Bootstrap variance of diversity and differentiation estimators in a subdivided population

R. J. PETIT* & O. PONS†

Institut National de la Recherche Agronomique, Laboratoire de Génétique et Amélioration des Arbres Forestiers, B.P.45, F-33611 Gazinet cedex, France and †Institut National de la Recherche Agronomique, Laboratoire de Biométrie, F-78352 Jouy-en-Josas cedex, France

We have recently proposed new estimators of the parameters of genetic diversity and differentiation and of their variances for a haploid locus in a population subdivided into a large number of subpopulations, with a two-stage sampling of populations and individuals. Here they are compared with bootstrap estimators. Several resampling methods are evaluated: sampling of populations only, individuals within populations only, or both. Theoretical results and a numerical example show that the most appropriate bootstrap variance estimators are obtained by resampling the populations alone and not both populations and individuals. However, some bias is apparent in the bootstrap methods, and the direct estimators proposed previously should therefore be preferred.

Keywords: bootstrap, genetic differentiation, genetic diversity, G_{ST}, resampling, variance.

Introduction

Resampling techniques are becoming widely used to assess confidence in phylogenetic reconstructions as well as in population genetics (Crowley, 1992). Indeed, direct analytical derivations of the appropriate variances can be extremely complex, and resampling techniques then provide rapid assessments of the precision of the studied statistics.

In particular, a variety of resampling methods have been used to detect genetic differentiation among populations (Crowley, 1992). To test whether there is a significant genetic structure, permutation procedures can be used where individuals are shuffled at random among populations, while keeping the sample sizes the same as in the original analysis (Palumbi & Wilson, 1990; Excoffier *et al.*, 1992; Hudson *et al.*, 1992). These methods do not replace the need to evaluate the precision of the measures of differentiation, such as F_{ST} (Wright, 1951; Weir & Cockerham, 1984) or G_{ST} (Nei, 1973; Pons & Petit, 1995). For multilocus isozyme data, a confidence interval can be obtained by jackknifing or bootstrapping over loci, as suggested by Weir (1990). But Van Dongen (1995) recently concluded that in general, resampling over individuals should be

preferred to resampling over loci, because the allele frequencies for the different loci are usually estimated from the same individuals and are therefore not independent. Resampling over individuals also has the advantage that the precision of the differentiation at each locus can be estimated, which allows for the possibility of detecting aberrant loci when several are available (McDonald, 1994) or of studying single-locus data, such as data based on mitochondrial or chloroplast DNA.

However, for estimators of diversity whose precision is affected by both sources of variation (sampling of individuals and of populations), it is unclear what sampling to use for the bootstrap. Several types of resampling have been used so far in the literature: for instance, Petit *et al.* (1993) used the two-stage bootstrap which mimics the original sampling, whereas Prout & Barker (1993) used a bootstrap with only the populations as units of resampling.

To decide between these alternative types of sampling, a satisfactory solution would be to compare the simulated bootstrap estimates with the analytically derived direct estimates. This seems important because, as outlined by Crowley (1992, p. 431), '*bootstrapping may have been swept into the mainstream of ecological and particularly evolutionary research somewhat ahead of a full, balanced evalu-*

*Correspondence. E-mail: remy@pierroton.inra.fr

ation of its capabilities, and shortcomings'. This analytical treatment, which considers both the sampling of individuals and the sampling of populations, is now available for a haploid and a diploid locus as well as for ordered alleles (Pons & Chaouche, 1995; Pons & Petit, 1995, 1996). The usual but sometimes implicit assumption is made that the observed populations are independent, which is both a genetic and a sampling assumption (for further discussion on this topic, see Nei, 1986; Pons & Petit, 1996). The same assumption is required in the uniform resampling methods, such as the bootstrap discussed here. We also assume that the number of sampled populations is large but nevertheless much smaller than the total number of existing populations.

