

Appendix 1: Software settings

The software package we use has many purposes. Consequently it provides many choices to the analyst, who must set its flags deliberately and carefully. In all the analyses of this paper, we chose the following settings:

- **Sample the occurrences with replacement.** This prevents changes in the probability of a species occurring as shuffling takes place. That targets the diversity in the computer rather than including any unsampled species that might remain to be discovered in the real world. Because we know exactly how many species are in the computer's data base, we can precisely judge the performance of the estimators.
- **Shuffle only pooled.** This focuses the estimators on those ecoregions already included in a sample subset of the pool of ecoregions. Otherwise, the software would select occurrences from the entire set of 110 ecoregions when building the virtual lists of every ecoregion subset. The walls of habitat difference among ecoregions would be destroyed (virtually) and the software would be able to cheat, knowing about all occurrences at every step although it seemed to admit to knowing only a subset of them.
- **Shuffle incidences only.** This flag must be used when a presence/absence data matrix is being analyzed. It prevents species from registering multiple occurrences in a single ecoregion.
- **Retain sample sizes.** Also required with a presence/absence data set.
- **Shuffle sample order.** Exception: not chosen in every case. We chose it for random runs, including those mimicking the analysis that would occur after exactly eleven spread-list ecoregions had been surveyed. But we did not choose it for runs from a kernel ecoregion (maintaining the order of the ecoregions was the point of these runs).

ESTIMATING DIVERSITY IN UNSAMPLED HABITATS

Appendix 2: Kernel selection

We wanted the kernels to be a stratified random sample of the 110 ecoregions. To achieve this, we apportioned the kernels evenly among ecoregions in the northern, western, southern, eastern, and central portions of the continent. We began by placing all ecoregions in four ordered lists: north to south, west to east, east to west, and south to north. Then we selected the first ecoregion in the north to south list, placed it in the north (N) bin, and eliminated it from all four lists. Then we selected the first ecoregion in the west to east list, placed it in the west (W) bin, and eliminated it from all four lists. We continued until all 110 ecoregions had been placed in four (N, W, E, and S) bins. N and W had 28 ecoregions apiece, E and S had 27. For the central (C) bin, we removed the six ecoregions added last to N, the six added last to W and the five each added last to E and to S. Thus we built five bins, each containing 22 ecoregions. By rolling a die, we chose four ecoregions from each bin as kernels. The dense grouping of ecoregions on the Pacific coast of the United States caused one kernel from the E bin to be unacceptably close to the center of the continent, so this ecoregion was added to the C kernels and a new one was selected from the E bin. Thus, the procedure yielded 21 kernels.