Estimating vocal repertoire size is like collecting coupons: A theoretical framework with heterogeneity in signal abundance

Arik Kershenbaum [a,b,*], Todd M. Freeberg [c], David E. Gammon [d]

[a] National Institute for Mathematical and Biological Synthesis, Knoxville, TN, USA
[b] Department of Zoology, University of Cambridge, Cambridge, England
[c] Department of Psychology, University of Tennessee, Knoxville, TN, USA
[d] Department of Biology, Elon University, Elon, NC, USA

HIGHLIGHTS

- Vocal repertoire size is an important measure of animal behaviour.
- Asymptotic curve fitting to find repertoire size is not based on probability theory.
- We derived the precise expression using the "Coupon Collector's problem".
- We then tested this model using data from two avian and one mammalian species.
- Coupon Collection gives better estimates of repertoire size than traditional methods.

ARTICLE INFO

ABSTRACT

Vocal repertoire size is an important behavioural measure in songbirds and mammals with complex vocal communication systems, and has traditionally been used as an indicator of individual fitness, cognitive ability, and social structure. Estimates of asymptotic repertoire size have typically been made using curve fitting techniques. However, the exponential model usually applied in these techniques has never been provided with a theoretical justification based on probability theory, and the model has led to inaccurate estimates. We derived the precise expression for the expected number of distinct signal types observed for a fixed sampling effort: a variation of what is known in the statistical literature as the "Coupon Collector's problem". We used empirical data from three species (northern mockingbird, Carolina chickadee, and rock hyrax) to assess the performance of the Coupon Collector model compared to commonly used techniques, such as exponential fitting and repertoire enumeration, and also tested the different models against simulated artificial data sets with the statistical properties of the empirical data. We found that when signal probabilities are dissimilar, the Coupon Collector model provides far more accurate estimates of repertoire size than traditional techniques. Enumeration and exponential curve fitting greatly underestimated repertoire size, despite appearing to have reached saturation. Application of the Coupon Collector model can generate more accurate estimates of repertoire size than the commonly used exponential model of repertoire discovery, and could go a long way towards re-establishing repertoire size as a useful indicator in animal communication research.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Assessing the repertoire size of animal vocalisations has long been recognised as important to understanding the development, function, and evolution of animal communication (Bradbury and Vehrencamp, 2011). Repertoire size is often considered to be an indicator of communicative complexity (Freeberg et al., 2012), and communicative complexity has numerous implications for the understanding of animal signalling and social evolution. It has been proposed that communicative complexity may drive the evolution of social complexity, and vice versa (McComb and Semple, 2005; Pollard and Blumstein, 2012). Communicative complexity may be an indicator of individual fitness (Hiebert et al., 1989; Darolová et al., 2012), may influence mate choice and the evolution of mating systems (Searcy, 1992; Nowicki et al., 2000), and may play a role in complex sender–receiver competitive games (Owren et al., 2010). Repertoire size is also an

important metric for assessing the tradeoff between function and variant diversity in communication (Bradbury and Vehrencamp, 2011), and the neurological constraints limiting the use of highly complex signalling, such as syntax (Lipkind et al., 2013). Although most previous research related to repertoire size has concentrated on the repertoire of birdsong, the work we describe here can be generalised to consider other non-song modalities such as visual signals (Peters and Ord, 2003), chemical signals (delBarco-Trillo et al., 2012), and even to entire behavioural repertoires (Sempo and Detrain, 2004). As such, hereafter we use the term "signals", rather than "songs", except where specifically referring to mating/territorial songs.

Despite the importance of repertoire size in animal species, repertoire size estimation for any given species or individual is a non-trivial problem in practical situations. For species with large repertoires, large sampling effort is required to enumerate all distinct vocalisations. Some species of songbird, furthermore, produce much of the diversity of their song repertoire quickly (immediate variety), whereas other species produce the diversity of their song repertoire over much longer time frames (eventual variety; Catchpole and Slater, 2003). Even if true repertoire size is small, enough data need to be collected to ensure that all possible types have been recorded (Hesler et al., 2012). In most cases, comprehensive sampling is impractical, and so estimation techniques must be used. Most work on estimating repertoire size has been carried out with songbirds, since song repertoires play such a crucial role in the social and reproductive life of many species (Catchpole and Slater, 2003). Songbird repertoire sizes range from a single stereotyped song such as in the white-crowned sparrow *Zonotrichia leucophrys* (Soha and Marler, 2001), through tens of distinct vocalisations, e.g. in the European blackbird *Turdus merula*, up to species such as the northern mockingbird *Mimus polyglottos* and brown thrasher *Toxostoma rufum*, which can make use of hundreds or even thousands of different songs (Kroodsma, 1977; Gammon and Altizer, 2011). Some songbird species, furthermore, possess diverse repertoires of calls, indicating that repertoire estimation is important to non-song vocal behaviour as well (e.g., call repertoires in crows and jays, chickadees and titmice, and parrots). Moreover, songbirds are not the only taxon where repertoire estimation is important. Some mammal species have extensive vocal repertoires, such as the rock hyrax *Procavia capensis* (Kershenbaum et al., 2012), free-tailed bats *Tadarida brasiliensis* (Bohn et al., 2009), marmots *Marmota flaviventris* (Blumstein, 2007), and pilot whales *Globicephala macrorhynchus* (Sayigh et al., 2012), and little is known about their social and evolutionary significance.