We will use these results to illustrate that, in complex situations, it is necessary to ascertain whether bootstrap simulations yield the required estimates or not. Although two-stage bootstrap methods have been studied in the case of a finite number of finite populations drawn with replacement (Rao & Wu, 1988; Sitter, 1992), no results are known for a large number of populations and data sampled without replacement (in the original data sampling scheme). Here, three situations will be considered: sampling of individuals only, of populations only, and of both. We will present the exact or approximate bootstrap estimators and will compare them to the direct estimators obtained analytically in Pons & Petit (1995) and which will be referred to simply as the 'direct estimators'.

These results will be illustrated using data on isozyme polymorphism in sessile oak (Zanetto & Kremer, 1995). The comparison of the bootstrap estimators with the corresponding bootstrap simulated values will provide an evaluation of the approximations we used in the analytical derivations of the variances.

Bootstrap estimation

We consider a total population subdivided into a large number of independent populations and in which I alleles are segregating at a haploid locus. For this situation, diversity and differentiation parameters were defined in Pons & Petit (1995). In particular, the diversity h_k of the kth population is given in their *eqn 1*, the average within-population diversity h_S in *eqn 2*, the total diversity h_T in *eqn 3* and the differentiation parameter G_{ST} in *eqn 4*. A two-stage random sampling is used to estimate these parameters: n independent populations are drawn with the same probability from the general popula-

tion, then n_k individuals are drawn independently and uniformly from the kth population. Within the kth population, the proportion x_{ki} of individuals having the ith allele is observed, corresponding to an unknown frequency p_{ki}. Estimators of the parameters are defined in Pons & Petit (1995) as $\hat{h}_k$, $\hat{h}_S$, $\hat{h}_T$ and $\hat{G}_{ST}$ by their *eqns 5, 6, 9 and 10*.

Here, we study three different bootstrap sampling procedures for estimating the variance of $\hat{h}_S$, $\hat{h}_T$ and $\hat{G}_{ST}$. We also study the within- and between-population components of these variances, which will be useful for understanding what is estimated under each resampling scheme. The first bootstrap method is a resampling of the individuals in the observed populations: in the kth population, n_k individuals are drawn uniformly and with replacement from the initial sample of the kth population. In the second resampling procedure, only the populations are drawn: a bootstrap sample is obtained by drawing uniformly n populations with replacement from the observed set of populations, then by taking the initial observed values for each sampled population. The third bootstrap procedure corresponds to a two-stage bootstrap resampling: n populations are sampled uniformly and with replacement from the observed set of populations and, whenever the kth population is selected, n_k individuals are drawn uniformly and with replacement from its initial sample.

For the vth bootstrap resampling procedure ($v = 1, 2$ or 3) and for a parameter θ, which will here be h_S, h_T or G_{ST}, the bootstrap estimator $\hat{\theta}^*_{(v)}$ of θ is the mean, under the bootstrap resampling distribution and conditionally on the observed variables, of the variable $\theta^*_{(v)}$ defined in the same way as the direct estimator $\hat{\theta}$ but for the vth bootstrap variable (Efron & Tibshirani, 1993). The bootstrap estimator $\widehat{\text{Var}}^*_{(v)}(\hat{\theta})$ of the variance of $\hat{\theta}$ is the variance of $\theta^*_{(v)}$, under the vth bootstrap resampling distribution and conditionally on the observed variables. Because the resampling distributions are multinomials with parameters depending on the observed variables, the bootstrap estimators $\hat{\theta}^*_{(v)}$ and $\widehat{\text{Var}}^*_{(v)}(\hat{\theta})$ are functions of the proportions of individuals having each allele in the different populations, when bootstrapping within the populations, and of the sample sizes n_k and n. Here, we give the expressions of these bootstrap estimators without proofs, which may be obtained from the second author (Pons, 1997). We will need the following biased estimators of h_S and h_T, where $x_{\bullet i} = n^{-1} \sum_k x_{ki}$:

$$\bar{h}_S = n^{-1} \sum_k \left(1 - \sum_i x_{ki}^2\right), \quad \bar{h}_T = 1 - \sum_i x_{\bullet i}^2. \quad (1)$$

