

STATISTICAL REPORTS

Ecology, 85(7), 2004, pp. 1822–1825
© 2004 by the Ecological Society of America

ESTIMATING THE RATE OF SPECIES INTRODUCTIONS FROM THE DISCOVERY RECORD

ANDREW R. SOLOW^{1,3} AND CHRISTOPHER J. COSTELLO²

¹*Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543 USA*

²*University of California, Santa Barbara, California 93106 USA*

Abstract. The discovery record of introduced species reflects a combination of the introduction process and the discovery process. For this reason, the discovery record does not provide a direct proxy for the record of introductions. We describe a general method for estimating the rate of introductions from the discovery record. The method is based on a statistical model of the discovery record that includes both the introduction and discovery processes. The method is illustrated using the discovery record of introduced species in the San Francisco estuary (California, USA). The estimated mean rate of introductions increases from 0.3 introductions in 1850 to 2.3 introductions in 1995.

Key words: maximum likelihood estimation; Poisson distribution; species introductions, estimating rate.

INTRODUCTION

The introduction of nonindigenous species into marine and terrestrial environments can have significant ecological and economic consequences (Wilcove et al. 1998, Williamson 1999, Pimentel et al. 2000). To develop effective policies aimed at avoiding or mitigating these consequences, it is necessary to have a good understanding of the process of introduction. This paper focuses on the use of aggregate historical data in understanding this process. A difficulty in using the historical record is that, with some notable exceptions, the introduction process is unobservable. Instead, inferences about the introduction process must be based on the record of *discoveries* of introduced species. As Costello and Solow (2003) pointed out, the historical pattern of discoveries of introduced species, which reflects a combination of the introduction process and the discovery process, can give a misleading picture of the pattern of introductions. This raises the question: How can the discovery record be used to make inferences about the introduction process? The purpose of this paper is to take what appears to be the first step toward answering this question. To do so, we describe a statistical model of the discovery record that incorporates both the introduction and discovery processes and that, when fit to the historical record of discoveries, provides information about both processes.

The paper is organized in the following way. The basic model is described in the next section. Then this model is fit to the discovery record of introduced marine species in the San Francisco (California, USA) estuary. Next, the results of a small simulation study are presented. The final section contains some concluding remarks.

MODEL

Let the observable random variable Y_t be the number of nonindigenous species that are discovered in year t . For economy of expression, the qualifier “nonindigenous” will hereafter be omitted and we will simply refer to species. This section outlines a model of Y_t that incorporates both the process of introduction and the process of discovery.

To begin with, consider the introduction process. Let the random variable N_t be the number of species that are introduced in year t . In contrast to Y_t , N_t is unobservable. We will assume that N_t has a Poisson distribution with unknown mean introduction rate μ_t that depends on an unknown, possibly vector-valued parameter. We will also assume that the sequence N_1, N_2, \dots is independent. Under this model, any nonrandom variation in this sequence is due to variation in the mean introduction rate μ_t .

Turning to the discovery process, Y_t can be written as

$$Y_t = \sum_{s=1}^t Y_{st} \quad (1)$$

where the random variable Y_{st} is the number of species

Manuscript received 11 July 2003; revised 27 December 2003; accepted 5 January 2004; final version received 4 February 2004. Corresponding Editor: A. M. Ellison.

³ E-mail: asolow@whoi.edu

introduced in year s that are discovered in year t . The probability mass function (pmf) of Y_{st} is

$$\begin{aligned} \text{prob}(Y_{st} = y_{st}) &= \sum_{n_s=y_{st}}^{\infty} \text{prob}(Y_{st} = y_{st} | N_s = n_s) \text{prob}(N_s = n_s). \end{aligned} \quad (2)$$

The first term in the summation in Eq. 2 is the conditional pmf of Y_{st} given $N_s = n_s$. The conditional distribution of Y_{st} given $N_s = n_s$ is binomial with parameters n_s and p_{st} , where p_{st} is the probability that species introduced in year s is discovered in year t . This probability is given by

$$p_{st} = \pi_{st} \prod_{j=s}^{t-1} (1 - \pi_{sj}) \quad (3)$$

where π_{st} is the probability that a species introduced in year s is observed in year t . Here, we assume that observations of a species in different years are independent, although we do not assume that the probabilities of these observations are the same. A simple model for π_{st} is given below. The second term in the summation in Eq. 2 is given by the pmf of N_s , which by assumption is Poisson with mean μ_s . It is straightforward to show that, under this model, Y_{st} has a Poisson distribution with mean $p_{st}\mu_s$ (Ross 2000). It follows from Eq. 1 that Y_t also has a Poisson distribution with mean

$$\lambda_t = \sum_{s=1}^t \mu_s p_{st}. \quad (4)$$

Before proceeding, it is instructive to consider the simple case in which there is a constant annual introduction rate $\mu_t = \mu$ and a constant annual observation probability $\pi_{st} = \pi$. In this case, Y_t has a Poisson distribution with mean $\mu(1 - [1 - \pi]^t)$. If π is small, the mean number of discoveries of introduced species initially increases approximately linearly with t with slope $\mu\pi$ before eventually leveling out toward its asymptotic value of μ . This is an example of the way in which the pattern of discoveries can give a misleading picture of the pattern of introductions.

