2010

High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment

Robin Waples
NOAA, robin.waples@noaa.gov

MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2599–2601

NEWS AND VIEWS

PERSPECTIVE

High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment

ROBIN S. WAPLES
NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

Abstract

Recognition of the importance of cross-validation ('any technique or instance of assessing how the results of a statistical analysis will generalize to an *independent dataset*'; Wiktionary, en.wiktionary.org) is one reason that the U.S. Securities and Exchange Commission requires all investment products to carry some variation of the disclaimer, 'Past performance is no guarantee of future results.' Even a cursory examination of financial behaviour, however, demonstrates that this warning is regularly ignored, even by those who understand what an independent dataset is. In the natural sciences, an analogue to predicting future returns for an investment strategy is predicting power of a particular algorithm to perform with new data. Once again, the key to developing an unbiased assessment of future performance is through testing with independent data—that is, data that were in no way involved in developing the method in the first place. A 'gold-standard' approach to cross-validation is to divide the data into two parts, one used to develop the algorithm, the other used to test its performance. Because this approach substantially reduces the sample size that can be used in constructing the algorithm, researchers often try other variations of cross-validation to accomplish the same ends. As illustrated by Anderson in this issue of *Molecular Ecology Resources*, however, not all attempts at cross-validation produce the desired result. Anderson used simulated data to evaluate performance of several software programs designed to identify subsets of loci that can be effective for assigning individuals to population of origin based on multilocus genetic data. Such programs are likely to become increasingly popular as researchers seek ways to streamline routine analyses by focusing on small sets of loci that contain most of the desired signal. Anderson found that although some of the programs made an attempt at cross-validation, all failed to meet the 'gold standard' of using truly independent data and therefore produced overly optimistic assessments of power of the selected set of loci—a phenomenon known as 'high grading bias.'

Keywords: assignment tests, cross-validation, discriminant function analysis, independence, jackknife, sample size, split-sample

Received 6 April 2010; revision received 16 April 2010; accepted 18 April 2010

The basic problem posed by failure of proper cross-validation can be illustrated with an example using discriminant function analysis (DFA), which is conceptually very similar to population assignments based on genetic data (Hansen *et al.* 2001). Figure 1 shows results of a DFA (conducted using Systat 12) based on simulated data for individuals

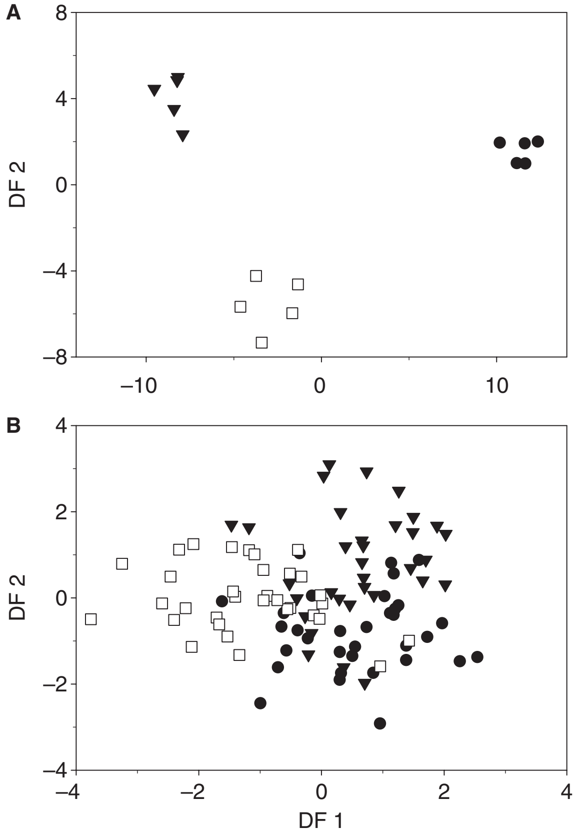

Fig. 1 Illusory appearance of differences between three arbitrary 'populations' of individuals based on discriminant function analysis of random data. The problem is more acute with small numbers of individuals per group (Panel A), especially when each is scored for a large number of variables.

Correspondence: Robin S. Waples, Fax: +1 206 860 3335;
E-mail: robin.waples@noaa.gov