According to each multinomial bootstrap distribution, we get for h_S the three bootstrap estimators:

$$\hat{h}^*_{S,(1)} = \hat{h}^*_{S,(3)} = \tilde{h}_S, \quad \hat{h}^*_{S,(2)} = \hat{h}_S. \tag{2}$$

For h_T, we have

$$\hat{h}^*_{T,(1)} = \hat{h}_T, \quad \hat{h}^*_{T,(2)} = \hat{h}^*_{T,(3)} = \tilde{h}_T. \tag{3}$$

The three bootstrap estimators of G_{ST} are therefore biased. If the number n of sampled populations is large, as is recommended to reduce the total variance of the estimates (Pons & Petit, 1995), $\hat{h}_S \approx \tilde{h}_S$ and $\hat{h}_T \approx \tilde{h}_T$, and the three procedures give similar results.

Before considering the bootstrap estimation of the variances, we define E_k and Var_k as the expectation and the variance under the multinomial distribution of parameters n_k and p_{ki} in the kth population. For h_S, the three bootstrap variances are

$$\widehat{Var}^*_{(1)}(\hat{h}_S) = n^{-2} \sum_k \widehat{Var}_k(\hat{h}_k), \tag{4}$$

$$\widehat{Var}^*_{(2)}(\hat{h}_S) = n^{-2} \sum_k (\hat{h}_k - \hat{h}_S)^2, \tag{5}$$

$$\widehat{Var}^*_{(3)}(\hat{h}_S) = n^{-2} \sum_k (\hat{E}_k(\hat{h}^2_k) - \hat{h}^2_S). \tag{6}$$

In eqn (4), $\widehat{Var}_k(\hat{h}_k)$ is the estimated variance of $\hat{h}_k$, within the kth population. It has the same form as $Var_k(\hat{h}_k)$ given by Pons & Petit (1995) but with x_{ki} instead of p_{ki}. This is then a biased estimator of the within-population variance of $\hat{h}_S$. Comparing eqn (5) to *eqn 12* in Pons & Petit (1995), it appears that the second bootstrap variance is an estimator of the total variance of $\hat{h}_S$ instead of an estimator of the between-population variance as could have been expected when populations alone are drawn. In eqn (6), $\hat{E}_k(\hat{h}^2_k)$ is an estimator of $E_k(\hat{h}^2_k) = Var_k(\hat{h}_k) + h^2_k$ obtained by replacing p_{ki} with x_{ki}, i.e. $\hat{E}_k(\hat{h}^2_k) = \widehat{Var}_k(\hat{h}_k) + \tilde{h}^2_k$. This bootstrap variance is therefore a biased estimator of the sum of the within-population and total variances of $\hat{h}_S$.

Closed forms of the bootstrap variance of $\hat{h}_T$ are more complicated and we use the same approximations as in Pons & Petit (1995) for large n. The three bootstrap variances of $\hat{h}_T$ are then approximated, up to the order n^{-2}, as

$$\widehat{Var}^*_{(1)}(\hat{h}_T) \approx 4n^{-2} \left\{ \sum_i x^2_{\cdot i} \sum_k n_k^{-1} x_{ki} - \sum_{ij} x_{\cdot i} x_{\cdot j} \sum_k n_k^{-1} x_{ki} x_{kj} \right\}, \tag{7}$$

$$\widehat{Var}^*_{(2)}(\hat{h}_T) \approx 4n^{-2} \sum_{ij} x_{\cdot ij} x_{\cdot j} \sum_k (x_{ki} - x_{\cdot i})(x_{kj} - x_{\cdot j}), \tag{8}$$

and

$$\widehat{Var}^*_{(3)}(\hat{h}_T) \approx 4n^{-2} \sum_{ij} x_{\cdot i} x_{\cdot j} \times$$
$$\left\{ \sum_k (x_{ki} - x_{\cdot i})(x_{kj} - x_{\cdot j}) - \sum_k n_k^{-1} x_{ki} x_{kj} \right\}$$
$$+ 4n^{-2} \sum_i x^2_{\cdot i} \sum_k n_k^{-1} x_{ki}. \tag{9}$$

