Vol. 216: 309–311, 2001 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 6

NOTE

Estimating the taxonomic composition of a sample when individuals are classified with error

Andrew Solow*, Cabell Davis, Qiao Hu

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

ABSTRACT: When a sample of individual organisms is classified taxonomically with the possibility of classification error, the taxonomic counts of the classified individuals are biased estimates of the true counts in the sample. This note describes a simple method for correcting for this bias based on the classification probabilities of the classifier. The method is illustrated using some data from the Video Plankton Recorder.

KEY WORDS: Multinomial distribution · Sample composition · Video Plankton Recorder

This note is concerned with the following common situation in biological oceanography. Interest centers on the taxonomic composition of a community or other group of organisms. A number of individuals are sampled from the community and sorted or classified. Increasingly, this kind of classification is done automatically. For example, in the illustration below, individuals were sampled optically by the Video Plankton Recorder (Davis et al. 1992), the images were processed by an image analyzer, and the processed images were classified by a neural net classifier. However, the problem considered here is not tied to any particular technology. In a typical situation, the composition of the sample is estimated by the composition of the classified individuals. However, unless the classifier is perfect, this estimate is biased. For example, suppose that a sample contains 90 individuals of one taxon and 10 individuals of another and that, for both taxa, the probability of correct classification is 0.9. The expected number of individuals classified as the first taxon is $(90 \times 0.9) + (10 \times 0.1) = 82$ and the expected number classified as the second taxon is $(10 \times 0.9) + (90 \times 0.1) = 18$. Despite the 90% accuracy of the classifier, the abundance of the rare taxon in the sample is, on average, overestimated by a factor of nearly 2. This is a simple case, involving only 2 taxa. As illustrated below, the bias in estimating the taxonomic composition of a sample containing several taxa due to misclassification error can be quite complicated. The purpose of this note is to describe a simple method for correcting this bias.

Method. The problem can be formulated in the following way. Let $\boldsymbol{n} = (n_1, n_2 \ldots, n_s)^t$ be the unknown vector of true taxonomic counts in a sample where n_j is the number of individuals of taxon j (j = 1, 2, …, s) and where here and below the superscript t denotes the transpose of a vector or matrix. Although the elements of $\boldsymbol{n}$ are unknown, their total $N = \sum_{j=1}^{s} n_j$ is known. The individuals are classified by a classifier. Let p_{jk} be the probability that an individual of taxon j is classified as taxon k. These probabilities pertain solely to the classifier and, in general, have no connection to the relative abundances of the taxa in the sample or in nature. In this note, we will assume that these probabilities are known to high precision (e.g., through applying the classifier to a large training sample for which the true classifications are known). The alternative case is discussed briefly at the end of the note. Let $\boldsymbol{P} = [p_{jk}]$ be the s-by-s matrix of these probabilities.

Let the random variable M_j be the number of individuals in the sample classified as taxon j. This random variable can be written as

$$M_j = \sum_{i=1}^{s} M_{ij} \quad (1)$$

where the random variable M_{ij} is the number of individuals of taxon i classified as taxon j. The expected value of M_{ij} is equal to $n_i p_{ij}$, from which it follows that the expected value of M_j is $E(M_j) = \sum_{i=1}^{s} n_i p_{ij}$.

In matrix notation

$$E(\boldsymbol{M}) = \boldsymbol{P}^t n \quad (2)$$

*E-mail: asolow@whoi.edu

where $\boldsymbol{M} = (M_1, M_2 \ldots, M_s)^t$ and E denotes expectation. It follows that an unbiased estimate of $\boldsymbol{n}$ is

$$\hat{\boldsymbol{n}} = \left(\boldsymbol{P}^t\right)^{-1}\boldsymbol{m} \quad (3)$$

where $\boldsymbol{m} = (m_1, m_2 \ldots, m_s)^t$ is the observed value of $\boldsymbol{M}$.

In addition to providing an unbiased estimate of the taxonomic counts, it is useful to have some idea of the precision of the estimate. The standard errors of the estimated taxonomic counts are given by the square roots of the diagonal elements of the variance matrix of $\hat{\boldsymbol{n}}$. For convenience, let $\boldsymbol{Q} = (\boldsymbol{P}^t)^{-1}$so that $\hat{\boldsymbol{n}} = \boldsymbol{Q}\boldsymbol{m}$. The s-by-s variance matrix of $\hat{\boldsymbol{n}}$ is given by

$$\mathrm{Var}(\hat{\boldsymbol{n}}) = \boldsymbol{Q}\,\mathrm{Var}(\boldsymbol{M})\boldsymbol{Q}^t \quad (4)$$

where Var($\boldsymbol{M}$) is the s-by-s variance matrix of $\boldsymbol{M}$. To find Var($\boldsymbol{M}$), note that $M_{i1}, M_{i2}, \ldots M_{is}$ have a multinomial distribution with n_i trials and probabilities $p_{i1}, p_{i2}, \ldots p_{is}$. The multinomial distribution is the extension of the familiar binomial distribution to the case where more than 2 outcomes are possible. The variance of M_{ij} is $n_i\, p_{ij}\, (1 - p_{ij})$ and the covariance between M_{ij} and M_{ik} is $-n_i\, p_{ij}\, p_{ik}$ provided $j \neq k$. Moreover, M_{ij} and $M_{i'k}$ are independent for all j and k provided $i \neq i'$. It follows from Eq. (1) and these results that the elements of Var($\boldsymbol{M}$) are given by

$$\mathrm{Var}(M_j) = \sum_{i=1}^{s} n_i p_{ij}(1 - p_{ij})$$

$$\mathrm{Cov}(M_j, M_k) = -\sum_{i=1}^{s} n_i P_{ij} P_{ik} \quad (5)$$

An approximate standard error for the estimate of n_j is found by substituting $\hat{\boldsymbol{n}}_i$ for n_j in Eq. (5) to form an estimate of Var($\boldsymbol{M}$), substituting this estimate for Var($\boldsymbol{M}$) in Eq. (4), and taking the square root of the corresponding term on the diagonal.

