

# Museum collections, species distributions, and rarefaction

Andrew R. Solow<sup>1\*</sup> and David L. Roberts<sup>2</sup>

<sup>1</sup>Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA and <sup>2</sup>Royal Botanic Garden, Kew, Richmond, Surrey TW9 3AE, UK

\*Correspondence: Andrew R Solow, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. E-mail: asolow@whoi.edu

## ABSTRACT

Biological specimens in museums and herbaria are sometimes used to compare the geographical distribution of different species. In doing so, it is necessary to account for differences in the numbers of specimens. We show how rarefaction can be used for this purpose. Rarefaction is a simple mathematical method originally designed to compare species richness in communities that differed in the number of sampled individuals. We present an example involving two *Phragmipedium* orchid species. In this case, rarefaction suggests that the apparent difference in range can be explained by the difference in the numbers of specimens.

## Keywords

Museum and herbaria collections, orchids, rarefaction, sampling, species distributions.

## INTRODUCTION

Biological collections in museums and herbaria can provide important information about the geographical distributions of plant and animal species (Suarez & Tsutsui, 2004). Indeed, for some species, collections are the only source of such information. The value of this information is growing with the demand for rapid and inexpensive conservation assessments (IUCN, 2001; Willis *et al.*, 2003; Nic Lughadha *et al.*, 2005). As with other kinds of biological data, in using museum and herbarium specimens to infer species distributions, care must be taken to avoid bias due to sampling effects. One such bias can arise in comparing the distributions of two or more species based on collections containing different numbers of individuals. This is analogous to the well-known problem of comparing the observed species richness of two or more communities when the numbers of individual organisms sampled from the communities are different. The recognition of this latter problem led to the development of the method known as rarefaction. The basic idea of rarefaction is due to Sanders (1968) — see Gotelli & Colwell (2001) for a recent discussion. Here, we point out that the same formalism can be used in comparing species distributions when the numbers of specimens differ.

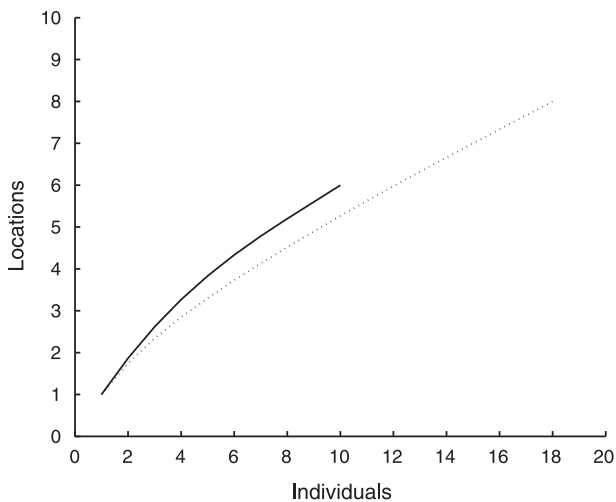
## RAREFYING COLLECTION LOCATIONS

To fix ideas, it is useful to consider a specific example. Table 1 reports the number of specimens by collection location in Ecuador of two species of the orchid genus *Phragmipedium* Rolfe reported in the W<sup>3</sup>TROPICOS VAST database (<http://mobot.mobot.org/W3T/Search/vast.html>) maintained by the Missouri Botanical Garden. On the basis of the number of

locations, *Phragmipedium longifolium* (Warsz. ex Rchb.f) Rolfe has the wider distribution with eight compared to *Phragmipedium hirtzii* Dodson with six. On the other hand, there are a total of 18 specimens of *P. longifolium* compared to only 10 of *P. hirtzii*. In the same way that the number of observed species is expected to increase with the number of observed individuals, the number of locations at which a species is observed is expected to increase with the number of specimens collected. For this reason, it is of interest to ask how the distributions of these species would compare when the sampling effort, as measured by the number of specimens collected, is the same. The purpose of rarefaction is to perform just such a comparison.

