_1(k) p_2(\ell)} - 1.$$

As $\log x \le x - 1$, where log is the natural logarithm, the relationship between Watanabe's measure $W(c_1, c_2)$ and the quadratic measure $V(c_1, c_2)$ is:

$$0 \le W(c_1, c_2) =$$

$$= \sum_k \sum_\ell p_{1,2}(k, \ell) \log \frac{p_{1,2}(k, \ell)}{p_1(k) p_2(\ell)} \le V(c_1, c_2).$$

Also, If c_1 and c_2 are independent, then:

$$p_{1,2}(k, \ell) = p_1(k)\, p_2(\ell),$$

for all k and ℓ, which implies $V(c_1, c_2) = 0$. We note that, equivalently,

$$V(c_1, c_2) = \sum_k \sum_\ell p_{1|2}(k \mid \ell) p_{2|1}(\ell \mid k) - 1.$$

2.3 Measuring diversity

Let $X = \{x_1, \ldots, x_N\}$ be a set of entities. Each entity x_i has the components $\{y_{ij}, j \in I_i\}$. Let n_i be the number of elements of I_i. Thus, the entity x_i has n_i components. Let n be the number of the elements of the set $I = I_1 \cup \ldots \cup I_N$. The sets $I_1, \ldots, I_N$ are not necessarily disjoint, which means that different entities could have common components. We have:

$$n_i \le n, \quad (i = 1, \ldots, N); \quad n \le n_1 + \ldots + n_N.$$

For each entity x_i, let p_{ij} be a probability distribution on its components $\{y_{ij}, j \in I_i\}$. Thus,

$$p_{ij} > 0, \quad (j = 1, \ldots, n_i), \quad \sum_{j=1}^{n_i} p_{ij} = 1,$$

for each $i = 1, \ldots, N$. The diversity of the components of the entity x_i is measured by the indicator:

$$D(x_i) = n_i \left(1 - \sum_{j \in I_i} p_{ij}^2 \right). \tag{1}$$

called here the Rich-Gini-Simpson indicator. The ratio $D(x_i)/n_i$ is the old Gini-Simpson indicator ([13], [14]) for the probability distribution $\{p_{ij}, j \in I_i\}$. The Gini-Simpson index, very popular with many ecologists, was proved recently to behave badly when the number of entities is very large. The Rich-Gini-Simpson indicator, however, preserves all qualities of the classic Gini-Simpson index while behaving very well even if the number of entities is very large.

If $\lambda_1, \ldots, \lambda_N$ are positive relative weights assigned to the entities, such that:

$$\lambda_i > 0, (i = 1, \ldots, N); \quad \sum_{i=1}^N \lambda_i = 1,$$

the alpha, gamma, and beta diversities are defined by:

$$\alpha\mathrm{Div} = \sum_{i=1}^N \lambda_i\, D(x_i), \tag{2}$$

$$\gamma\mathrm{Div} = n \left[1 - \sum_{j \in I} \left(\sum_{i=1}^N \lambda_i p_{ij} \right)^2 \right], \tag{3}$$

$$\beta\mathrm{Div} = \gamma\mathrm{Div} - \alpha\mathrm{Div} > 0, \tag{4}$$

respectively. The αDiv measures the average within-entity diversity. The γDiv measures the diversity of the set of averaged entities. The βDiv measures the average between-entity diversity, showing the change in the average diversity when we compare the local entities x_1,…,x_N to the global set of entities X.

The formulas (2)-(4), have been given in [15], only for the particular case when $N = 2$, $n_1 = n_2$, $I_1 = I_2$, $\lambda_1 = \lambda_2 = 1/2$, and in the general case in [16], but both in [15] and [16], the Shannon entropy was used as a measure of diversity instead of (1). The concepts of alpha, beta, and gamma diversities were introduced in [17], using the multiplicative partitioning of diversity, and extended to the additive partitioning of diversity (4) in [18] and [16].

