Capitalising on Opportunistic Data for Monitoring Biodiversity

Christophe Giraud, Clément Calenge, Romain Julliard

Running title: Capitalising on opportunistic data

Number of words: ∼8600

Date of submission: July 9, 2014

Capitalising on Opportunistic Data for Monitoring Biodiversity

Christophe Giraud, Clément Calenge & Romain Julliard

1. *C. Giraud, CMAP, UMR 7641, Ecole Polytechnique, France – Laboratoire de Mathématiques d'Orsay, UMR 8628, Université Paris-Sud, France.*

2. *C. Calenge (*clement.calenge@oncfs.gouv.fr*), Office national de la chasse et de la faune sauvage, Direction des études et de la recherche, Saint Benoist, BP 20. 78612 Le Perray en Yvelines, France.*

3. *R. Julliard, CESCO, UMR 7204, MNHN-CNRS-UPMC, CP51, 55 rue Buffon, 75005 Paris, France.*

Abstract

With the internet, a massive amount of information on species abundance can be collected under citizen science programs. However, these data are often difficult to use directly in statistical inference, as the data collection is generally opportunistic under such programs, and the distribution of the sampling effort is often not known. In this paper, we developed a statistical framework to combine such "opportunistic data" with data collected using schemes characterized by a known sampling effort. We illustrated the framework with typical bird datasets from the Aquitaine region, south-western France. We demonstrated that such a framework can provide estimates that are always more precise than the ones obtained from the dataset with a known sampling effort alone. The gain in precision may be considerable if the opportunistic data are abundant. We also show that estimates could be obtained even for species recorded only in the opportunistic scheme. Opportunistic data combined with a relatively small amount of data collected with a known effort may thus provide access to precise estimations of quantitative changes in abundance. This should significantly change the organisation of large scale monitoring schemes, particularly for the rarer species. The framework can be readily used to monitor temporal changes but with more restrictive conditions for monitoring spatial changes. The framework presented in this paper will be improved in the future to allow a more easy application to the estimation of the spatial distribution of a species.

Keywords: carnivorous species, opportunistic data, species distribution map, sampling effort, detection probability

1 Introduction

How species abundance varies in space and time is a major issue both for basic (biogeography, macroecology) and applied (production of biodiversity state indicators) ecology. Professionals working on biodiversity thus spend considerable resources collecting data that are suitable for estimating this variation (Yoccoz et al., 2001). Most of the scientific literature recommends the implementation of both a statistically valid sampling design and a standardised protocol for collecting such data (Williams et al., 2002). Many methods have been developed to estimate species abundance in a defined location, e.g., using mark-recapture methods (Seber, 1982) or distance sampling approaches (Buckland et al., 1993). However, these approaches require an intense sampling effort and are not always practical. Many authors have noted that most frequently, interest will not be in abundance itself, but either in the rate of population change, i.e., the ratio of abundance in the same location at two different time points, or in the relative abundance, i.e., the ratio of abundance at two separate locations (MacKenzie and Kendall, 2002).

Relative abundance is frequently monitored with the help of simpler schemes. For instance, a set of sites is randomly sampled in the area of interest, and counts of organisms are organised on these sites using a given protocole. At a given location, the resulting count can be used as an index of the true abundance. Indeed, assuming constant detectability over space and time, the average number of animals counted per sampled site is proportional to the true abundance of the species in the area (MacKenzie and Kendall, 2002). Log-linear models can be used to represent this average number of animals detected per site as a function of space and/or time (and, possibly, other factors such as the habitat; see for example van Strien and Pannekoek, 2001), and thereby, to infer population trends. Thus, such programs have been implemented in many countries to monitor the changes in the abundance of several groups of species, such as birds (e.g., Breeding Bird Survey, see Julliard et al., 2004) or butterflies (e.g., Butterfly Monitoring Scheme van Swaay et al., 2008). Estimates of relative abundance have also been commonly used for mapping the spatial distribution of several species (Gibbons et al., 2007).

Note that in many cases, it is practically impossible to control the effort put into taking the sample (called "sampling effort" or "observational effort" in this paper). This generally occurs when the spatial scale of interest is large and when volunteers collect the data (e.g. the Breeding Bird Survey, see below). In such cases, the effort is generally important in places where the volunteers are more numerous. However, it is possible to account for an unequal sampling effort in the estimation of the relative abundance of the species of interest under the assumption that the size of a sample caught from a population is proportional to this effort, all other things being equal. An unbiased estimation of the relative abundance can be obtained by including the logarithm of this known effort as an offset in the log-linear model used for estimation. Such "catch-effort" approaches are often used to estimate fish stock in fisheries (e.g. see Bishop et al., 2004; Maunder and Punt, 2004).