For the first bootstrap variance of $\hat{h}_T$, the right-hand side of eqn (7) is an estimator of $Var_{intra}(\hat{h}_T)$, the within-population variance of $\hat{h}_T$ given by *eqn 14* in Pons & Petit (1995), but $n_k - 1$ in $Var_{intra}(\hat{h}_T)$ is now replaced by n_k. Thus, it is biased and it may differ substantially from the direct estimator for small population sample sizes. If the sample sizes of the bootstrap populations are modified and set to $n_k - 1$ for the kth population, the bootstrap estimator of h_T is still $\hat{h}_T$ and its bootstrap variance estimates the within-population variance of $\hat{h}_T$. By a comparison of our direct estimators of $Var(\hat{h}_T)$, we can see that the right-hand side of eqn (8) is an estimator of the total variance of $\hat{h}_T$. This is therefore also the case for the second bootstrap variance estimator. Finally, by the third procedure the estimated bootstrap variance is approximately the sum of the total and within-population bootstrap variances of $\hat{h}_T$.

We get similar results for the bootstrap estimators of the covariance between $\hat{h}_S$ and $\hat{h}_T$, $\widehat{Cov}^*_{(v)}(\hat{h}_S, \hat{h}_T)$. Thus, the bootstrap procedures provide estimators of the variance matrix of $(\hat{h}_S, \hat{h}_T)$ when sampling the populations alone, and of its within-population variance when sampling only the individuals. Such results also hold for $\hat{G}_{ST}$, and in particular the bootstrap estimator of its variance is approximately

$$(\tilde{h}_T)^{-2} \widehat{Var}^*_{(2)}(\hat{h}_S) - 2\hat{h}_S(\tilde{h}_T)^{-3} \widehat{Cov}^*_{(2)}(\hat{h}_S, \hat{h}_T)$$
$$+ \hat{h}^2_S(\tilde{h}_T)^{-4} \widehat{Var}^*_{(2)}(\hat{h}_T).$$

If n is large, this expression is close to the direct estimator of $Var(\hat{G}_{ST})$ defined by Pons & Petit (1995). However, for small n, the bias of this bootstrap estimator will become apparent and the direct estimation is preferable.

Numerical example

The data set originates from a large study of gene diversity of sessile oak (*Quercus petraea* (Matt.) Liebl.) in Europe using several isozyme markers (Zanetto & Kremer, 1995). A total of 81 populations were sampled over most of the European range of this species. We selected a single locus (acid phosphatase, EC 3.1.3.2). Sessile oak is a diploid species and a total of five alleles and 12 genotypes were detected at this locus in the survey.

For the purposes of illustrating the approach described in this paper, all the analyses are made at the genotypic level, where each genotype is equivalent to an allele in a haploid locus. Alternatively, Hardy–Weinberg equilibrium could have been assumed, to consider the data as haploid. This has no consequence in regard to the question studied here.

The mean number of individuals per population is 114.6, for a total of 9281 individuals analysed. Alleles *2* and *4* are largely predominant, and genotypes *22*, *24* and *44* make up 98 per cent of all genotypes found. In Table 1, simulated bootstrap estimates of the three parameters are compared with the direct estimates (Pons & Petit, 1995). In accordance with the theoretical results presented previously, bootstrapping over individuals only provides an unbiased estimate of h_S but a biased one for h_T, and bootstrapping over populations only provides a biased estimate of h_S and an unbiased one for h_T, whereas the two-step bootstrap provides

biased estimates for both h_S and h_T. All bootstrap estimates for G_{ST} are therefore biased.