As proposed in [16], the ratio between the alpha diversity and the gamma diversity may be used as an index of similarity between the entities of the set X. We denote:

$$Sim = \frac{\alpha\mathrm{Div}}{\gamma\mathrm{Div}}. \tag{5}$$

3. APPLICATION

The methodology discussed in the previous section was applied to the analysis of the relationship between the endemic and immigrant plant species found on the islands of the Galápagos Archipelago. This archipelago is made up of fairly recently formed islands of volcanic origin. The endemic species are considered unique to the Galápagos Islands, and are therefore only found in this archipelago. It is assumed that these endemic species evolved on these islands. The immigrant species category includes both the species which colonized the islands a long time ago, and arrived on the islands by natural means (dispersed by wind, birds, ocean currents, etc.) and the more recently arrived species, presumably brought to the islands as a result of human activities ([19], [20]). From a practical standpoint, it is often very difficult, if not impossible, to determine with certainty the exact ways in which various species arrived at a particular location.

The 29 islands listed are our entities (x_1: Baltra; x_2: Bartolomé; x_3: Caldwell; x_4: Champion; x_5: Coamaño; x_6: Daphne Major; x_7: Darwin; x_8: Eden; x_9: Enderby; x_{10}: Española; x_{11}: Fernandina; x_{12}: Gardner (near Española); x_{13}: Gardner (near Santa Maria); x_{14}: Genovesa; x_{15}: Isabela; x_{16}: Marchena; x_{17}: Onslow; x_{18}: Pinta; x_{19}: Pinzón; x_{20}: Las Plazas; x_{21}: Rábida; x_{22}: San Cristóbal; x_{23}: San Salvador; x_{24}: Santa Cruz; x_{25}: Santa Fé; x_{26}: Santa Maria; x_{27}: Seymour; x_{28}: Tortuga; x_{29}: Wolf).

There are seven characteristics (c1: 'Total number of species observed on the respective island'; c2: 'Number of endemic species observed on the respective island'; c3: 'Number of immigrant species observed on the respective island'; c4: 'Area (in km^2) of the respective island'; c5: 'Elevation (in m) of the respective island'; c6: 'Distance (in km) from Santa Cruz (taken to be the central island of the archipelago)'; and c7: 'Area (in km^2) of the adjacent island'). The incidence matrix is given in Table 1. The columns c1, c2, c4, c6 (and the sixth component of c5) of the data set were taken from [2] and [21] (pp. 291-293), whereas the columns c5 (except the sixth component of it) and c7 are taken from [3].

We note that with respect to the seven characteristics listed, each entity is distinct, which makes it almost impossible to detect any similarity or causality among the columns. Also, the scale of values is different for different characteristics. Using the standard techniques from statistical inference, however, we can apply the regression analysis and calculate the correlations between the given characteristics.

3.1 Regression analysis

In the paper [22], published in 1963, regression analysis was applied to an incidence matrix containing only 17 of the islands listed in Table 1. The focus was on the total number of species (characteristic c1) and the main conclusion was that the island elevation (characteristic c5) had the most significant impact on the values taken by c1. Ten years later, regression analysis was applied again in [2], this time using the data referring to 29 islands of the Galápagos Archipelago, and the main conclusion was that the number of plant species (characteristic c1) was influenced mainly by the area of the island (characteristic c4). If we apply regression analysis using the computer software Minitab version 11 on Windows, focusing not only on the total number of species (c1) but also on the number of endemic species (c2) and the number of immigrant species (c3), we obtain the following main results. Let us name: c1 'Species', c2 'Endemic', c3 'Immigrant', c4 'Area', c5 'Elevation', c6 'DistSC', and c7 'AdjArea'. After storing the numerical values of the characteristics from Table 1 in the Minitab columns c1–c7, using the Minitab command 'regress c1 4 c4–c7', where 4 specifies the number of variables used in regression analysis, we obtain:

Species=51.4−0.007 Area+0.023 Elevation−0.395 DistSC−0.038 AdjArea

Endemic=16.6−0.004 Area+0.063 Elevation−0.095 DistSC−0.009 AdjArea

Immigrant=34.8−0.003 Area+0.172 Elevation −0.300 DistSC− 0.029 AdjArea.

