THE ESTIMATION OF BIODIVERSITY AND THE CHARACTERIZATION OF THE DYNAMICS: AN APPLICATION TO THE STUDY OF A PEST

Carlos Bouza HERRERA[1]
Leo SCHUBERT[2]

- ABSTRACT: In this paper the estimation of biodiversity is developed by deriving an estimator of an index. It is considered as a parameter and should be estimated when a simple random sample of plots is selected. Its variances is obtained. Pest dynamics is studied using a Data Mining tool. We used data generated by a large study on the behavior of a sugar cane pest, in order to evaluate the accuracy of the estimators and determine characteristics of the pest dynamics.

- KEY WORDS: Unbiased estimators; accuracy; bootstrap; sampling errors; data mining; decision tree.

1 Introduction

A common interest of ecologists is to characterize biodiversity and establish the structure of communities. Diversity can be described by the number of individuals in a population (N) , the number of species (K) and their distribution as expressed by the proportions $W_i=N_i/N$, $i=1,...,K$. Generally, the number of species N_i is unknown and W_i must be estimated. The study of biodiversity needs a single numerical measure index. Non-statistically minded persons have proposed many of them. Hence, their meaning is fuzzy and their properties out of the usual statistical thinking. Heltshe-Bitz (1979) made an empirical study on the behavior of a biodiversity index. Other authors have used Jacknife in the study of variability, for example Adams-McCune (1979) and Zahl (1977). Patil-Taillie (1982) used some general assumptions for deriving families characterized by the index structure. Beran (1994) proposed a set of properties that establishes the admissibility of an index. In the sequel a population index is considered as a parameter. In this paper we derive some point estimators. The variances of the estimators with a linear structure are determined. A comparison among the estimators cannot be developed analytically. We use data generated by a study of a pest in the western provinces of Cuba during three years. Ecologists studied the effect of the pest on sugar-cane yields. This research was developed within a project of the Centro Nacional de Sanidad Agropecuaria (National Center of Agricultural and Cattle Health) supported by the Ministerio de Agricultura (Ministry of Agriculture) during the period 1997-2001. The measurements made in 3604 plots were used for constructing an

---

[1] Departamento de Matematica Aplicada, Faculdad de Matematica y Computacion. Universidad de la Habana, CP 10.400, Habana, Cuba. E-mail:*bouza@matcom.uh.cu*

[2] Fach Hocshule Konstanz, Germany,. E-mail: *schubert@fh-konstanz.de*

artificial population. The variables were the number of eggs, larvae and beetles.

Two of the biodiversity measures (index) can be estimated unbiasedly. Their variances are determined. The structure of one of them suggests that using Bootstrap for estimation it is an adequate option. The third index has a non-linear structure. The estimation of its variance is made using Bootstrap. A Monte Carlo experiment allows the comparison of the behavior of the estimators in terms of accuracy. Samples were selected from the artificial population and the variances were estimated. Confidence intervals (CI) were constructed in each generated sample. We measured the percent of times in which the computed CI covered the value of each index.

The dynamics of this population was studied using Data Mining techniques. These methods allow to explore large quantities of data for discovering patterns and rules. Our goal was to establish behavioral regularities in the population structure. The undirected knowledge methodology was used. Decision Tree tools were utilized for the analysis. The module Answer Tree of SPSS was the software put to use.

2 Estimation of indexes

Diversity is very important in the study of population dynamics. The practical need of using it is one of the reasons for biodiversity measurement. Researchers have defined an index for measuring diversity in their data. Patil-Taillie (1982) proposed a statistical frame for modeling the diversity measurement. Hence, characterizing it statistically will help to understand its differences and properties. A biodiversity index should give an idea of the equilibrium. An index is generally treated in the literature as a descriptive measure. Patil-Taillie (1982) and Beran (1999) established this fact clearly. Therefore, inferences cannot be developed using the data obtained from random samples. In this paper we consider a population index as a parameter and develop an estimator.

Our approach is to consider the existence of at most K different species in each site. We model this fact by considering that the population is described by the partition $U = U_1+ +...+U_K$. U_j is the set of individuals of species j in the community and N_j is its size. Hence $N = N_1 +...+N_K$.

The proposals of Simpson (1949), Brillouin (1962), McIntosh (1967) and Fayer (1972) are frequently used to measure biodiversity. Pielou (1977) proposed a method for pooling a sample of m quadrant. His experiments established that the index of McIntosh (1967) was a function of Simpson's index. Heltshe-Bitz (1979) derived experimentally that the index proposed by Brillouin (1962) was insensitive to changes in the distribution.

These studies determined the tractability of an index from a statistical point of view. We studied:

2.1 The index of Fayer

The ecologist must fix a rank for every species. Denoting it by R_i

$$\lambda^*_F = [N(K+1)-J(K-J)]/2 - \sum^K_{i=1} N_i R_i = \lambda^*_0 - \sum^K_{i=1} N_i R_i$$

is the index. $J \in [0,K)$ is an integer and R_1 ,... , R_K are the ranks of the species in decreasing order of importance. For typifycation we will consider a simple transformation of this index given by

$$\lambda_F = \frac{N(K+1)-J(K-J)}{2N} - \sum_{i=1}^{K} W_i R_i = \lambda_0 - \sum_{i=1}^{K} W_i R_i \ .$$

2.2 The index of Simpson

It is:

$$\lambda_S = 1 - \Sigma^K_{i=1} N_i (N_i - 1) \ .$$

We propose using the transformation of Simpson's index given by:

$$\lambda'_S = 1 - \Sigma^K_{i=1} W_i^{\,2} \ .$$

When the number of individuals is large enough, we have that $\lambda'_S \cong \lambda_S$. Note that $\lambda'_S=1-H$, where H is Herfindall's concentration index, which is of commonly used as a descriptive measure in economy, see Bruckman (1969 and 1999). When $H=1$, the concentration of the shares is maximum and $H=1/N$ when it is minimum. H is robust for the unknowledge of the shares, the W_i's, for small units. Hence, we can argue about the robustness of λ'_S. This is the most popular index among biologists because of its easy computation and interpretation.