In addition to such data characterized by a known sampling effort, a large amount of data can also be collected by non-standardised means, with no sampling design and no standardised protocol. In particular, the distribution of the observers and of their sampling effort is often unknown (Dickinson et al., 2010). These so-called "opportunistic data" have always existed, and with the recent development of citizen science programs, we observe a massive increase in the collection of these data on a growing number of species (e.g., Dickinson et al., 2010; Hochachka et al., 2012; Dickinson et al., 2012). Additionally, as the use of online databases facilitates the exchange and storage of data, such opportunistic data may now include millions of new observations per year that are collected in areas covering hundreds of thousands of square kilometres (e.g., the global biodiversity information facility, including more than 400 million records at the time of writing, see Yesson et al., 2007).

The temporal and spatial distributions of the observations in such data reflect unknown distributions of both observational efforts and biodiversity. Thus, a report of a high number of individuals of a given species at a given location compared to other locations could be because the focus species is abundant at this location or because numerous observers were present at this location. Using such opportunistic data to estimate variation in the space and time of species abundance is therefore complex, since any modelling approach should include a submodel of the observation process (Kéry et al., 2009; Hochachka et al., 2012) or an attempt to manipulate the data to remove the bias caused by unequal effort (down-weighting the records in oversampled regions, manipulating background data, etc.; see a discussion in Phillips et al., 2009).

MacKenzie et al. (2005) noted that "In some situations, it may be appropriate to share or borrow information about population parameters for rare species from multiple data sources. The general concept is that by combining the data, where appropriate, more precise estimates of the parameters may be obtained." In this paper, we adopt this strategy and develop a general framework for using data characterized with a known sampling effort and data collected with an unknown effort simultaneously. We focus on multi-species and multi-site data that correspond to the data typically collected in this context.

The purpose of this study is to estimate the relative densities of the species at different sites. We base this estimation on two datasets containing the number of animals detected by observers for each species of a pool of species of interest and each spatial unit of a study area of interest: (i) one dataset is collected under a program characterized by a known sampling effort, possibly variable among spatial units, (ii) one "opportunistic" dataset characterized by a completely unknown sampling effort. For simplicity, we assume as a first step that the detectability of any given species is constant (i.e., no observer effects, no habitat-related detectability, etc., but see the discussion). However, we suppose that this detectability is varying across species. We show that, under this assumption, the information concerning both the distribution of the observational effort and the biodiversity can be efficiently retrieved from "opportunistic" data by combining them with "known-effort" data. Moreover, we prove that such a combination returns more precise estimates than when using the "known-effort" data alone. Our statistical framework allowing this win-win combination can open numerous avenues for application. We utilized data on French birds, which are typical of existing data, to illustrate the numerous qualities of this framework. Note however that the work presented in this paper is preliminary, and that further work will be required to account for a detectability depending on the habitat or on the observers.

2 Single scheme statistical modelling

2.1 Statistical framework

We want to estimate the density (number of individuals per unit of surface area) of I species in J sites. We emphasise that the "sites" j can either refer to different spatial sites, to different times, or to different combinations of sites and times. Let A_{ij} be the true density of individuals of a species i at site j. We suppose that a study has been conducted and that each species has been counted at each site. Let X_{ij} be the resulting count of individuals of the species i by the observers in the site j. We further assume that there is no identification error in these data.

As noted in the introduction section, we allow the observational effort to vary among sites. Indeed, depending on the type of study carried out, the sites may be characterised by a variable number of observers, a variable number of traps, different transect lengths, different observational durations, etc. The observational effort may be a (possibly complex) function of all these elements. We note E_j, the observational effort in site j (assumed identical for all species).

Moreover, we allow for a variable detectability among species. Indeed, depending on the type of study, some species may be more conspicuous than others, some more easily trapped, etc. It is therefore important to take this point into account by including a species detection probability P_i in our model. In the particular case of opportunistic data, this probability also includes the reporting rate, which is the probability that a detected species will be reported by the observer. The reporting rate reflects the fact that the attention of the observer may systematically vary among species. We suppose that the detectability of a given species is the same in all sites.

Our statistical framework supposes that the observed number X_{ij} is the outcome of a sampling process of intensity $A_{ij}E_jP_i$. For example, X_{ij} can be modelled by a Poisson distribution of intensity $A_{ij}E_jP_i$, but other distributions are possible, such as quasi-Poisson or negative binomial distributions.