Obviously, we can get the corresponding regression equations for c1, c2, and c3 as functions of only one or some of the characteristics c4,…,c7. It is often forgotten that these linear functions are only a rough approximation of the way c1, c2, and c3 depend on the other characteristics. The real dependence among characteristics is not strictly deterministic but probabilistic. We get:

Species = 65.3 + 0.0815 Area
Endemic = 21.8 + 0.0193 Area
Immigrant = 43.5 + 0.0692 Area
Species = 16.0 + 0.197 Elevation
Endemic = 8.96 + 0.0499 Elevation
Immigrant = 7.0 + 0.147 Elevation
Species = 106 − 0.315 DistSC
Endemic = 21.2 − 0.0721 DistSC
Immigrant = 74.5 − 0.242 DistSC
Species = 95.9 − 0.136 AdjArea
Endemic = 28.5 − 0.00244 AdjArea
Immigrant = 67.4 − 0.0112 AdjArea
Species = 21.2 + 0.0189 Area + 0.169 Elevation
Endemic = 9.44 + 0.00172 Area + 0.0473 Elevation
Immigrant = 11.8 + 0.0172 Area + 0.121 Elevation

We can see from the above linear functions that

the island elevation is the most important characteristic as far as the number of different plant species is concerned. The elevation of islands in the Galápagos Archipelago is related to the diversity of habitats found on these islands. The higher the elevation, the greater the diversity of habitats available for various plant species, and the greater the number of vegetation zones present on the islands [23]. A greater habitat variety is generally linked to a greater species diversity. Also, regression analysis shows that where there are many endemic species there are also many immigrant species, because:
Endemic = 9.41 + 0.292 Immigrant
This suggests that habitats which can sustain large numbers of immigrant species can also support, at the same time, large numbers of endemic species. Thus, as long as a habitat contains sufficient relevant resources, a great species diversity, consisting of both endemic and immigrant species, can be found there.

3.2 Correlations

Pearson's correlation is shown in Table 2 for all pairs of characteristics. The correlation between endemic and immigrant species is positive and very strong (0.955), confirming the coexistence between these two categories of plant species. Again, the correlation between the number of plant species and elevation is larger than that between the number of plant species and the area of the respective island. Both the numbers of endemic species and immigrant species have weak negative correlations with the area of the adjacent island and the distance to the centre of the Archipelago, respectively.

3.3 Coarse-graining

We switch from the initial numerical values of the characteristics to symbols after grouping the 29 islands, as uniformly as possible, in four disjoint classes (the four-symbol coding) and in two disjoint classes (the binary coding). This is equivalent to introducing the symbol 1 for 'very small', 2 for 'small', 3 for 'large', and 4 for 'very large', in the four-symbol coding, and, respectively 0 for 'small', and 1 for 'large' in the binary coding.

For c1 (Species), for instance, we use the Minitab command 'describe c1' which gives the minimum value 2.0, the first quartile 11.0, the median 44.0, the third quartile 100.5, and the maximum value 444.0. Then, the four-symbol coding uses the command:

```
code (0.0:11.0)1 (11.0:44.0)2 (44.0:100.5)3 (100.5:500.0)4 c1 c11
```

The column c11 will contain the values of the characteristic c1 for the 29 islands after the four-symbol coding. Similarly, using:

```
code (0.0:44.0)0 (44.0:500)1 c1 c12
```

the column c12 will contain the values of the characteristic c1 for the 29 islands after the binary coding. If two coding intervals overlap, then the first assignment matters. Thus 44.0, from the last command, gets the symbol 0. Table 3 contains the incidence matrix after the four-symbol coding and the binary coding are performed, respectively.

3.4 Similarity

The values of the Hamming distance $d_H(c_i, c_j)$ between the pairs of characteristics from the four-symbol incidence matrix and from the binary incidence matrix are given in Table 4. As the Hamming distance counts the number of different values of two vectors with the same number of components, it measures the dissimilarity between the corresponding characteristics. Therefore, a small value of it means a large similarity. The largest similarity is between the endemic species and immigrant species. Also, both the endemic species and the immigrant species have more similarity with the corresponding island area than with the corresponding island elevation. Let us note that if the Hamming distance is calculated for the initial incidence matrix, we obtain that there is no similarity between the number of endemic species and the elevation of the corresponding island (the Hamming distance between them is equal to 29), whereas the similarity between the same characteristics is 18 (the Hamming distance is 11) after the four-symbol coding and 25 (the Hamming distance is 4) after the binary coding. This shows that the coarse-graining is necessary if we want to detect any kind of similarity.