2. Previous work

Estimating the repertoire size of individuals or a group of individuals has been accomplished using different methods. In one of the earliest methods (henceforth WM), Wildenthal (1965) proposed measuring the total number of distinct signals observed as sample size increases, and fitting these data to an exponential function of the form

$$E = N\left(1 - e^{-M/N}\right) \tag{1}$$

where N is the true repertoire size, M is the number of observations (samples) made, and E is the number of distinct signals observed. Wildenthal originally proposed this relationship based on the observation that, "when the data were plotted… the curves obtained approximate exponential curves" (Wildenthal, 1965). Davidson and Wilkinson (2002) subsequently proposed a correction to the Wildenthal model which appeared to give a better fit to

the empirical data. In the Davidson and Wilkinson (2002) model (henceforth, DW), the rate of increase of distinct signal types observed is lower than expected by the WM, by a factor A, where $A > 1$

$$E = N\left(1 - e^{-M/AN}\right) \tag{2}$$

The DW is purely empirical, as no known probabilistic process would be expected to produce a relationship as indicated in Eq. (2). However, even the derivation of the WM remains unexplained, and to date no examination has been made of why the number of observed signals should follow this exponential relationship, or what the theoretical justification for such a model might be.

A different approach, examined by Garamszegi et al. (2005) applied the capture–recapture principle to the observation of signal types. Using the capture–recapture approach, the researcher models the observation of signals in the same way as observation of marked individuals in population sampling; an established technique used for the estimation of population size. This technique also gives repertoire size estimates that are empirically accurate under certain conditions, and is based on the theoretical similarity between sampling signals, and the sampling of marked individuals.

Botero et al. (2008) compared these estimation techniques using artificial repertoires with known repertoire sizes, and concluded that in many cases, simple enumeration of the number of distinct signals observed (henceforth, EN) is the easiest and most accurate estimator of true repertoire size. Since then, other researchers have proposed additional techniques that appear empirically valid, such as rarefaction (Peshek and Blumstein, 2011), but no bottom-up analysis of the process of signal sampling has been performed to derive the precise relationship between sampling effort and repertoire estimate. In practice, many researchers simply use the enumeration of distinct observed signals, on the assumption that sufficient signals have been sampled to represent the entire repertoire accurately (e.g. Searcy, 1992; Nowicki et al., 2000; Pfaff et al., 2007; Hesler et al., 2011).

We show that this problem of sampling signals from a repertoire is a variant of the "Coupon Collector's problem" (Erdös and Rényi, 1961; Jocković and Mladenović, 2011), and we use probability theory to derive the precise expression for the expected number of distinct observed signals, given a particular sampling effort. We show that this result closely approximates Wildenthal's and Davidson–Wilkinson's exponential models only when each signal occurs with a similar probability. When individual signals do not occur homogenously, exponential fitting is likely to underestimate total repertoire size. Although no closed-form expression exists for the repertoire size estimate, non-linear least-squares fitting provides an accurate estimate, if sufficient data are available to estimate the signal probabilities. Finally, we compare our method to other common methods for assessing repertoire size, using examples of animal vocalisations: mockingbird song, chickadee calls, and hyrax songs. Matlab scripts implementing this method are available for download at http://sourceforge.net/projects/couponcollectoralgorithm/.

3. Mathematical theory of the model

Sampling signals from a repertoire can be represented as a variant of the Coupon Collector's problem, solved by Erdös and Rényi (1961) and cited in Jocković and Mladenović (2011). Suppose we collect coupons (sometimes the problem is posed as collecting baseball cards, or similar) of which there are N different types. We randomly select coupons with replacement, i.e. the coupons are