Once parametric models for the mean introduction rate μ_t and the observation probability π_{st} have been specified, the complete model can be fit to an observed time series of Y_1, Y_2, \dots, Y_m by the method of maximum likelihood (ML). The Poisson log-likelihood function is

$$\log L = \sum_{t=1}^m (y_t \log \lambda_t - \lambda_t) \quad (5)$$

where y_t is the observed value of Y_t and where λ_t is given in Eq. 4 and p_{st} given in Eq. 3. The ML estimates of the unknown parameters are found by maximizing Eq. 5 over the unknown parameters in μ_t and π_{st} . This maximization must be done numerically. A good dis-

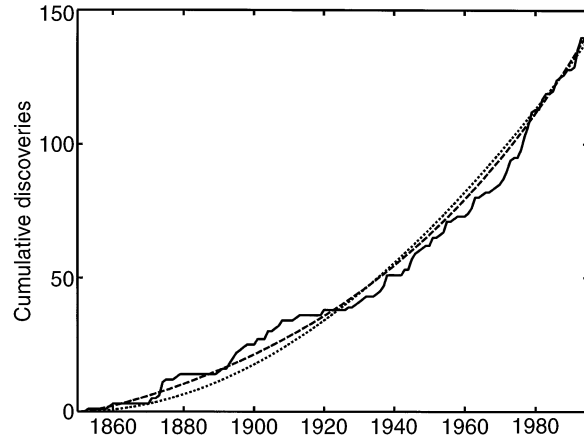


FIG. 1. The cumulative record of discoveries of introduced species in the San Francisco estuary, California, USA, 1850–1995 (solid line). Also shown are fitted values allowing for an increasing introduction rate (dashed line) and assuming a constant introduction rate (dotted line).

cussion of maximum-likelihood estimation and related methods is contained in Azzalini (1996).

APPLICATION

In this section we illustrate the use of the model outlined in the previous section by applying it to some data on the discoveries of introduced species in the San Francisco estuary (California, USA). These data were compiled by Cohen and Carlton (1995); see also Cohen and Carlton (1998). The data, which are plotted in cumulative form in Fig. 1, cover the period 1850–1995 (see Supplement). Following Cohen and Carlton (1995, 1998), we have retained for analysis 140 species with well-dated discoveries that did not arise from an extraordinary observational effort.

To fit the model, it is necessary to specify parametric forms for μ_t and π_{st} . To allow for the possibility of a monotonic trend in the mean introduction rate, we will assume that

$$\mu_t = \exp(\beta_0 + \beta_1 t) \quad (6)$$

where β_0 and β_1 are unknown parameters. As discussed in Costello and Solow (2003), π_{st} will reflect both the observational effort in year t and the abundance in year t of a species introduced in year s . A simple model that incorporates these factors is

$$\pi_{st} = \frac{\exp[\gamma_0 + \gamma_1 t + \gamma_2 \exp(t - s)]}{1 + \exp[\gamma_0 + \gamma_1 t + \gamma_2 \exp(t - s)]} \quad (7)$$

where γ_0 , γ_1 , and γ_2 are unknown parameters. Under this model, the logistic transformation $\log(\pi_{st}/[1 - \pi_{st}])$ is a linear function of t , reflecting a monotonic trend in observational effort, and $\exp(t - s)$, reflecting exponential post-introduction growth in abundance. The logistic transformation is commonly used to model the dependence of a probability on covariates.

The maximum-likelihood (ML) estimates of the parameters of the full model are

$$\hat{\beta}_0 = -1.1 \quad \hat{\beta}_1 = 0.014 \quad \hat{\gamma}_0 = -1.46$$

$$\hat{\gamma}_1 = 0.00001 \quad \hat{\gamma}_2 = 0.0000004$$

and the maximized value of the log likelihood is -122.56 . The point estimates of γ_1 and γ_2 are extremely small and, in fact, are not significantly different from 0. Specifically, fitting the model under the null hypothesis that $\gamma_1 = \gamma_2 = 0$ decreases the log likelihood to only -122.58 for an approximate significance level or P value of around 0.98. The ML estimates of β_0 , β_1 , and γ_0 under this restricted model are essentially the same as reported above. The corresponding estimate of the cumulative mean discovery rate is plotted in Fig. 1. Under this fitted model, the estimated mean introduction rate, which is the quantity of primary interest, rises from around 0.3 introductions per year in 1850 at an annual rate of 1.4% to around 2.3 introductions per year in 1995. The estimated annual observation probability is around 0.19, implying a mean delay of around five years between introduction and discovery.