3.5 Interdependence

Watanabe's measure of global interdependence may be easily calculated for the coded incidence matrices. For the four-symbol coding we obtain the values:
$W(c1, c4) = 0.616697; W(c1, c5) = 0.495887;$
$W(c1, c6) = 0.0954175; W(c1, c7) = 0.102132;$
$W(c2, c4) = 0.700586; W(c2, c5) = 0.728960;$
$W(c2, c6) = 0.172981; W(c2, c7) = 0.182142;$
$W(c3, c4) = 0.559629; W(c3, c5) = 0.393402;$
$W(c3, c6) = 0.151990; W(c3, c7) = 0.200251;$
$W(c1, c4, c5) = 1.57796; W(c1, c6, c7) = 0.98434;$
$W(c2, c4, c5) = 1.67118; W(c2, c6, c7) = 1.14579;$
$W(c3, c4, c5) = 1.55065; W(c3, c6, c7) = 1.02288.$

These values show that the total number of plant species, the number of endemic species, and the number of immigrant species, respectively, have a stronger interdependence with the corresponding island area and elevation and a much weaker interdependence with the distance to the centre of the archipelago and the area of the adjacent island.

If we use the quadratic measure of interdependence we obtain for the binary incidence matrix

from Table 3:
$V(c2,c4) = 0.742880; V(c2,c5) = 0.523900;$
$V(c3,c4) = 0.630861; V(c3,c5) = 0.429693,$
which show that the number of plant species in the Galápagos Islands depends more on the area than on the elevation and this dependence is larger for the endemic species than for immigrant species.

3.6 Probabilistic causality

For the binary incidence matrix from Table 3 we obtain the following values for the one-cause and two-cause probabilistic causality:
$R(c1 \mid c4) = 0.637702; R(c1 \mid c5) = 0.420846;$
$R(c1 \mid c6) = 0.000802; R(c1 \mid c7) = 0.007940;$
$R(c2 \mid c4) = 0.637702; R(c2 \mid c5) = 0.420846;$
$R(c2 \mid c6) = 0.021317; R(c2 \mid c7) = 0.000802;$
$R(c3 \mid c4) = 0.527942; R(c3 \mid c5) = 0.340000;$
$R(c3 \mid c6) = 0.007349; R(c3 \mid c7) = 0.001066;$
$R(c1 \mid c4,c5) = 0.75544; R(c1 \mid c6,c7) = 0.01833;$
$R(c2 \mid c4,c5) = 0.75544; R(c2 \mid c6,c7) = 0.02666;$
$R(c3 \mid c4,c5) = 0.58118; R(c3 \mid c6,c7) = 0.01992.$

These values show, convincingly, that the area of the corresponding island and its elevation (in this order) are the essential factors in the probabilistic relationship with the number of different types of plant species.

3.7 Species diversity

We apply here the formalism from section 2.3. In this context, the entities are the 29 islands, taken as habitats. We focus on the columns c1, c2, c3, c4, and c5 from Table 1. Each habitat has two components (component 1 is the set of endemic plant species, component 2 is the set immigrant plant species), except x_3, x_9 and x_{17} which have only component 1. Therefore, $N = 29$, $n_i = 2$, $I_i = \{1,2\}$, for i =1,2,4,. . . ,8,10,. . . ,16,18,. . . ,29, and n_i =1, $I_i = \{1\}$, for i =3,9,17, whereas I ={1,2}. The probability p_{i1} is the relative frequency of the endemic species and p_{i2} is the relative frequency of the immigrant species in the habitat x_i. These relative frequencies are obtained by dividing the values from the line i of the columns c2 and c3, respectively, with the corresponding value from column c2. Thus, $p_{11} = 23/58 = 0.3966$, and $p_{12} = 35/58 = 0.6034$, and so on, down to $p_{29,1} = 12/21 = 0.5714$, and $p_{29,2} = 9/21 = 0.4286$. When the area of the habitats is taken into account, the corresponding relative weights are obtained from column c4 of Table 1 according to the formula:

$$\lambda_i = \frac{\text{area}(x_i)}{\text{area}(x_1) + \ldots + \text{area}(x_{29})}, \quad i = 1, \ldots, 29.$$

When the elevation of the habitats is taken into account, the corresponding relative weights are obtained from column c5 of Table 1 according to the formula:

$$\lambda_i = \frac{\text{elev}(x_i)}{\text{elev}(x_1) + \ldots + \text{elev}(x_{29})}, \quad i = 1, \ldots, 29.$$

For instance, $\lambda_1 = 25.09/7851.18$, for area weighting, and $\lambda_1 = 150/10493$, for elevation weighting.