Illustration. As an illustration, we analyzed data from a 24 h deployment of the Video Plankton Recorder. Details not relevant to the present problem will be given elsewhere. The data consisted of a total of 18 902 images of individual organisms. Each image was classified to 1 of 7 taxonomic groups by visual inspection (which we took to be free of error) and also by a preliminary neural net classifier based on an independent sample. We treated the 8533 images from odd-numbered hours as the training set for estimating the classification probabilities. The resulting estimate of the matrix $\boldsymbol{P}$ was:

$$\begin{pmatrix}
0.710 & 0.059 & 0.010 & 0.010 & 0.007 & 0.031 & 0.175 \\
0.073 & 0.873 & 0.001 & 0.007 & 0.008 & 0.013 & 0.024 \\
0.078 & 0.012 & 0.556 & 0.035 & 0.066 & 0.179 & 0.074 \\
0.030 & 0.028 & 0.054 & 0.560 & 0.019 & 0.177 & 0.132 \\
0.205 & 0.054 & 0.107 & 0.046 & 0.366 & 0.157 & 0.065 \\
0.158 & 0.025 & 0.076 & 0.064 & 0.175 & 0.449 & 0.054 \\
0.289 & 0.096 & 0.033 & 0.065 & 0.018 & 0.072 & 0.427
\end{pmatrix}$$

For example, of 2727 individuals in the training set that were visually classified as taxon 1, 1935 (71%) were also classified as taxon 1 by the neural net classifier. We treated the N = 10 369 images from the even-numbered hours as the field sample. The classified counts for these images were

$\boldsymbol{m}$ = (2891 1965 495 1399 676 1191 1752)^t

We applied the correction method outlined above to these classified counts and found

$\hat{\boldsymbol{n}}$ = (2431 1696 194 1994 1119 846 2088)^t

The true taxonomic counts in this sample were

$\boldsymbol{n}$ = (2560 1759 210 1679 1003 941 2217)^t

With the exception of taxon 4, the corrected counts were closer—sometimes dramatically—to the true counts than were the classified counts. The standard errors of the elements of $\hat{\boldsymbol{n}}$ found by the procedure outlined above were around 93, 32, 43, 58, 90, 101, and 123. Again, with the exception of taxon 4, the true counts were within 2 standard errors of the estimated counts.

Discussion. It is important to emphasize that the problem considered in this note concerns the estimation of the composition of a sample. The more interesting problem of how the composition of the sample is used to estimate the composition of the field population is a separate one. Although the method described here was developed for use in conjunction with an automatic classifier, it can be applied to any kind of classification scheme including rapid visual sorting.

The analysis presented here assumes only that good estimates of the classification probabilities in $\boldsymbol{P}$ are available. In almost all cases, these estimates will be found by classifying a training sample of individuals for which the correct classifications are known. This training sample can be the same sample used to construct the classifier, provided that the elements of $\boldsymbol{P}$ are estimated by cross-validation. When $\boldsymbol{P}$ is estimated from a training sample, the quality of the estimate depends on the size (and composition) of the training sample and is therefore under the analyst's control. For the illustration given in the previous section, the large size of the training sample ensures that the estimate of $\boldsymbol{P}$ is good. For example, the standard errors of the elements of the estimate of $\boldsymbol{P}$ are all less than 0.03 and most are much smaller.

When the training sample is smaller, the estimate of $\boldsymbol{P}$ has larger variability, and the estimate $\hat{\boldsymbol{n}}$ may be biased. Limited experience suggests that this bias is negligible provided the training sample contains at least 100 to 200 individuals of each taxon. For the data set described in the previous section, this amounts to a training sample of around 1000 individuals. Variability in the estimate of $\boldsymbol{P}$ has a larger effect on the standard errors of the elements of $\hat{\boldsymbol{n}}$. Specifically, standard errors constructed as outlined above will be too narrow. It is

straightforward, although computationally demanding, to account for this additional variability through a bootstrap procedure (Efron & Tibshirani 1993). The bootstrap can also be used to correct for bias. This extension will be pursued elsewhere.

The main cost of the method described here is the construction of a training sample of a size sufficient to estimate ***P*** with reasonable accuracy. An alternative is to reduce the classification error rates by improving the classifier itself. Of course, the construction of a classifier also requires a training sample, and, as noted above, the same training sample can also be used to estimate ***P***. Thus, the choice between these alternatives will depend on the scope for improved classification and on the relative data requirements for improved classification and for estimating ***P***.

Editorial responsibility: Kenneth Sherman (Contributing Editor), Narragansett, Rhode Island, USA

LITERATURE CITED

Davis CS, Gallager SM, Berman MS, Haury LR, Strickler JR (1992) The Video Plankton Recorder (VPR): design and initial results. Arch Hydrobiol Beih 36:67–81

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London

Submitted: November 1, 1999; Accepted: March 16, 2001
Proofs received from author(s): June 8, 2001