Estimating the rate of species introductions from the discovery record

Andrew R Solow, Woods Hole Oceanographic Institution, Woods Hole, MA  02543
Christopher J Costello, University of California, Santa Barbara, CA  93106

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. This paper describes 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. The estimated mean rate of introductions increases from 0.3 introductions in 1850 to 2.3 introductions in 1995.

Keywords

maximum likelihood estimation; Poisson distribution; species introductions
1. Introduction

The introduction of non-indigenous species into marine and terrestrial environments can have significant ecological and economic consequences (Wilcove, et al., 1998; Williamson, 1999; Pimental, 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. In Section 3, this model is fit to the discovery record of introduced marine species in the San Francisco estuary. The results of a small simulation study are presented in Section 4. Section 5 contains some concluding remarks.
2. Model

Let the observable random variable $Y_t$ be the number of non-indigenous species that are discovered in year $t$. For economy of expression, the qualifier non-indigenous 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 $\mu_t$ that depends on an unknown, possibly vector-valued parameter. We will also assume that the sequence $N_1, N_2, ...$ is independent. Under this model, any non-random variation in this sequence is due to variation in the mean introduction rate $\mu_t$.

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 introduced in year $s$ that are discovered in year $t$. The probability mass function (pmf) of $Y_{st}$ is:
\[ \text{prob}(Y_{st} = y_{st}) = \sum_{n_s = y_{st}}^\infty \text{prob}(Y_{st} = y_{st} \mid N_s = n_s) \text{prob}(N_s = n_s) \quad (2) \]

The first term in the summation in (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:

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

where \( \pi_{st} \) 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 \( \pi_{st} \) is given below. The second term in the summation in (2) is given by the pmf of \( N_s \), which by assumption is Poisson with mean \( \mu_s \). It is straightforward to show that, under this model, \( Y_{st} \) has a Poisson distribution with mean \( p_{st} \mu_s \) (Ross, 1992). It follows from (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 $\pi$ 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 $\mu$. 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 $\mu_t$ and the observation probability $\pi_{st}$ have been specified, the complete model can be fit to an observed time series of $Y_1, Y_2, ..., Y_m$ by the method of maximum likelihood. The Poisson log likelihood function is:

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

(5)

where $y_t$ is the observed value of $Y_t$ and where $\lambda_t$ is given in (4) and $p_{st}$ given in (3). The ML estimates of the unknown parameters are found by maximizing (5) over the unknown parameters in $\mu_t$ and $\pi_{st}$. This maximization must be done numerically. A good discussion of maximum likelihood (ML) estimation and related methods is contained in Azzalini (1999).

3. 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. These data were compiled by Cohen and Carlton (1995) – see, also, Cohen and Carlton (1998). The data, which are plotted in cumulative form in Figure 1, cover the period 1850-1995. 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 $\mu_t$ and $\pi_{st}$. 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 $\beta_0$ and $\beta_1$ are unknown parameters. As discussed in Costello and Solow (2003), $\pi_{st}$ 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:

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

where $\gamma_o$, $\gamma_1$, and $\gamma_2$ 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 ML estimates of the parameters of the full model are:

\[
\hat{\beta}_o = -1.1 \quad \hat{\beta}_1 = 0.014 \quad \hat{\gamma}_o = -1.46 \quad \hat{\gamma}_1 = 0.00001 \quad \hat{\gamma}_2 = 0.0000004
\]

and the maximized value of the log likelihood is -122.56. The point estimates of \( \gamma_1 \) and \( \gamma_2 \) 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 \( \beta_o, \beta_1, \) and \( \gamma_o \) under this restricted model are essentially the same as reported above. The corresponding estimate of the cumulative mean discovery rate is plotted in Figure 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 5 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 \( \beta_o \) and \( \gamma_o \) 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 Figure 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 \(\mu_t\). For the final model fit the data in Figure 1, this estimate is around 150 species. Thus, the estimated number of undiscovered species in 1995 is only around 10.

### 4. Some simulation results

In this section, we present some results from a simulation study of the performance of ML estimation under the model outlined above. The study proceeded in the following way. For fixed values of \(\beta_o\) and \(\beta_1\), an introduction record of length \(m\) with mean introduction rate (6) was simulated. For fixed values of \(\gamma_o, \gamma_1,\) and \(\gamma_2\), observations of each simulated introduced species were simulated with probabilities (7) and the simulated sighting record was formed. The model was fit to the simulated sighting record and the entire procedure was repeated 100 times. In Table 1, the means and standard deviations 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 $\beta_1$, which is the parameter of greatest interest, is small, indicating that this parameter is well-estimated.

4. 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 $\mu_t$ and the observation probability $\pi_{st}$. 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 $\mu_t$. 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 used in the model for $\pi_{st}$. We are currently exploring extensions along these lines.
Acknowledgements

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


Table 1

Estimated means and standard deviations of ML estimates for selected parameter values.
In each case, the sighting record has length \( m = 145 \) and the means and standard
deviations were estimated from 100 simulated sighting records.

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \gamma_o )</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>-1.0</td>
<td>0</td>
<td>-1.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mean</td>
<td>-0.96</td>
<td>0.00</td>
<td>-1.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.21</td>
<td>0.003</td>
<td>0.47</td>
<td>0.002</td>
<td>0.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \gamma_o )</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>-1.0</td>
<td>0</td>
<td>-1.1</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>mean</td>
<td>-0.98</td>
<td>0.00</td>
<td>-1.14</td>
<td>0.021</td>
<td>0.00</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.28</td>
<td>0.003</td>
<td>0.78</td>
<td>0.008</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \gamma_o )</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>-1.0</td>
<td>0.02</td>
<td>-1.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mean</td>
<td>-1.00</td>
<td>0.020</td>
<td>-1.07</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.13</td>
<td>0.002</td>
<td>0.43</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \gamma_o )</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true value</td>
<td>-1.0</td>
<td>0.02</td>
<td>-1.1</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>mean</td>
<td>-0.96</td>
<td>0.00</td>
<td>-1.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.21</td>
<td>0.003</td>
<td>0.32</td>
<td>0.006</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Figure 1. The cumulative record of discoveries of introduced species in San Francisco estuary, 1850-1995 (solid). Also shown are fitted values allowing for an increasing introduction rate (dashed) and assuming a constant introduction rate (dotted).