Using two decimals, the relative area weights are:

$$\begin{aligned}
&\lambda_1 = 0.00, \quad \lambda_2 = 0.00, \quad \lambda_3 = 0.00,\\
&\lambda_4 = 0.00, \quad \lambda_5 = 0.00, \quad \lambda_6 = 0.00,\\
&\lambda_7 = 0.00, \quad \lambda_8 = 0.00, \quad \lambda_9 = 0.00,\\
&\lambda_{10} = 0.01, \quad \lambda_{11} = 0.08, \quad \lambda_{12} = 0.00,\\
&\lambda_{13} = 0.00, \quad \lambda_{14} = 0.00, \quad \lambda_{15} = 0.59,\\
&\lambda_{16} = 0.02, \quad \lambda_{17} = 0.00, \quad \lambda_{18} = 0.01,\\
&\lambda_{19} = 0.00, \quad \lambda_{20} = 0.00, \quad \lambda_{21} = 0.00,\\
&\lambda_{22} = 0.07, \quad \lambda_{23} = 0.07, \quad \lambda_{24} = 0.12,\\
&\lambda_{25} = 0.00, \quad \lambda_{26} = 0.02, \quad \lambda_{27} = 0.00,\\
&\lambda_{28} = 0.00, \quad \lambda_{29} = 0.00.
\end{aligned} \tag{6}$$

Using two decimals, the relative elevation weights are:

$$\begin{aligned}
&\lambda_1 = 0.01, \quad \lambda_2 = 0.01, \quad \lambda_3 = 0.01,\\
&\lambda_4 = 0.00, \quad \lambda_5 = 0.00, \quad \lambda_6 = 0.01,\\
&\lambda_7 = 0.02, \quad \lambda_8 = 0.00, \quad \lambda_9 = 0.01,\\
&\lambda_{10} = 0.02, \quad \lambda_{11} = 0.14, \quad \lambda_{12} = 0.00,\\
&\lambda_{13} = 0.02, \quad \lambda_{14} = 0.01, \quad \lambda_{15} = 0.16,\\
&\lambda_{16} = 0.03, \quad \lambda_{17} = 0.00, \quad \lambda_{18} = 0.07,\\
&\lambda_{19} = 0.04, \quad \lambda_{20} = 0.00, \quad \lambda_{21} = 0.03,\\
&\lambda_{22} = 0.07, \quad \lambda_{23} = 0.09, \quad \lambda_{24} = 0.08,\\
&\lambda_{25} = 0.02, \quad \lambda_{26} = 0.06, \quad \lambda_{27} = 0.01,\\
&\lambda_{28} = 0.02, \quad \lambda_{29} = 0.02.
\end{aligned} \tag{7}$$

The values of the diversity index $D(x_i)$, measuring the local plant diversity on the islands x_i, (i = 1,2,. . . ,29), are:

$$\begin{aligned}
D(x_1) &= 0.95719 & D(x_2) &= 0.87409\\
D(x_3) &= 0.00000 & D(x_4) &= 0.92160\\
D(x_5) &= 1.00000 & D(x_6) &= 0.95062\\
D(x_7) &= 0.84000 & D(x_8) &= 1.00000\\
D(x_9) &= 0.00000 & D(x_{10}) &= 0.78478\\
D(x_{11}) &= 0.93884 & D(x_{12}) &= 0.82878\\
D(x_{13}) &= 0.64000 & D(x_{14}) &= 0.99750\\
D(x_{15}) &= 0.76280 & D(x_{16}) &= 0.99039\\
D(x_{17}) &= 0.00000 & D(x_{18}) &= 0.91679\\
D(x_{19}) &= 0.84877 & D(x_{20}) &= 0.75000\\
D(x_{21}) &= 0.97959 & D(x_{22}) &= 0.71301\\
D(x_{23}) &= 0.89986 & D(x_{24}) &= 0.67273
\end{aligned}$$

$$D(x_{25}) = 0.99063 \quad D(x_{26}) = 0.76213$$
$$D(x_{27}) = 0.92562 \quad D(x_{28}) = 1.00000$$
$$D(x_{29}) = 0.97959$$

Applying the formulas (1)-(5) we obtain the following numerical values:

(a) For the relative area weights (6):

$$\alpha\text{Div} = 0.780552, \quad \gamma\text{Div} = 0.792200,$$
$$\beta\text{Div} = 0.011648, \quad \text{Sim} = 0.985297.$$

(b) For the relative elevation weights (7):

$$\alpha\text{Div} = 0.825584, \quad \gamma\text{Div} = 0.927908,$$
$$\beta\text{Div} = 0.102324, \quad \text{Sim} = 0.889726.$$