Under the assumption that $\gamma_1 = \gamma_2 = 0$, we fit the model under the null hypothesis that $\beta_1 = 0$ (i.e., that the rate of introductions has been constant). The corresponding ML estimates of β_0 and γ_0 are 2.2 and -6.4 , respectively, for an estimated mean introduction rate of around 8.6 species per year and an annual observation probability of around 0.002 (i.e., a mean delay of around 600 years). The maximized value of the log likelihood for this model is -126.21 for an approximate P value of around 0.001, so that the null hypothesis of a constant introduction rate can be rejected. Although this model is rejected, the corresponding cumulative mean discovery rate, which is also shown in Fig. 1, captures the overall behavior of the data reasonably well. This underlines the point that the discovery record should not be treated uncritically as a proxy for the introduction record.

Finally, as an illustration of the use of this kind of modeling, the estimate of the total number of species is given by $\sum_{t=1}^m \hat{\mu}_t$, where $\hat{\mu}_t$ is the ML estimate of μ_t . For the final model fit to the data in Fig. 1, this estimate is ~ 150 species. Thus, the estimated number of undiscovered species in 1995 is only ~ 10 .

SOME SIMULATION RESULTS

In this section, we present some results from a simulation study of the performance of maximum-likelihood (ML) estimation under the model outlined above. The study proceeded in the following way. For fixed values of β_0 and β_1 , an introduction record of length m with mean introduction rate (Eq. 6) was simulated. For fixed values of γ_0 , γ_1 , and γ_2 , observations of each simulated introduced species were simulated with probabilities (Eq. 7), and the simulated sighting record was formed (see Supplement). The model was fit to the simulated sighting record and the entire procedure was re-

TABLE 1. Estimated means and standard deviations of maximum-likelihood estimates for selected parameter values.

Simulation result	β_0	β_1	γ_0	γ_1	γ_2
True value	-1.0	0	-1.1	0	0
Mean	-0.96	0.00	-1.18	0.00	0.00
1 SD	0.21	0.003	0.47	0.002	0.002
True value	-1.0	0	-1.1	0.02	0
Mean	-0.98	0.00	-1.14	0.021	0.00
1 SD	0.28	0.003	0.78	0.008	0.003
True value	-1.0	0.02	-1.1	0	0
Mean	-1.00	0.020	-1.07	0.00	0.00
1 SD	0.13	0.002	0.43	0.001	0.001
True value	-1.0	0.02	-1.1	0.02	0
Mean	-0.96	0.00	-1.18	0.00	0.00
1 SD	0.21	0.003	0.32	0.006	0.001

Note: In each case, the sighting record has length $m = 145$ yr, and the means and standard deviations were estimated from 100 simulated sighting records.

peated 100 times. In Table 1, the means and 1 SD of the ML estimates are reported for a small number of cases chosen to be similar to the model fitted to the sighting record from San Francisco. In no case was a significant bias found. Also, the standard deviation of the ML estimate of β_1 , which is the parameter of greatest interest, is small, indicating that this parameter is well estimated.

DISCUSSION

The purpose of this paper has been to describe and illustrate an approach to modeling the aggregate discovery record of introduced species. The approach explicitly incorporates models of the introduction and discovery processes. By doing so, it provides a direct estimate of the introduction process that accounts for the effect of the discovery process.

A key step in applying this general approach is specifying models of the mean introduction rate μ_t and the observation probability π_{st} . In the application described in the previous section, we adopted relatively simple descriptive models. In some situations, it may be possible to use other kinds of information in model specification. For example, if time-series data are available on the main vectors of introduction (e.g., the volume of foreign shipping), then this could be used as a covariate in the model for μ_t . Similarly, if time-series data relating to observational effort (e.g., the number of indigenous species discovered in each year) are available, then this could be used in the model for π_{st} . We are currently exploring extensions along these lines.

ACKNOWLEDGMENTS

The helpful comments of two anonymous reviewers are acknowledged with gratitude.

LITERATURE CITED

Azzalini, A. 1996. Statistical inference. CRC Press, Boca Raton, Florida, USA.

- Cohen, A. N., and J. T. Carlton. 1995. Nonindigenous aquatic species in a United States estuary: a case study of the biological invasions of the San Francisco Bay and Delta. U.S. Fish and Wildlife Service, Washington, D.C., USA.
- Cohen, A. N., and J. T. Carlton. 1998. Accelerating invasion rate in a highly invaded estuary. *Science* **279**:555–558.
- Costello, C. J., and A. R. Solow. 2003. On the pattern of discovery of introduced species. *Proceedings of the National Academy of Sciences (USA)* **100**:3321–3323.
- Pimentel, D., L. Lack, R. Zuniga, and D. Morrison. 2000. Environmental and economic costs of nonindigenous species in the U.S. *BioScience* **50**:53–67.
- Ross, S. M. 2000. *Introduction to probability models*. Academic Press, San Diego, California, USA.
- Wilcove, D., D. Rothstein, J. Dubow, A. Phillips, and A. Losos. 1998. Quantifying threats to imperiled species in the United States. *BioScience* **48**:607–615.
- Williamson, M. 1999. Invasions. *Ecography* **22**:5–12.

SUPPLEMENT

Matlab codes for running the maximum-likelihood estimate, data files for the San Francisco estuary (1851–1995), and simulated records for the simulation exercise are available in ESA's Electronic Data Archive: *Ecological Archives* E085-049-S1.