In Table 2, we computed total, inter- and intra-population variances for the estimates of h_S, h_T and G_{ST}, following the method of Pons & Petit (1995). The estimators of h_S and h_T have similar variances but the estimate of G_{ST} (which directly derives from the other two parameters) is less precise. These estimates are then compared to those obtained by a bootstrap procedure, either empirically (1000 bootstrap simulations) or using eqns (4–9). The three types of bootstrap were considered, as discussed above. Overall, the bootstrap estimates are in excellent agreement with the results obtained in the simulations (Table 2). Hence, the approximations (7–9) are acceptable in this example. The same approximations led to the expression of $\text{Var}_{\text{intra}}(\hat{h}_T)$, $\text{Var}_{\text{inter}}(\hat{h}_T)$, $\text{Cov}_{\text{intra}}(\hat{h}_S, \hat{h}_T)$ and $\text{Cov}_{\text{inter}}(\hat{h}_S, \hat{h}_T)$ in Pons & Petit (1995). By analogy, these expressions must also be sufficiently precise. The comparison of the variances obtained using either bootstrap (through simulations or estimations) or direct estimates clearly shows that a two-step bootstrap yields the sum of the total and intrapopulation variances instead of the total variance, in agreement with the theory developed in the previous section. Moreover, the bootstrap over populations gives estimates of the total variances (and not of the interpopulation variances). Finally, the bootstrap over individuals does estimate the intrapopulation variance, though it appears to give an inflated estimate in the case of G_{ST}.

Because the population sample sizes are large, the bootstrap over individuals is not greatly modified by the sampling of n_k-1 individuals instead of n_k, as proposed above (results unchanged for $\text{Var}(\hat{h}_T)$ for

Table 1 Direct estimates and bootstrap simulated estimates of the parameters (using 1000 bootstrap samples). The direct estimates are $\hat{h}_S$, $\hat{h}_T$ and $\hat{G}_{ST}$ (Pons & Petit, 1995) and the biased estimates are $\tilde{h}_S$, $\tilde{h}_T$ given by eqn (1) and $\tilde{G}_{ST} = 1 - \tilde{h}_S/\tilde{h}_T$

		h_S	h_T	G_{ST}
Direct estimates	Unbiased	0.6305	0.6491	0.0286
	Biased	0.6249	0.6488	0.0368
Bootstrap simulations	Individuals	0.6250	0.6492	0.0373
	Populations	0.6306	0.6488	0.0280
	Pop. + Ind.	0.6252	0.6490	0.0367

Table 2 Direct and bootstrap variance estimates $\times 10^4$ of gene diversity estimates for the complete data set (using 1000 bootstrap samples)

		$\hat{h}_S$	$\hat{h}_T$	$\hat{G}_{ST}$
Direct estimates	Variance intrapop.	0.074	0.053	0.051
	Variance interpop.	0.119	0.121	0.338
	Total variance	0.193	0.174	0.388
Bootstrap procedure	Method			
Individuals	Simulation	0.083	0.058	0.077
	Estimation	0.079	0.056	0.073
Populations	Simulation	0.184	0.166	0.361
	Estimation	0.191	0.172	0.373
Pop. + Ind.	Simulation	0.257	0.236	0.452
	Estimation	0.270	0.228	0.456

Table 3 Direct and bootstrap variance estimates $\times 10^4$ of gene diversity estimates for a subset of 20 randomly sampled individuals in each of the 81 populations (using 1000 bootstrap samples)

		$\hat{h}_S$	$\hat{h}_T$	$\hat{G}_{ST}$
Direct estimates	Variance intrapop.	5.27	2.29	4.07
	Variance interpop.	0.97	2.09	3.66
	Total variance	6.24	4.38	7.73
Bootstrap procedure	Method			
Individuals	Simulation	6.69	3.34	10.84
	Estimation	6.44	3.11	10.45
Populations	Simulation	6.02	4.54	6.99
	Estimation	6.16	4.32	7.39
Pop. + Ind.	Simulation	11.11	7.45	16.54
	Estimation	12.60	7.43	17.92

the bootstrap over individuals: 0.058×10^{-4} by both methods). Another example was studied using a subset of the complete data set where 20 individuals were selected at random and without replacement in each of the 81 populations. Direct and bootstrap variance estimates were then computed as before. The results indicate that, with this sample size, the procedure of bootstrapping over individuals no longer provides a good estimate of the intrapopulation variance of G_{ST} (Table 3). Hence, we recommend the use of the direct analytical estimates in these situations.