2.2 Identifiability

The individual estimation of the $IJ + I + J$ parameters A_{ij}, E_j and P_i is impossible with the sole IJ values of X_{ij}. Additional information is therefore required to enable this estimation.

First, we consider the case of schemes where the effort E_j is known up to a (possibly unknown)

multiplicative constant c. In such studies, one can estimate the IJ parameters $\widetilde{A}_{ij} = A_{ij}P_i/c$ by the ratio $X_{ij}/(cE_j)$. The quantity $\widetilde{A}_{ij}$ is equal to the density A_{ij} times a (unknown) constant depending only on the species i. Therefore, the parameters $\widetilde{A}_{ij}$ give access to the ratio of densities between two sites $A_{ij}/A_{ij'} = \widetilde{A}_{ij}/\widetilde{A}_{ij'}$ for all species i and all pairs of sites j, j'. In general, we choose a reference site, say $j' = 1$, and we focus on the ratios $A_{ij}/A_{i1} = \widetilde{A}_{ij}/\widetilde{A}_{i1}$ which are called the relative densities. Such schemes then enable us to estimate the relative density of the monitored species. As noted previously, accounting for unequal observational effort using such multiplicative models is very common to estimate the fish stock in fisheries research (Maunder and Punt, 2004). Note that formally, we can write the intensity $A_{ij}E_jP_i$ in the form $\widetilde{A}_{ij}\widetilde{E}_j\widetilde{P}_i$ where $\widetilde{E}_j = cE_j$ and $\widetilde{P}_i = 1$. The parameters $\widetilde{E}_j$ and $\widetilde{P}_i$ are known, we then only have to estimate the IJ parameters $\widetilde{A}_{ij}$ from the IJ observations X_{ij}.

However, in the case of schemes characterized by an unknown effort E_j, the same reasoning as above would lead to estimating the $IJ + J$ parameters $\widetilde{A}_{ij} = A_{ij}P_i$ and E_j with the sole IJ observations X_{ij}, which is impossible. To overcome this issue, we propose in the next section to combine the "unknown effort" data with the "known effort" monitoring data.

3 Double scheme statistical modelling

3.1 Statistical framework

Suppose we have two datasets collected via two schemes that both aim to estimate the same density A_{ij}. These two datasets are labelled by $k \in \{0, 1\}$. The efforts and the detection probabilities are usually not the same in the two datasets, so we write E_{jk} (respectively P_{ik}) for the effort for the site j (resp. probability detection of the species i) in the dataset k. Therefore, under the same assumptions as above, the count X_{ijk} for the species i at site j in the dataset k can be viewed as the outcome of a sampling process of intensity $A_{ij}E_{jk}P_{ik}$. For simplicity, we will assume in this paper that

$$\begin{aligned} X_{ijk} &\sim \text{Poisson}(A_{ij}E_{jk}P_{ik}), \\ \text{for } i = 1, \ldots, I,\ j = 1, \ldots, J &\text{ and } k = 0, 1. \end{aligned} \tag{1}$$

However, we stress that if overdispersion is present in the data, this Poisson distribution could be replaced by a quasi-Poisson process or by a negative binomial process. We consider below that the dataset $k = 0$ is collected by a monitoring scheme with a known sampling effort while the dataset $k = 1$ corresponds to a scheme characterized by an unknown sampling effort. We then investigate how one can estimate each parameter involved in the model (1) and to what extent adding the dataset $k = 1$ would enhance the precision of the density estimation.

3.2 Identifiability

In the above setting, there are $2IJ$ observations for $IJ + 2(I + J)$ parameters. For $IJ > 2(I + J)$, which typically holds for large J and $I \geq 3$, we have more observations than parameters. Nevertheless, as explained in the web appendix A, the model is not identifiable without $J + I + 1$ additional identifiability conditions. As in Section 2.2, we introduce some renormalisation $\widetilde{A}_{ij}$, $\widetilde{E}_{jk}$ of $\widetilde{P}_{ik}$ of A_{ij}, E_{jk} and P_{ik}, which enables us to easily express these identifiability conditions while preserving the identity $\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik} = A_{ij}E_{jk}P_{ik}$.