arbitrarily grouped into three 'populations.' In this example, no real population differences exist, as trait values for each character in each individual were randomly assigned by drawing from a Standard Normal distribution. In Panel A, each group had five individuals scored for 33 different continuous characters. The 12 characters with the largest variance between means of the three arbitrary groups were used in the DFA (thus mimicking what is done with a locus-selection program), which produces discriminant functions that are linear combinations of the variables that maximize group differences. In this example, with few individuals in each group and lots of variables to choose from, it is easy to find a few that (entirely by chance) have large inter-group differences. The DFA then weights those variables most heavily, with the result that the three arbitrary groups appear to represent very distinct populations. This impression is reinforced if one considers the fraction of individuals (100%) that can be correctly assigned to their 'population' of origin based on their trait values. However, this high apparent power is completely illusory, as all the between-individual and between-population differences are random. Fortunately, DFA has a couple of ways of alerting one to overly optimistic estimates of self-assignment success. First, a multivariate test can evaluate whether overall group differences are larger than can be attributed to chance. For data in Fig. 1A, the *P*-value for Wilk's lambda is >>0.05, as would be expected for random data. With genetic data, the analogue would be a multilocus test of heterogeneity of allele frequencies; if this is not significant, any attempt to evaluate power to discriminate the 'populations' is suspect. Second, DFA has a simple method of cross-validation (jackknifing, or leave-one-out, termed LOO in Anderson 2010). When the discriminant functions for Fig. 1A were recalculated after sequentially leaving each individual out of the analysis, the percent of individuals correctly allocated dropped to 20%—not significantly different from the 1/3 expected by chance. Panel B shows a similar analysis but with 33 individuals in each random group. In this case, there is still a suggestion of spurious intergroup differences (partially non-overlapping discriminant scores; 72% self-assignment accuracy that drops to 38% with jackknifing; non-significant Wilk's lambda). However, results are not nearly as overly optimistic because with more individuals per group there is a much smaller chance for random intergroup differences to be large.

An intriguing point made by Anderson (2010) is that the locus-selection programs lead to overly optimistic assessments of power in spite of some attempts by program authors to deal with cross-validation issues. The key is that the locus-selection process involves two major steps in developing the algorithm: (i) estimating allele frequencies based on samples of individuals from target populations, and (ii) identifying loci with the highest power to detect population differences identified in Step 1. To ensure independence, data used to assess power of the selected set of loci cannot have been used in *either* Step 1 or Step 2 (done correctly, this is termed 'double cross-validation' by Anderson). It appears that the locus-selection programs have either (i) used the holdout set for Step 2, (ii) incompletely implemented the jackknife (LOO) option, or (iii) not attempted cross validation at all. In Fig. 1, the jackknife option is effective in revealing spurious estimates of self-assignment accuracy because the entire process of calculating group means and calculating new discriminant functions is repeated when each individual is sequentially left out of the analysis. However, to do this properly with the locus-selection programs, the process of locus selection (not just assignment to population) would have to be repeated with each individual removed from the analysis. This would likely result in different mixes of loci being selected for use with each individual, which would complicate interpretation of results and presumably explains why double cross-validation LOO is not implemented in these programs.

Anderson's paper raises three important points that should be kept in mind by those interested in evaluating power of genetic methods.

1. The problem with the software programs is not in the process for selecting informative loci, but rather with the method to assess power in future applications. That is, these programs can be effective in identifying subsets of loci that can help reduce costs and streamline analyses, but they tend to provide an overly optimistic assessment of power.
2. Achieving optimal cross-validation can be difficult, as different methods have advantages and disadvantages (Stone 1977; Efron 1982; Goutte 1997). For example, the split-sample method ensures independence [and hence is termed the 'gold standard' here and 'obviously correct' by Anderson (2010)]; however, it is wasteful of data and for some applications has less desirable properties than *k*-fold cross-validation (in which the data are split into *k* groups of roughly equal size and the algorithm is developed by sequentially leaving out one group and using the others as the training sample). If *k* equals the total sample size, the latter method is equivalent to LOO. Although LOO is widely used, it affects sample size (e.g. each jackknifed group mean in Fig. 1A is based on four rather than five individuals) and hence changes the variance structure of the data. The different cross-validation methods reflect the inherent tension between the desire for complete independence and the desire to use as much data as possible to construct the algorithm. Anderson's innovative suggestion to combine the LOO and split-sample approaches provides a nice balance in dealing with this tradeoff and merits consideration for broader application.
3. Problems with high-grading bias and related issues are most severe when (i) sample sizes of individuals are small; (ii) large numbers of characters are used; and (iii) true differences between groups are small. If populations are genetically divergent, power is already high and the relative influence of high-grading bias will be small. However, the increasing availability of large number of genetic markers for non-model organisms has encouraged researchers to try to address challenging

problems in conservation and evolution that could not be attempted previously because the underlying signal is weak. These applications require particular attention to issues related to cross-validation.

Acknowledgements

I thank Jeff Hard for conducting the DFA analyses.

References

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, **10**, 701–710.

Efron B (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Goutte C (1997) Note on free lunches and cross-validation. *Neural Computation*, **9**, 1245–1249.

Hansen MM, Kenchington E, Nielsen EE (2001) Assigning individual fish to population using microsatellite DNA markers. *Fish and Fisheries*, **2**, 93–112.

Stone M (1977) Asymptotics for and against cross-validation. *Biometrika*, **64**, 29–35.

Note: A typographical error occurs in Table 1 of Anderson (2010). There, the definition of CARNI is given as the fraction of simulated individuals *incorrectly* assigned to their population of origin; however, the CARNI is the fraction of simulated individuals *correctly* assigned to their population of origin. The author regrets any confusion this typographical error may have created.

R. S. W.'s research focuses on population genetics and conservation genetics, with emphasis on marine and anadromous species. He is interested in developing and applying population genetic principles to real-world problems in ecology, conservation, and management.

doi: 10.1111/j.1365-294X.2010.04675.x