2.3 The index of Beran

Beran (1994) analyzed the concept of diversity and proposed desirable properties of a diversity index $d(U_1,\dots,U_K)$. When we deal with one variable, they are:

- $\mathrm{d}(U_1,\dots,U_K)\geq 0$.
- $d(U_1,\dots,U_K)$ is symmetric in W_i.
- $d(U_1,\dots,U_K)$ is continuous in W_i.
- $d(U_1,\dots,U_K)= d(U_1,\dots,U_K,U_{K+1},\dots,U_{K+\mathrm{m}})$ if $W_j>0$ for any $j>K$.
- $d(U_1,\dots,U_K)$ is minimum if $W_{\mathrm{j}}=1$ for a certain j.
- $d(U_1,\dots,U_K)$ is maximum if $W_j=1/K$ for any j.
- $d(U_1,\dots,U_K)\in [\, d(U_1,\dots,U_{K'}), d(U_1,\dots,U_{K''})]$ in another case with $K' \leq K \leq K''$.
- $d(U_1,\dots,U_K)> d(U_1,\dots,U_{K'})$ if $K>K'$ and $W_j=1/K$, $W'_j=1/K'$.
- $1\leq d(U_1,\dots,U_K)\leq K$.

Beran (1999) introduced three other properties when the index depends on two classification variables, which is not our case. He derived the index

$$\lambda_B = \Pi_{i=1}^{K}\ W_i^{\,-W_i} = \exp\,[-\Sigma_{i=1}^{K} W_i \ln(W_i\,)]$$

and proved that it is the only admissible index.

The sampling procedure that we will consider is a random sampling design implemented as follows:

Sampling Procedure

a) Select a random sample of n plots using simple random sampling.

b) Evaluate the number n_i of individuals in U_i .

The distribution of $(n_1 ,... ,n_K)$ is a multinomial. Hence, $E(n_j)=nW_j$ and $p_j= n_j / n$ unbiasedly estimates the population proportion of species j. We assume that $R_i=Rank\ (N_i) \cong Rank\ (nW_i)$ and that the relation $R_i \cong Rank\ (n_i)=r_i$. Then

$$\lambda^0{}_F=\lambda_0-\sum^K{}_{i=1}\, p_i\, r_i$$

and

$$\lambda^0{}_S=1-\sum^K{}_{i=1}\, p_i{}^2$$

are naive estimators of λ_F and λ'_S .

Note that $E(\lambda^0{}_F)\cong\lambda_F$, with variance

$$V(\lambda^0{}_F)= \sum^K{}_{i=1} R_i^2 W_i(1-W_i)/n.$$

An unbiased estimator of it is easily derived as

$$V^o(\lambda^0{}_F)= \sum^K{}_{i=1} r_i^2 p_i(1-p_i)/n.$$

The transformed Simpson's index is seriously biased because:

$$E(\lambda^0{}_S)=1-\sum^K{}_{i=1} E(p_i{}^2)=1-\sum^K{}_{i=1} W_i{}^2+W_i(1-W_i)/n.$$

A biased corrected estimator of it is $\lambda^*{}_S=\lambda^0{}_S+\sum^K{}_{i=1} v_i$, where v_i is an unbiased estimator of $W_i(1-W_i)$. Then, we propose using $\lambda^*{}_S=\lambda^0{}_S+\sum^K{}_{i=1} p_i(1-p_i)/n$. The derivation of the variance of this estimator requires cumbersome calculations because they involve moments of high order of the multinomial. It is given by :

$$V(\lambda^*{}_S)=\sum^K{}_{i=1} W_i(1-W_i)/n + 4\sum^K{}_{i=1} W_i^3 /n -\sum^K{}_{i\neq j} W_iW_j\,(W_i +(1-W_j\,))/n.$$

Its estimation leads to a very complicated formula. We recommend using Bootstrap for estimating the error.

For the index of Beran

$$\lambda_B{}^0 =\Pi_{i=1}{}^K\, p_i{}^{-p_i} = exp\ [-\sum_{i=1}{}^K p_i\, ln(p_i\,)]$$

is a naive estimator, which is consistent. Its variance has a very complicated structure because of its non-linearity. We will use a Bootstrap procedure for estimating it.

3 Study of the biodiversity

We studied biodiversity in the population of M=3,604 plots where the number of eggs, larvae and beetles were measured. The selection of the sample was made using Sampling Procedure 1. The sampling fractions were $f=m/M$=0.01, 0.05 and 0.10. The corresponding values of n_i, r_i and p_i were computed in each sample. The procedure was repeated S=100 times. In our Monte Carlo experiment N_i, R_i and W_i were known; hence the values of each index in the population were computed. The classic Bootstrap methodology was used for obtaining error estimations for the three proposed estimators and the number of bootstrap samples were $b^*=100$. The Bootstrap sub-samples were selected using simple random sampling with replacement from the original sample. We computed

$$V^*(\lambda_A^0)=\Sigma^{b^*}_{b=1}(\lambda_A^0(b)-\lambda_{A(BS)})^2/b^*$$

where $\lambda_A^0(b)$ is the estimate computed in the b^{th}-Bootstrap sample for $A=F, S, B$ and

$$\lambda_{A(BS)}=\Sigma^{b^*}_{b=1}(\lambda_A^0(b)/b^*$$

is the Bootstrap mean. The accuracy of each estimator was evaluated by

$$A(\lambda_A^0)=100\Sigma^{S}_{s=1}|\lambda_A^0(s)-\lambda_A|/S\lambda_A$$

and

$$CV(\lambda_A^0)=100[\Sigma^{S}_{s=1}V^0(\lambda_A^0(s))/S\lambda_A^2]^{1/2}.$$

The first measure serves for evaluating the absolute deviations of the estimators and the second one the relative accuracy in a sense related to the coefficient of variation. Both are relative measures.