**Table 1** The number of specimens by locations in Ecuador of the orchids *Phragmipedium hirtzii* Dodson and *Phragmipedium longifolium* (Warsz. ex Rchb.f) ROLFE

| Location            | <i>P. hirtzii</i> | <i>P. longifolium</i> |
|---------------------|-------------------|-----------------------|
| Alto Tambo          | 0                 | 1                     |
| El Pailon           | 1                 | 0                     |
| Lita-Alto Tambo     | 3                 | 3                     |
| Lita-San Lorenzo    | 3                 | 1                     |
| Nayon               | 0                 | 1                     |
| Nono-Nanegal        | 0                 | 1                     |
| Quinde (Bilsa)      | 1                 | 0                     |
| Quito-Punto Viejo   | 0                 | 1                     |
| Reserva Awa         | 1                 | 1                     |
| San Lorenzo         | 1                 | 0                     |
| Santo Domingo-Quito | 0                 | 9                     |



**Figure 1** Rarefaction curves based on the data in Table 1 for *Phragmipedium hirtzii* Dodson (solid) and *Phragmipedium longifolium* (Warsz. ex Rchb.f) Rolfe (dashed).

Suppose that a species is collected at a total of  $k$  locations and let  $m_j$  be the number of specimens from location  $j$  ( $j = 1, 2, \dots, k$ ) and let  $m = \sum_{j=1}^k m_j$  be the total number of specimens. Define the random variable  $L(n)$  as the number of locations represented in a random sample of  $n$  of these specimens ( $n \leq m$ ). The expected value of  $L(n)$  is:

$$E(L(n)) = k - \sum_{j=1}^k \binom{m - m_j}{n} / \binom{m}{n} \quad (1)$$

where, for non-negative integers  $a$  and  $b$ ,  $\binom{a}{b} \equiv 0$  when  $a < b$ . This result is given in Walton (1986), along with higher-order moments of  $L(n)$ . A plot of  $E(L(n))$  against  $n$  is called a rarefaction curve. The result in Eqn. 1 is exact and there is no need to resort to simulation — as is commonly done in rarefying species richness (e.g. Gotelli & Colwell, 2001) — to estimate  $E(L(n))$ .

Returning to the data in Table 1, Fig. 1 shows the rarefaction curves for the two *Phragmipedium* species. The largest value of  $n$  at which these species can be compared is 10, corresponding to the total number of specimens of *P. hirtzii* in a total of six locations. For this value of  $n$ , the expected number of locations for *P. longifolium* is around 5.3. Although these data are far too few to draw firm conclusions, in this case, correcting for the number of specimens suggests that the distributions of these species may be more similar than suggested by the uncorrected location data.

## DISCUSSION

The expression in Eqn. 1 is the exact answer to the question: What is the expected number of locations represented in  $n$  specimens

of a species sampled at random from the larger collection of  $m$  specimens? Although this is not nearly as interesting as a statement about the geographical distribution of the species in nature, it is still useful in understanding the information contained in such collections. To go further, it is necessary to make some assumption about the collection process itself. For example, to treat  $E(L(n))$  as a predictor of the number of locations that would be found in a field sample of  $n$  individuals, it is necessary to assume that the specimens themselves were collected at random. To go beyond this to the construction of a confidence interval requires further assumptions — see Colwell *et al.* (2004) for one such attempt in the context of species richness.

Turning to the application of the previous section, it is worth noting that, while *P. longifolium* has been recognized since 1852, *P. hirtzii* was only identified in 1988. This may explain, in part, the relative paucity of specimens of this species. When sampling effort is consistently lower for recently identified species, there will be a tendency to underestimate their ranges on the basis of collection locations alone. This underlines the need to control for sampling effort in interpreting observed differences in distribution.

## ACKNOWLEDGEMENTS

The helpful comments of Mark Burgman are acknowledged with gratitude.

## REFERENCES

- Colwell, R.K., Mao, C.X. & Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2171–2727.
- Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.
- IUCN (2001) *Red list categories and criteria*, Version 3.1. IUCN, Gland, Switzerland.
- Nic Lughahda, E. & others. (2005) Measuring the fate of plant diversity: towards a foundation for future monitoring and opportunities for urgent action. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **B360**, 359–372.
- Sanders, H. (1968) Marine benthic diversity: a comparative study. *American Naturalist*, **102**, 243–282.
- Suarez, A.V. & Tsutsui, N.D. (2004) The value of museum collections for research and society. *Bioscience*, **54**, 66–74.
- Walton, G.S. (1986) The number of observed classes from a multiple hypergeometric distribution. *Journal of the American Statistical Association*, **81**, 169–171.
- Willis, E., Moat, J. & Paton, A. (2003) Defining a role for herbarium data in Red List assessments: a case study of *Plectranthus* from eastern and southern tropical Africa. *Biodiversity and Conservation*, **12**, 1537–1552.