(c) For the relative equal weights:

$$\lambda_i = 1/29, \qquad (i = 1, \ldots, 29),$$

we obtain the corresponding values:

$$\alpha\text{Div} = 0.790528, \quad \gamma\text{Div} = 0.999980,$$
$$\beta\text{Div} = 0.209453, \quad \text{Sim} = 0.790544.$$

A on-going debate in island biogeography focuses on whether the area or the elevation is more relevant for the species diversity found in various habitats. The answer is not an easy one. When equal weights are taken into account, then diversity is measured by taking only the species abundance into account. From the above data it is obvious that there is not much difference in the within-habitat diversity for the three kinds of relative weights taken into account. The gamma diversity, the similarity, and the between-habitat diversity, however, are more different. Compared to the case of equal weights, there is less between-habitat diversity when the elevation is taken into account and even less between-habitat diversity when the area is taken into account. Similarity between local habitats and the region containing all habitats is larger when the elevation is taken into account and very large, close to the maximum value 1, when the area is taken into account. We cannot conclude, however, that area is more relevant than elevation for the simple reason that, in this case, the range of the area and the range of the elevation are very different. The equal weights correspond to the uniform distribution, which has the Shannon entropy equal to:

$$H_{uniform} = \ln 29 = 3.36730.$$

The distribution (7) of the island elevation is not uniform, its entropy being:

$$H_{elevation} = -\sum_{i=1}^{29} \lambda_i \ln \lambda_i = 2.79883,$$

whereas the distribution (6) of the island area is even less uniform, with an entropy equal to:

$$H_{area} = -\sum_{i=1}^{29} \lambda_i \ln \lambda_i = 1.44442.$$

Indeed, according to Table 1, Isabela, the entity x_{15}, has a tremendously large area compared to all the other islands, overshadowing the contribution of the smaller islands to the diversity of the whole region when areas are taken into account. This is why, in the case (a) mentioned above, the gamma diversity is almost equal to the alpha diversity, and the similarity between the averaged local islands and the entire Archipelago is very high. There are also differences between the local elevation of the different islands of the Archipelago, but the distribution (7) of the relative weights is more uniform than the distribution (6) of the relative area weights, as shown by the corresponding values of the Shannon entropy given above.

4. CONCLUSION

In an incidence matrix where the rows are entities, the columns are characteristics and the entries are real numbers, we can do the regression analysis and calculate correlations. If, however, we want to analyze the similarity, interdependence, and probabilistic causality between characteristics or between entities, we have to coarse-grain first the range of values taken on by characteristics, classifying the entities into classes. This paper deals with coarse-graining based on keeping the right balance between uncertainty and ambiguity induced by this grouping of entities into disjoint classes. After coarse-graining, we may use the resulting coded incidence matrix in order to detect the dissimilarity (using the Hamming distance), interdependence (using both Watanabe's logarithmic measure and a new quadratic measure proposed in this paper), and probabilistic causality (using Tuldava's measure and a generalization of it) between characteristics or between entities. Also, if we have a set of entities, like the local habitats of a larger region, and a probability distribution reflecting the abundance of some components (species living in those habitats, for example) of these entities, a new weighted measure of diversity, called the Rich-Gini-Simpson quadratic index, is defined in this paper and used for calculating the average within-component diversity and the average between-component diversity in the given set of entities. This new index preserves all the properties of the classic Gini-Simpson index, which is used by many ecologists, but also behaves very well when the number of distinct species is very large, a case where the Gini-Simpson index is not suitable to be used, as shown recently by Jost ([24], [25]). An application involving the endemic and immigrant plant species from 29 Galápagos Islands is discussed, and special attention is given to the relevance of the area versus elevation in measuring the local and global plant species diversity.

References

[1] J. Tuldava, "Informational Measures of Causality", **Journal of Quantitative Linguistics**, Vol.2, 1995. pp.11-14.

[2] M.P. Johnson, P.H Raven, "Species Number and Endemism: The Galápagos Archipelago Revisited", **Science**, Vol.179, 1973, pp.893-895.

[3] E.F. Connor, D. Simberloff, "Species Number and Compositional Similarity of the Galápagos Flora and Avifauna", **Ecological Monographs**, Vol.48, 1978, pp.219-248.

[4] S. Guiasu, "Grouping Data by Using the Weighted Entropy", **Journal of Statistical Planning and Inference**, Vol.15, 1966, pp.63-69.