The results of the Monte Carlo experiments are given in Table 1.

Tabela 1 - Analysis of the behavior of index estimators in the study of a sugar cane pest

Index	Value of the Index	m/M=0.01		m/M=0.05		m/M=0.10	
		CV(λ_A)	A(λ_A)	CV(λ_A)	A(λ_A)	CV(λ_A)	A(λ_A)
Fayer	0.57	27.21	12.46	15.62	11.16	10.54	7.81
Simpson	0.51	24.67	16.97	22.89	13.52	11.71	10.14
Beran	0.52	32.33	11.95	28.97	13.24	18.85	11.74

Note that the estimator of Beran's index is estimated more accurately only for the largest sampling fraction (f=0.10). Its accuracy can be considered as a direct function of f. Fayer's index has the opposite behavior in the experiment.

CI's were computed in each generated sample assuming that the normal approximation was valid. The percent of the samples that contained the true index value was computed. In any case $1-\alpha=0.95$. The CI's were determined by using the population variance, when it was possible to be computed. In the case of Fayer, we computed the CI's using Bootstrap and sample estimation of the variance. The results are given in Table 2.

Tabela 2 - Proportion of samples in which the Confidence Interval contained de Index with α=0.05

Confidence Interval	Fayer	Simpson	Beran
$\lambda_A^0 \pm 1.96\sqrt{V(\lambda_A^0)}$	0.87	0.83	-
$\lambda_A^0 \pm 1.96\sqrt{V^0(\lambda_A^0)}$	0.91	-	-
$\lambda_A^0 \pm 1.96\sqrt{V^*(\lambda_A^0)}$	0.96	0.94	0.82

The results show that Fayer index estimator has closer behavior to that expected in theory. Its best behavior was when the Bootstrap estimated variance was used. Beran's index has the worst results. It should be expected because its structure does not suggest that the usual Bootstrap approximation to the normal should be acceptable.

4 Study of the population dynamics

Data Mining (DM) is a collection of techniques that allow the analyzis of large data sets. Its development is a result of the possibilities given by modern computers. The existence of large databases in companies is largely due to the storing capabilities of their computers. They accumulate data, but their use in management does not go beyond the usual needs of controlling. The identification of the problem to be studied or its detection is the first step in DM investigations. In the analysis, statistical, artificial intelligence and data base techniques are used for 'drilling' into the data in order to obtain valuable information. The result of DM application should be the characterization of the problem and the support of an adequate decision-making process. Its increasing use in marketing research has its roots in such fact. A lot of applied scientific research provides large date bases and the problem concerning data analysis in order to determine patterns and rules is also present. Scientists identify the problem and they want to determine the patterns reflected in their observations. In fact, a measurement process originates data and the objective of applied research is to look for the establishment of the regularities of the phenomena. Hence, DM techniques may be used as a useful tool for the study of experimental data.

DM is mainly used for two groups of tasks:

a) Classifying, affinity grouping or clustering.

b) Describing, estimating or predicting.

The techniques in each group deal with similar problems. Clustering is used for determining groups by analyzing how close the units are. A certain distance or similarity measure is utilized for constructing homogenous groups. Examining the features of new units for placing them in a previously known group is the classification task. Affinity grouping is related to the need to determine why some units of different nature appear together.

Describing may be the main objective of an investigation: to obtain some knowledge of the characteristics of a phenomenon. Estimation deals with the determination of a set of values that characterizes data, while prediction establishes what may be expected in the future by analyzing the existing results.

Delineating the picture of population dynamics requires the observation of a large number of specimens. In practice it poses the need to analyze a relatively large Data Base. Ecologists were interested in developing a theoretical model that describes the behavior of the pest. Its evaluation should permit to establish what should be used for controlling pest once a field is sampled. Some of the biological controls usable for pest control are efficient because of their consume, or parasitace, eggs, larvae or beetles. The theoretical model should be reflected in the data analysis. We considered the use of Data Mining for describing the regularities exhibited by the collected data. Researchers from the Centro National de Sanidad Agrícola of Cuba obtained the data during a three-years study of a sugar-cane pest. They measured the number of eggs, larvae and beetles in the same sampled plots. The results generated 3604 valid observations.

The characteristics of the eggs-larvae-beetles relations in the same plot should permit to learn about the behavior of its structure. This knowledge should sustain the establishment of environmental policies for pest control. The dynamics was studied using the number of eggs and beetles as explanatory variables. The number of larvae was regarded as unimportant for considering it as a explanatory variable by the ecologists.

An analysis of our objectives and the characteristics of DM techniques suggested using the tool known as Answer-Tree. The trees may be constructed using different algorithms. In our case we used CHAID and C&RT because of the nature of the data. CHAID (CHi squared Automatic Interaction Detection) is due to Hartigan, see Hartigan (1985). Different improvements of it have been introduced. The algorithm looks for detecting the significance of statistical relationships among the variables using a Chi-squared test at each step. The input variables are split generating at least two new child-nodes from each parent node. The significance of the split is tested. If it is significant the split is performed.

CR&T (Classification & Regression Tree) is an algorithm which works looking for the minimization of a measure of the difference. It uses the input and builds a binary tree. At each step it fixes which independent variable is the best splitter. Then the classes are determined. Breiman *et al*. (1984) developed it within the frame of machine learning. The measure of the difference is considered an ‘impurity’ or ‘diversity’ measure. The best variable is that with the smaller value of the measure selected by the specialists from those offered by the software.