We assume henceforth that the effort E_{j0} is known up to a multiplicative constant c for the dataset $k = 0$ and that the effort E_{j1} is unknown for $k = 1$. As in Section 2.2, we define $\widetilde{E}_{j0} = cE_{j0}$ (which is known) and $\widetilde{P}_{i1} = 1$ for all i. We could have set $\widetilde{P}_{i0} = 1$ instead of $\widetilde{P}_{i1} = 1$, but the latter choice is more suited for handling species i monitored in the dataset $k = 1$ but not in the dataset $k = 0$, as we will show later. We must still set one more constraint. We choose $\widetilde{P}_{10} = 1$ for convenience. These $I + J + 1$

constraints combined with the identity $\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik} = A_{ij}E_{jk}P_{ik}$ lead to the change of variables:

$$\begin{aligned} \widetilde{A}_{ij} &= A_{ij}P_{i1}P_{10}/(cP_{11}), \\ \widetilde{E}_{jk} &= cE_{jk}P_{1k}/P_{10} \\ \widetilde{P}_{ik} &= P_{ik}P_{11}/(P_{i1}P_{1k}). \end{aligned} \tag{2}$$

In terms of these new variables, we have the simple statistical model $X_{ijk} \sim \text{Poisson}(\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik})$ with $\widetilde{E}_{j0} = cE_{j0}$ for all j, $\widetilde{P}_{i1} = 1$ for all i and $\widetilde{P}_{10} = 1$. These $J+I+1$ quantities are known, and the resulting statistical model is identifiable.

Let us interpret these new quantities. The parameter $\widetilde{A}_{ij}$ is proportional to the density A_{ij} by an unknown factor $P_{i1}P_{10}/(cP_{11})$ depending only on the species i. As in Section 2.2, this parameter gives access to the relative density of each species i in each site j. The parameter $\widetilde{E}_{j1}$ is equal, up to a constant, to the effort E_{j1}; therefore, it provides the relative effort for each site j in the dataset 1. Finally, $\widetilde{P}_{i0}$ is proportional to the ratio P_{i0}/P_{i1} by an unknown factor P_{11}/P_{10}, so we can compare the ratios P_{i0}/P_{i1} across the different species. The ratio P_{i0}/P_{i1} reflects the systematic difference of attention toward some species among the observers of the two schemes.

In addition, we emphasise that we can consider the case where some species i are not monitored in the dataset 0 but are recorded in the dataset 1. This case can be handled by merely adding the constraints $\widetilde{P}_{i0} = P_{i0} = 0$ for the concerned species i.

3.3 Estimation

We can estimate the parameters $\widetilde{A}_{ij}$, $\widetilde{E}_{jk}$ and $\widetilde{P}_{ik}$ by the maximum likelihood estimators $(\widehat{A}_{ij}, \widehat{E}_{jk}, \widehat{P}_{ik})$ with the constraints $\widehat{E}_{j0} = \widetilde{E}_{j0}$ for all j, $\widehat{P}_{i1} = 1$ for all i and $\widehat{P}_{10} = 1$. This estimation can be carried out with the help of a generalised linear model. Indeed, with the notations $a_{ij} = \log(\widetilde{A}_{ij})$, $e_{jk} = \log(\widetilde{E}_{jk})$ and $p_{ik} = \log(\widetilde{P}_{ik})$, the model (1) can be recast as a classical generalised linear model from the Poisson family with a log link:

$$X_{ijk} \sim \text{Poisson}(\lambda_{ijk}), \quad \text{with } \log(\lambda_{ijk}) = a_{ij} + e_{jk} + p_{ik}.$$

Indeed, we only have to define $e_{j0} = \log \widetilde{E}_{j0}$ as an offset in the model, $p_{i1} = 0$ for all i, and fit the resulting model with any statistical package.

3.4 Improvement resulting of the combination of the two datasets

It is now important to determine whether combining the dataset 1 with unknown effort E_{j1} to the dataset 0 with known effort improves the estimation of the relative densities, when compared to the estimations obtained with the single dataset 0. An improvement is expected simply by looking at the balance between the number of observations and the number of free parameters. With the dataset 0, we have IJ observations, and we want to estimate IJ free parameters; whereas with the two datasets 0 and 1, we have $2IJ$ observations for $IJ + J + I - 1$ free parameters. The balance between the number of observations and the number of free parameters is better in the second case. Below, we quantify the improvement more precisely.