[5] C.E. Shannon, "A Mathematical Theory of Communication", **Bell System Technical Journal**, Vol.27, 1948, pp.379-423, 623-656.

[6] V. Tozzini, "Coarse-grained Models for Proteins", **Current Opinion in Structural Biology**, Vol.15, 2005, pp.144-150.

[7] Y. Yamamoto, R.L. Hughson, "Coarse-graining Spectral Analysis: New Method for Studying Heart Rate Variability", **Journal of Applied Physiology**, Vol.71, 1991, pp.1143-1150.

[8] Y. Fujuwara, H. Fujisaka, "Coarse-graining and Self-similarity of Price Fluctuations", **Physica A: Statistical Mechanics and Its Applications**, Vol.294, 2001, pp.439-446.

[9] R.W. Hamming, "Error Detecting and Error Correcting Codes", **Bell System Technical Journal**, Vol.29, 1950, pp.147-160.

[10] S. Watanabe, **Knowing and Guessing**, Wiley, New York: Wiley, 1969.

[11] R.C. Guiasu, S. Guiasu, **Entropy in Ecology and Ethology**, New York: Nova Science Publishers, 2003.

[12] R.C. Guiasu, S. Guiasu, "Comparison Between Some Probabilistic Indicators Useful in the Behavioural Classification", **International Journal of Mathematical and Statistical Sciences**, Vol.7, 1998, pp.39-59.

[13] C. Gini, "Variabilità e Mutabilità", **Studi Economico Giuridici della Facolta di Giurisprudenza dell'Università di Cagliari**, a.III, parte II, 1912.

[14] E.H. Simpson, "Measurement of Diversity", **Nature**, Vol.163, 1949, p.688.

[15] R.H. MacArthur, H. Recher, M. Cody, "On the Relation Between Habitat Selection and Species Diversity", **American Naturalist**, Vol.100, 1966, pp.319-332.

[16] R. Lande, "Statistics and Partitioning of Species Diversity and Similarity Among Multiple Communities", **Oikos**, Vol.76, 1996, pp.5-13.

[17] R.H. Whitaker, "Evolution and Measurement of Species Diversity", **Taxon**, Vol.21, 1972, pp.213-251.

[18] J.D. Allan, "Components of Diversity", **Oecologia**, Vol.18, 1975, pp.359-367.

[19] D.M. Porter, "Geography and Dispersal of Galápagos Islands Vascular Plants", **Nature**, Vol.264, 1976, pp.745-746.

[20] J. Kricher, **Galápagos: A Natural History**, Princeton: Princeton University Press, 2006.

[21] D.F. Andrews, A.M. Herzberg, **Data. A Collection of Problems from Many Fields for the Student and Research Worker**, New York: Springer, 1985.

[22] T.H. Hamilton, I. Rubinoff, R.H. Barth Jr., and G.L. Bush, "Species Abundance: Natural Regulation of Insular Variation", **Science**, Vol.142, 1963, pp.1575-1577.

[23] J. Fitter, D. Fitter, and D. Hosking, **Wildlife of the Galápagos**, Princeton: Princeton University Press, 2000.

[24] L. Jost, "Entropy and Diversity", **Oikos**, Vol.113, 2006, pp.363-375.

[25] L. Jost, "Partitioning diversity into independent alpha and beta components", **Ecology**, Vol.88, 2007, pp.2427-2439.

Table 1: The initial numerical incidence matrix.