The question posed to the data, at each step, was which values of the other variables determine homogeneous groups. The final result should permit to establish which composition of the units supports saying that 'from the point of view of its population their dynamics, are similar'. Then the elements in a group should be treated similarly when an eradication campaign is planned.

The output obtained using the Answer-Tree 2.0 module, provided by SPSS 8.0, is given in Figures 1 - 4. Their analysis yielded the following results.

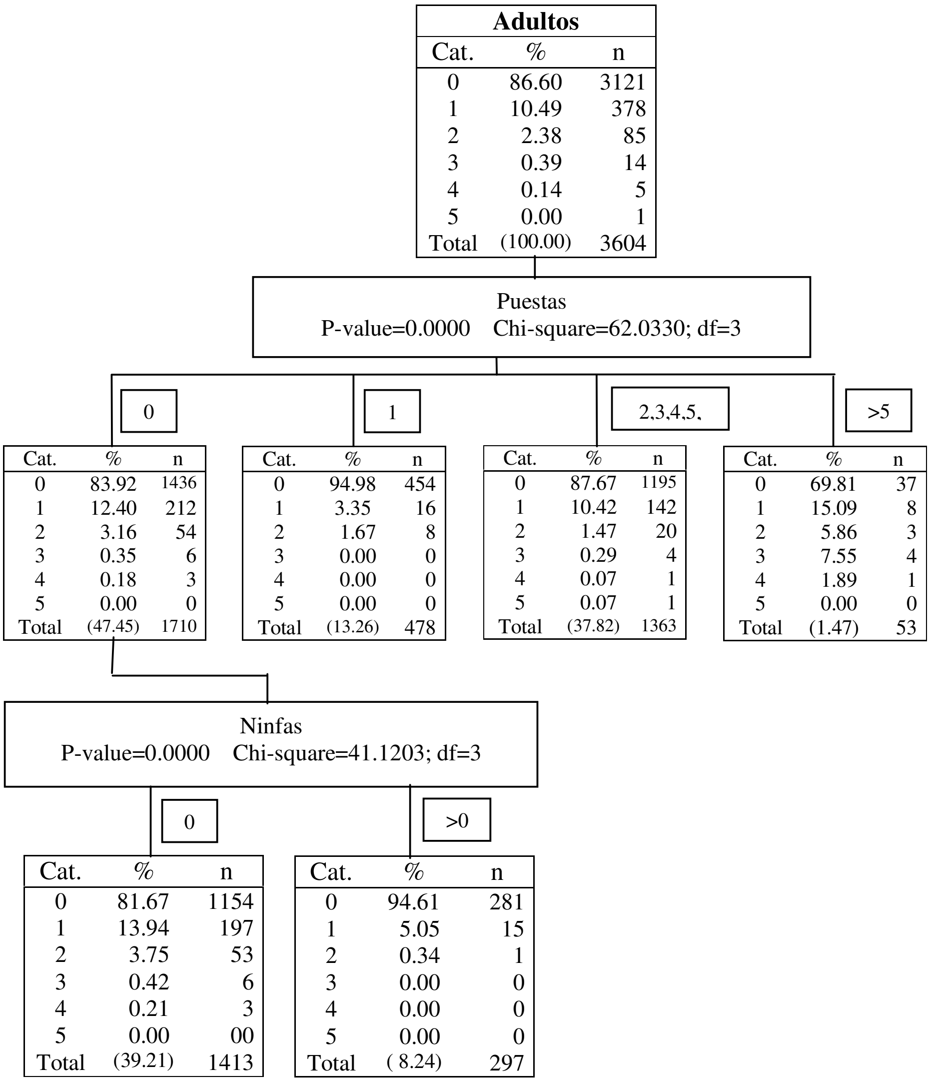
5 Explanatory variable: Beetles

5.1 Beetles computed with CHAID

Node 1 of Figure 1 gives the structure of the sample. Note that their existence has the probability $P[X_B \geq 1]=0,134$. The most important variable for the target variable (beetles) is the number of eggs. Four nodes appear in the following level. Node 1 $[X_E=0]$, node 2 $[X_E=1]$, node 3 $[X_E \leq 5]$ and node 4 $[X_E \geq 6]$. The most important nodes are 1 and 3. When the number of eggs is zero the probability of observing at least one beetle is

$P(X_B \geq 1 \mid X_E=0)=1-0.8392=0.1208$.

This node may be split by taking into account the existence of larvae. If no larvae are observed the conditional probability is $P(X_B>0 \mid X_E=X_L=0)=0.1833$. This means that 18.33% of the sample plots without eggs and larvae will not mean that the pest does not invade the plot. When larvae are observed the probability is reduced to $P(X_B \geq 0 \mid X_E=0$ and $X_L>0)=0.0539$. Hence, the probability of observing beetles is considerably large when there are no eggs, but there are larvae in the plot.



Adultos

Cat.	%	n
0	86.60	3121
1	10.49	378
2	2.38	85
3	0.39	14
4	0.14	5
5	0.00	1
Total	(100.00)	3604

Puestas
P-value=0.0000 Chi-square=62.0330; df=3

0

Cat.	%	n
0	83.92	1436
1	12.40	212
2	3.16	54
3	0.35	6
4	0.18	3
5	0.00	0
Total	(47.45)	1710

1

Cat.	%	n
0	94.98	454
1	3.35	16
2	1.67	8
3	0.00	0
4	0.00	0
5	0.00	0
Total	(13.26)	478

2,3,4,5,

Cat.	%	n
0	87.67	1195
1	10.42	142
2	1.47	20
3	0.29	4
4	0.07	1
5	0.07	1
Total	(37.82)	1363

>5

Cat.	%	n
0	69.81	37
1	15.09	8
2	5.86	3
3	7.55	4
4	1.89	1
5	0.00	0
Total	(1.47)	53

Ninfas
P-value=0.0000 Chi-square=41.1203; df=3

0

Cat.	%	n
0	81.67	1154
1	13.94	197
2	3.75	53
3	0.42	6
4	0.21	3
5	0.00	00
Total	(39.21)	1413