Conclusion

Resampling procedures, as emphasized by Crowley (1992), are often used without being validated, especially in the field of population genetics. Although seemingly appealing, intuitive resampling procedures that mimic the sampling of individuals and populations may turn out to be misleading. Here we have shown that the bootstrap variance estimators of $\hat{h}_S$, $\hat{h}_T$ and $\hat{G}_{ST}$ are obtained by resampling over populations only, instead of over both populations and individuals. Nevertheless, we recommend rather the use of the direct estimators (Pons & Petit, 1995) that we have indirectly validated here. These estimators do not require the computing time necessary for resampling procedures and they are unbiased.

Acknowledgements

References

CROWLEY, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Ann. Rev. Ecol. Syst.*, **23**, 405–447.

EFRON, B. AND TIBSHIRANI, R. J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

EXCOFFIER, L., SMOUSE, P. E. AND QUATTRO, J. M. 1992. Analysis of molecular variance inferred from metric distances among haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

HUDSON, R. R., BOOS, D. D. AND KAPLAN, N. L. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.*, **9**, 138–151.

McDONALD, J. H. 1994. Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms. In: Golding, B. (ed.) *Non-neutral Evolution. Theories and Molecular Data*, pp. 88–100. Chapman and Hall, New York.

NEI, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 3321–3323.

NEI, M. 1986. Definition and estimation of fixation indices. *Evolution*, **40**, 643–645.

PALUMBI, S. R. AND WILSON, A. C. 1990. Mitochondrial DNA diversity in the sea urchins *Strongylocentrotus purpuratus* and *S. droebachiensis*. *Evolution*, **44**, 403–415.

PETIT, R. J., KREMER, A. AND WAGNER, D. B. 1993. Geographic structure of chloroplast DNA polymorphisms in European oaks. *Theor. Appl. Genet.*, **87**, 122–128.

PONS, O. 1997. Bootstrap variance of diversity and differentiation estimators in a subdivided population: technical proofs. *Unpublished manuscript available from the author*.

PONS, O. AND CHAOUCHE, K. 1995. Estimation, variance and optimal sampling of gene diversity. II. Diploid locus. *Theor. Appl. Genet.*, **91**, 122–130.

PONS, O. AND PETIT, R. J. 1995. Estimation, variance and optimal sampling of gene diversity. I. Haploid locus. *Theor. Appl. Genet.*, **90**, 462–470.

PONS, O. AND PETIT, R. J. 1996. Measuring and testing genetic differentiation with ordered and unordered alleles. *Genetics*, **144**, 1237–1245.

PROUT, T. AND BARKER, J. S. F. 1993. *F*-statistics in *Drosophila buzzatii*: selection, population size and inbreeding. *Genetics*, **134**, 369–375.

RAO, J. N. K. AND WU, C. F. J. 1988. Resampling inference with complex survey data. *J. Am. Stat. Ass.*, **83**, 231–241.

SITTER, R. R. 1992. A resampling procedure for complex survey data. *J. Am. Stat. Ass.*, **87**, 755–765.

VAN DONGEN, S. 1995. How should we bootstrap allozyme data? *Heredity*, **74**, 445–447.

WEIR, B. S. 1990. *Genetic Data Analysis*. Sinauer Associates, Sunderland, MA.

WEIR, B. S. AND COCKERHAM, C. C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

WRIGHT, S. 1951. The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

ZANETTO, A. AND KREMER, A. 1995. Geographical structure of gene diversity in *Quercus petraea* (Matt.) Liebl. I. Monolocus patterns of variation. *Heredity*, **75**, 506–517.