	c1	c2	c3	c4	c5	c6	c7
x_1	58	23	35	25.09	150	0.6	903.82
x_2	31	21	10	1.24	109	26.3	572.33
x_3	3	3	0	0.21	114	58.7	170.92
x_4	25	9	16	0.10	46	47.4	0.18
x_5	2	1	1	0.05	10	1.9	903.82
x_6	18	11	7	0.34	119	8.0	903.82
x_7	10	7	3	2.33	168	290.2	2.85
x_8	8	4	4	0.03	10	0.4	903.82
x_9	2	2	0	0.18	112	50.2	0.10
x_{10}	97	26	71	58.27	198	88.3	0.57
x_{11}	93	35	58	634.49	1494	95.3	4669.32
x_{12}	58	17	41	0.57	49	93.1	58.27
x_{13}	5	4	1	0.78	227	62.2	0.21
x_{14}	40	19	21	17.35	76	92.2	129.49
x_{15}	347	89	258	4669.32	1707	28.1	634.49
x_{16}	51	23	28	129.49	343	85.9	59.56
x_{17}	2	2	0	0.01	25	45.9	170.92
x_{18}	104	37	67	59.56	777	119.6	129.49
x_{19}	108	33	75	17.95	458	10.7	903.82
x_{20}	12	9	3	0.23	10	0.6	903.82
x_{21}	70	30	40	4.89	367	24.4	572.33
x_{22}	280	65	215	551.62	716	66.6	0.57
x_{23}	237	81	156	572.33	906	19.8	1.24
x_{24}	444	95	349	903.82	864	0.0	0.03
x_{25}	62	28	34	24.08	259	16.5	903.82
x_{26}	285	73	212	170.92	640	49.2	0.01
x_{27}	44	16	28	1.84	100	9.6	25.09
x_{28}	16	8	8	1.24	186	50.9	4669.32
x_{29}	21	12	9	2.85	253	254.7	2.32

Table 2: The correlations between characteristics.

	c1	c2	c3	c4	c5	c6
c2	0.974					
c3	0.998	0.955				
c4	0.616	0.618	0.609			
c5	0.738	0.792	0.714	0.750		
c6	−0.186	−0.181	−0.186	−0.109	0.000	
c7	−0.139	−0.105	−0.147	0.037	0.257	−0.107

Table 3: The four-symbol coding and the binary coding.

	c1	c2	c3	c4	c5	c6	c7	c1	c2	c3	c4	c5	c6	c7
x_1	3	3	3	3	2	1	3	1	1	1	1	0	0	1
x_2	2	3	2	2	2	2	3	0	1	0	0	0	0	1
x_3	1	1	1	1	2	3	3	0	0	0	0	0	1	1
x_4	2	2	2	1	1	3	1	0	0	0	0	0	1	0
x_5	1	1	1	1	1	1	3	0	0	0	0	0	0	1
x_6	2	2	2	2	2	1	3	0	0	0	0	0	0	1
x_7	1	1	1	2	2	4	2	0	0	0	0	0	1	0
x_8	1	1	2	1	1	1	3	0	0	0	0	0	0	1
x_9	1	1	1	1	2	3	1	0	0	0	0	0	1	0
x_{10}	3	3	4	3	3	4	1	1	1	1	1	1	1	0
x_{11}	3	4	3	4	4	4	4	1	1	1	1	1	1	1
x_{12}	3	2	3	2	1	4	2	1	0	1	0	0	1	0
x_{13}	1	1	1	2	3	3	1	0	0	0	0	1	1	0
x_{14}	2	2	2	3	1	4	2	0	0	0	1	0	1	0
x_{15}	4	4	4	4	4	2	3	1	1	1	1	1	0	1
x_{16}	3	3	2	4	3	3	2	1	1	0	1	1	1	0
x_{17}	1	1	1	1	1	2	3	0	0	0	0	0	0	1
x_{18}	4	4	3	3	4	4	2	1	1	1	1	1	1	0
x_{19}	4	3	4	3	3	2	3	1	1	1	1	1	0	1
x_{20}	2	2	1	1	1	1	3	0	0	0	0	0	0	1
x_{21}	3	3	3	3	3	2	3	1	1	1	1	1	0	1
x_{22}	4	4	4	4	4	3	1	1	1	1	1	1	1	0
x_{23}	4	4	4	4	4	2	2	1	1	1	1	1	0	0
x_{24}	4	4	4	4	4	1	1	1	1	1	1	1	0	0
x_{25}	3	3	3	3	3	2	3	1	1	1	1	1	0	1
x_{26}	4	4	4	4	4	3	1	1	1	1	1	1	1	0
x_{27}	2	2	2	2	2	1	2	0	0	0	0	0	0	0
x_{28}	2	2	2	2	2	3	4	0	0	0	0	0	1	1
x_{29}	2	2	2	2	3	4	2	0	0	0	0	1	1	0

Table 4: The Hamming distance between characteristics.

	c2	c3	c4	c5	c6	c7	c2	c3	c4	c5	c6	c7
c_1	4	5	10	12	24	21	2	1	2	4	15	16
c_2		9	8	11	24	17		3	2	4	17	14
c_3			10	15	25	20			3	5	16	15
c_4				9	24	18				4	15	16
c_5					21	22					13	18
c_6						25						23