>0

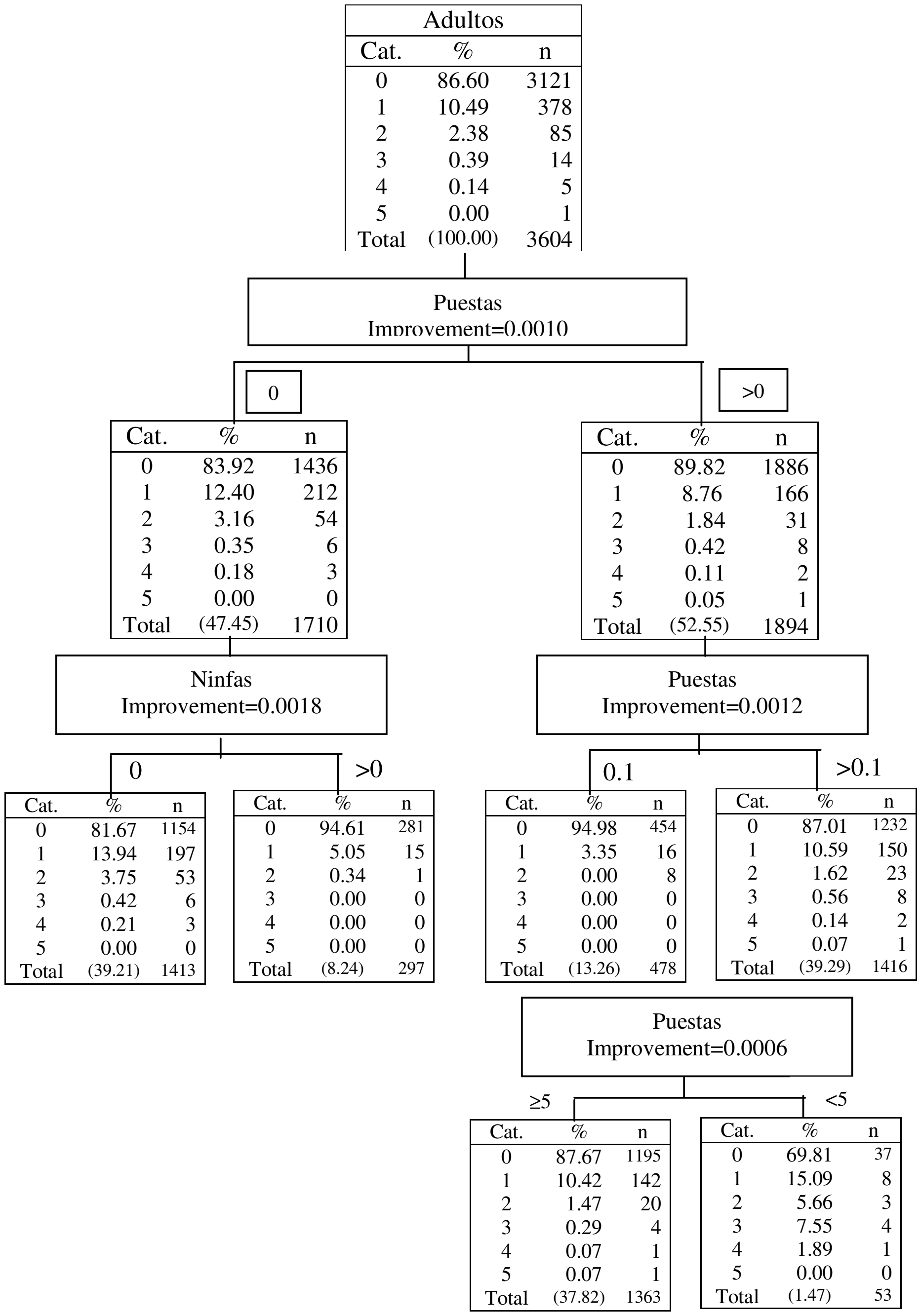
Cat.	%	n
0	94.61	281
1	5.05	15
2	0.34	1
3	0.00	0
4	0.00	0
5	0.00	0
Total	(8.24)	297

FIGURE 1 - Tree of Beetles computed with CHAID. Beetles=Adultos, Puestas=Eggs, Ninfas= Larvae.

5.2 Beetles computed with CR&T

The output obtained is given in Figure 2. The results obtained with this algorithm and with CHAID when X_E=0 are the same. The nature of the search developed by this algorithm permits to detect that the main importance of observing X_E>0 is described by node 6 because $P(X_B > 0 \mid X_E > 1) = 0.1399$.

]



Adultos

Cat.	%	n
0	86.60	3121
1	10.49	378
2	2.38	85
3	0.39	14
4	0.14	5
5	0.00	1
Total	(100.00)	3604

Cat.	%	n
0	83.92	1436
1	12.40	212
2	3.16	54
3	0.35	6
4	0.18	3
5	0.00	0
Total	(47.45)	1710

Cat.	%	n
0	89.82	1886
1	8.76	166
2	1.84	31
3	0.42	8
4	0.11	2
5	0.05	1
Total	(52.55)	1894

Cat.	%	n
0	81.67	1154
1	13.94	197
2	3.75	53
3	0.42	6
4	0.21	3
5	0.00	0
Total	(39.21)	1413

Cat.	%	n
0	94.61	281
1	5.05	15
2	0.34	1
3	0.00	0
4	0.00	0
5	0.00	0
Total	(8.24)	297

Cat.	%	n
0	94.98	454
1	3.35	16
2	0.00	8
3	0.00	0
4	0.00	0
5	0.00	0
Total	(13.26)	478

Cat.	%	n
0	87.01	1232
1	10.59	150
2	1.62	23
3	0.56	8
4	0.14	2
5	0.07	1
Total	(39.29)	1416

Cat.	%	n
0	87.67	1195
1	10.42	142
2	1.47	20
3	0.29	4
4	0.07	1
5	0.07	1
Total	(37.82)	1363

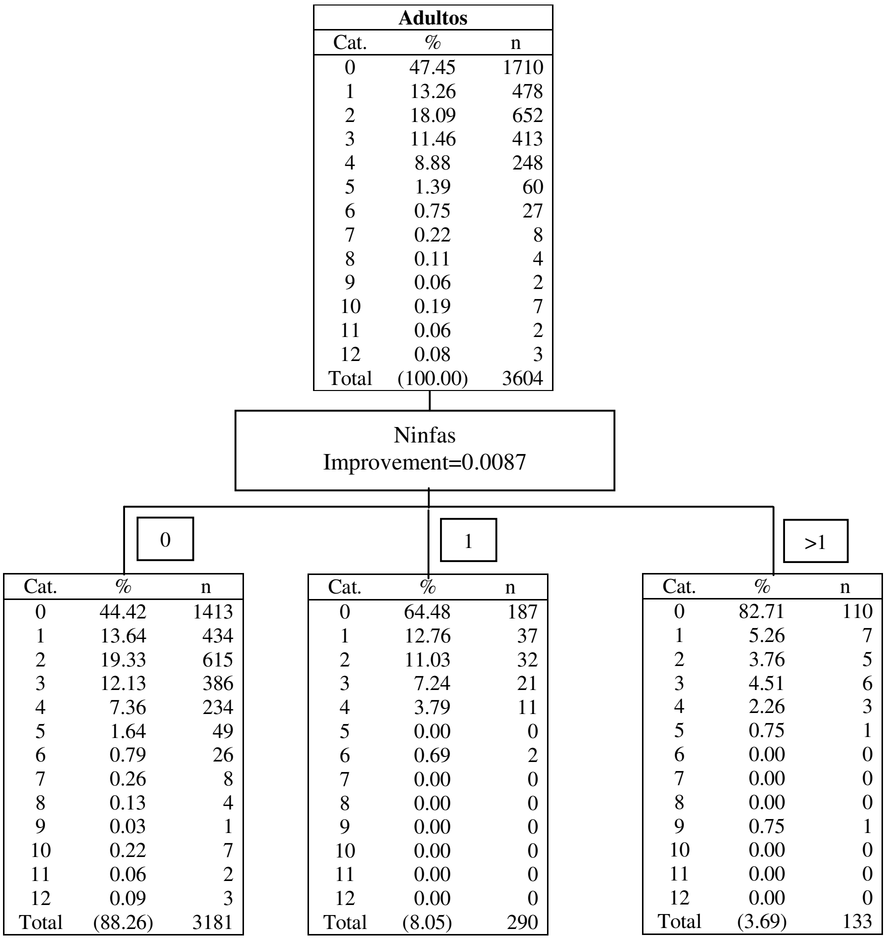
Cat.	%	n
0	69.81	37
1	15.09	8
2	5.66	3
3	7.55	4
4	1.89	1
5	0.00	0
Total	(1.47)	53

FIGURE 2 - Tree of beetles computed with C&RT. Beetles=Adultos, Puestas=Eggs, Ninfas=Larvae.

6 Exploratory variable: Eggs

6.1 Eggs computed with CHAID

The output obtained is given in Figure 3.



Adultos		
Cat.	%	n
0	47.45	1710
1	13.26	478
2	18.09	652
3	11.46	413
4	8.88	248
5	1.39	60
6	0.75	27
7	0.22	8
8	0.11	4
9	0.06	2
10	0.19	7
11	0.06	2
12	0.08	3
Total	(100.00)	3604

Cat.	%	n
0	44.42	1413
1	13.64	434
2	19.33	615
3	12.13	386
4	7.36	234
5	1.64	49
6	0.79	26
7	0.26	8
8	0.13	4
9	0.03	1
10	0.22	7
11	0.06	2
12	0.09	3
Total	(88.26)	3181

Cat.	%	n
0	64.48	187
1	12.76	37
2	11.03	32
3	7.24	21
4	3.79	11
5	0.00	0
6	0.69	2
7	0.00	0
8	0.00	0
9	0.00	0
10	0.00	0
11	0.00	0
12	0.00	0
Total	(8.05)	290

Cat.	%	n
0	82.71	110
1	5.26	7
2	3.76	5
3	4.51	6
4	2.26	3
5	0.75	1
6	0.00	0
7	0.00	0
8	0.00	0
9	0.75	1
10	0.00	0
11	0.00	0
12	0.00	0
Total	(3.69)	133

FIGURE 3 - Tree of Eggs computed with CHAID. Beetles=Adultos, Puestas=Eggs, Ninfas= Larvae.

A two-level tree is generated. The second level is larvae. The response to the increase of larvae is a decrease in the number of eggs. When X_L=0 (node 1), the structure is very similar to that observed in the whole population of eggs. Note that $P(X_E=0 \mid X_L \geq 2)=0.827$, which is almost the double of the population's probability (0.4745). The use of this algorithm permits to detect that when $X_E = X_L = 0$ splitting it into two child nodes of adults is significant.

6.2 Eggs computed with CR&T

The output obtained is given in Figure 4. A similar result is observed, but when there are not larvae, the numbers of beetles is a significant variable and two child nodes are generated. The conditional probability of observing no beetles is

$$P(X_B = 0 / X_L = 0) = 0.7564$$

and for the other child node

$$P(X_B > 0 / X_L = 0) = 0.1262.$$

Hence, the non-existence of larvae seems to be linked with the presence of beetles in more than 12% of the sampled plots. Node 4, X_B>0 , is split into nodes 7 and 8. Their analysis establishes that in the 9.85% of the plots with no larvae, only one beetle was observed. The split of node 2 into node with one larvae and another with more than one is not of real importance for describing the pest dynamics.

Though the classification of the plots was not important for the purposes of this research, we analyzed the probability of erroneous classification. The results appear in Table 3. Note that CHAID has a larger probability of correct classification than CR&T for eggs, but smaller for beetles, although the differences are not high. We reclassified all the plots.

Tabela 3 - Percent of correctly classified plots

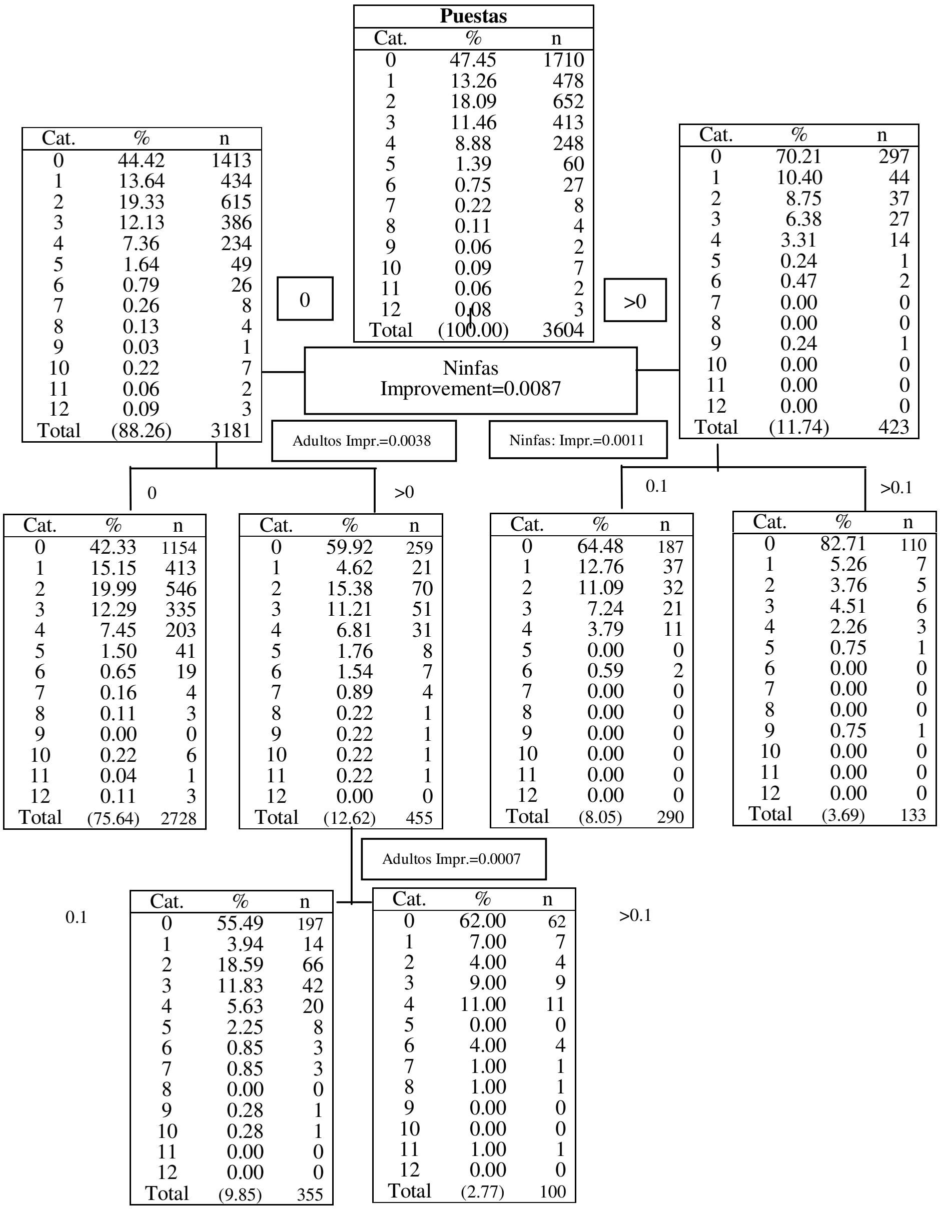
Explanatory Variable	Method CHAID	CR&T
Beetles	94.7	95.4
Eggs	96.0	93.4

The obtained results suggest a pest dynamics in the fields. Ecologists can evaluate the behavior of the infestation structure and elect an adequate control for it.

Conclusions

In Section 2 we derived unbiased estimators for λ_F and λ_S and their sampling errors when the design is simple random ample with replacement. Their comparison and the plug-in derived estimator of Beran's index was made by developing a Monte Carlo study based on an artificial population. The results suggest that:

1. The estimation of biodiversity for this pest is more accurate when the Fayer's index was used as a measure because:
   - Its point estimation seems to be more accurate,
   - The CI's computed contained the parameter in a very similar proportion to 1-α.



Puestas

Cat.	%	n
0	47.45	1710
1	13.26	478
2	18.09	652
3	11.46	413
4	8.88	248
5	1.39	60
6	0.75	27
7	0.22	8
8	0.11	4
9	0.06	2
10	0.09	7
11	0.06	2
12	0.08	3
Total	(100.00)	3604

Cat.	%	n
0	44.42	1413
1	13.64	434
2	19.33	615
3	12.13	386
4	7.36	234
5	1.64	49
6	0.79	26
7	0.26	8
8	0.13	4
9	0.03	1
10	0.22	7
11	0.06	2
12	0.09	3
Total	(88.26)	3181

Cat.	%	n
0	70.21	297
1	10.40	44
2	8.75	37
3	6.38	27
4	3.31	14
5	0.24	1
6	0.47	2
7	0.00	0
8	0.00	0
9	0.24	1
10	0.00	0
11	0.00	0
12	0.00	0
Total	(11.74)	423

Cat.	%	n
0	42.33	1154
1	15.15	413
2	19.99	546
3	12.29	335
4	7.45	203
5	1.50	41
6	0.65	19
7	0.16	4
8	0.11	3
9	0.00	0
10	0.22	6
11	0.04	1
12	0.11	3
Total	(75.64)	2728

Cat.	%	n
0	59.92	259
1	4.62	21
2	15.38	70
3	11.21	51
4	6.81	31
5	1.76	8
6	1.54	7
7	0.89	4
8	0.22	1
9	0.22	1
10	0.22	1
11	0.22	1
12	0.00	0
Total	(12.62)	455

Cat.	%	n
0	64.48	187
1	12.76	37
2	11.09	32
3	7.24	21
4	3.79	11
5	0.00	0
6	0.59	2
7	0.00	0
8	0.00	0
9	0.00	0
10	0.00	0
11	0.00	0
12	0.00	0
Total	(8.05)	290

Cat.	%	n
0	82.71	110
1	5.26	7
2	3.76	5
3	4.51	6
4	2.26	3
5	0.75	1
6	0.00	0
7	0.00	0
8	0.00	0
9	0.75	1
10	0.00	0
11	0.00	0
12	0.00	0
Total	(3.69)	133

Cat.	%	n
0	55.49	197
1	3.94	14
2	18.59	66
3	11.83	42
4	5.63	20
5	2.25	8
6	0.85	3
7	0.85	3
8	0.00	0
9	0.28	1
10	0.28	1
11	0.00	0
12	0.00	0
Total	(9.85)	355

Cat.	%	n
0	62.00	62
1	7.00	7
2	4.00	4
3	9.00	9
4	11.00	11
5	0.00	0
6	4.00	4
7	1.00	1
8	1.00	1
9	0.00	0
10	0.00	0
11	1.00	1
12	0.00	0
Total	(2.77)	100

FIGURE 4 - Tree of eggs computed with C&RT. Beetles=Adultos, Puestas=Eggs, Ninfas= Larvae.

2. The CI's based on Bootstrap variance cover the true index value in a larger proportion than the other alternatives.
3. The index of Simpson is the second best alternative.

The study of the Answer Tree suggested that CHAID should be recommended for analyzing the pest dynamics because its associated classification error probability is smaller than Cr&T's.

Acknowledgments: This paper was partly prepared during a visit by one of the authors to Germany supported by a DAAD's fellowship. This version has been improved thanks to helpful comments by two referees.

HERRERA, C.B.; SCHUBERT, L. A estimação da biodiversidade e a caracterização de dinâmicas: uma aplicação ao estudo de uma doença. . *Rev. Mat. Estat.,* São Paulo, v.21, n.3, p. 85-98, 2003.

- RESUMO: Neste artigo a estimação da biodiversidade é feita pela obtenção do estimador de um índice, que é considerado como um parâmetro e que pode ser estimado quando se tem uma amostra aleatória simples. A variância deste índice é obtida. A dinâmica da doença é estudada usando um procedimento *Data Mining*. Para avaliar a precisão dos estimadores e descrever as características da doença, foram utilizados dados gerados em um estudo do comportamento de uma doença da cana-de-açúcar.

- PALAVRAS CHAVES: Estimador imparcial; precisão; *bootstrap*; erros aleatórios; *data mining*; árvore de decisão.

References

ADAMS, J.E.; MCCUNE, E.D. Application of the generalized Jacknife to Shannon's measure of information used as an index of diversity. In GRASSIE, J.F. *et al*., (Ed.). *Ecological diversity in theory and practice.* Fairland: Maryland International Co-operative House, 1979. p.150-165.

BERAN, H. Eine system konforme maβzahlbildung für klassifikatorishe mermale. *Oper. Res. Spektrum*, Heildelberg, v.16, p.81-88, 1994.

BERAN, H. A generalization and optimization of a measure of concentration. In: LEOPOLD-WILDBURGER, U. *et al.* (Ed.). *Modeling and decisions in economics.* Heidelberg: Physika Verlag, 1999. p.119-138.

BREIMAN, L. *et al. Classification and regression trees*. Belmont: Wadsworth, 1984. p.269.

BRILLOUIN, L. *Science and information*. New York: Academic Press, 1962. p.265

BRUCKMAN , G. Einige bemerkungen zür statistischen messung der konzentration. *Metrika,* New York, v.14, p.183-213, 1969.

BRUCKMAN, G. A note on Herfindahl's index of concentration. In LEOPOLD-WILDBURGER. U. *et al*.(Ed.). *Modeling and decisions in economics.* Heidelberg: Physika Verlag, 1999. p.115-118.

FAYER, E.W. Diversity: a sampling study. *Am. Nat.*, Chicago, v.106, p.293-310, 1972.

HARTIGAN, J. A. Statistical theory in clustering . *J. Classif.*, New York, v.2. p.63-76, 1985.

HELTSHE, J. F.; BITZ, D.W. Comparing diversity measures in sampled communities. In: GRASSIE *et al*. (Ed.). *Ecological diversity in theory and practice*. Fairland: Maryland International Co-operative House, 1979. p.150-165.

MCINTOSH, R.P. An index of diversity and the relation to certain concepts of diversity. *Ecology*, Washington, v.48, p.392-404, 1967.

PATIL, G.P; TAILLIE, D.C. Diversity as a concept and its measurement. *J. Am. Stat. Assoc.*, Washington, v.77, p.548-567, 1982.

PIELOU, E.C. *Mathematical ecology*. New York: Wiley, 1977. p.175.

SIMPSON, E.M. Measurement of diversity. *Nature,* London, v.163, p.688, 1949.

ZAHL, S. Jacknifing an index of diversity. *Ecology*. Durham, v.58, p.907-913, 1977.

Received 28.02.2002

Received in revised form 03.10.2003