For simplicity, we assume in the following that the ratios P_{i0}/P_{i1} are known for all i. In terms of the normalised variables, this means that the $\widetilde{P}_{i0}$ are known. When we work with the single dataset 0, we estimate $\widetilde{A}_{ij}$ by the maximum likelihood estimator $\widehat{A}^0_{ij} = X_{ij0}/(\widetilde{E}_{j0}\widetilde{P}_{i0})$ whose variance is $\text{var}(\widehat{A}^0_{ij}) = \widetilde{A}_{ij}/(\widetilde{E}_{j0}\widetilde{P}_{i0})$.

We now investigate how the maximum likelihood estimator $\widehat{A}_{ij}$ associated with the model $X_{ijk} \sim \text{Poisson}(\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik})$ improves upon $\widehat{A}^0_{ij}$. We consider the case where the (unknown) effort E_{j1} in the dataset 1 is much larger than the effort E_{j0} in the dataset 0. In the web appendix A, we show that the limit variance of $\widehat{A}_{ij}$ when $E_{j1} \gg 1$ is given by

$$\text{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\to} \text{var}(\widehat{A}^0_{ij}) \times \frac{P_{i0}A_{ij}}{\sum_l P_{l0}A_{lj}}. \tag{3}$$

In particular, the variance of the estimation is reduced by a factor

$$\frac{\mathrm{var}(\widehat{A}_{ij})}{\mathrm{var}(\widehat{A}^0_{ij})} \overset{E_{j1}\gg 1}{\sim} \frac{P_{i0}A_{ij}}{\sum_l P_{l0}A_{lj}},$$

when working with the two datasets instead of the sole dataset 0.

In the simple casewhere the ratios P_{i0}/P_{i1} are the same for all the species i (which formally corresponds to $\widetilde{P}_{i0} = 1$ for all i), we have a closed-form formula for $\widehat{A}_{ij}$ (see Formula (2) in the web appendix A)

$$\widehat{A}_{ij} = \frac{X_{ij0} + X_{ij1}}{\sum_l (X_{ij0} + X_{lj1})} \times \frac{\sum_l X_{lj0}}{\widetilde{E}_{j0}}.$$

This formula reveals the contribution of each dataset to the estimation of the (normalised) density. Actually, the estimator $\widehat{A}_{ij}$ is the product of two terms, where the first term mainly depends on the opportunistic dataset 1 when the observational effort E_{j1} is large, whereas the second term only depends on the dataset 0

$$\widehat{A}_{ij} \overset{E_{j1}\gg 1}{\approx} \frac{X_{ij1}}{\sum_l X_{lj1}} \times \frac{\sum_l X_{lj0}}{\widetilde{E}_{j0}}.$$

Let us interpret these two terms. The ratio on the left-hand side provides an estimation of the proportion $\widetilde{A}_{ij}/\sum_l \widetilde{A}_{lj}$ of individuals in a site j that belong to a species i. This proportion is estimated by the ratio of the number X_{ij1} of individuals of the species i observed at site j in the opportunistic dataset to the total number $\sum_l X_{lj1}$ of individuals observed at site j in the same data. When the observational effort E_{j1} in the opportunistic dataset 1 is large, the ratio $X_{ij1}/\sum_l X_{lj1}$ provides a very accurate estimation of the density proportion $\widetilde{A}_{ij}/\sum_l \widetilde{A}_{lj}$, and we have (see Formula (3) in the web appendix A)

$$\widehat{A}_{ij} \overset{E_{j1}\gg 1}{\sim} \frac{\widetilde{A}_{ij}}{\sum_l \widetilde{A}_{lj}} \times \frac{\sum_l X_{lj0}}{\widetilde{E}_{j0}}. \tag{4}$$

The term on the right-hand side provides an estimation of the total (normalised) density $\sum_l \widetilde{A}_{lj}$ at the site j. This total (normalised) density is estimated from the dataset 0 by dividing the total number $\sum_l X_{lj0}$ of individuals counted at the site j in the dataset 0 by the (normalised) observational effort $\widetilde{E}_{j0}$. Let us now explain the reduction of variance observed in (3). The formula (4) shows that we estimate $\widetilde{A}_{ij}$ by first estimating the total (normalised) density $\sum_l \widetilde{A}_{lj}$ with the dataset 0 and then renormalise this estimation with the ratio $\widetilde{A}_{ij}/\sum_l \widetilde{A}_{lj}$ which has been accurately estimated with the dataset 1. The reduction of variance observed in (3) then results from the use of the whole counts $\sum_l X_{lj0}$ at site j in the dataset 0 for estimating $\widetilde{A}_{ij}$ instead of the sole counts X_{ij0} of the species i at site j.

3.5 Species not monitored in the scheme characterized by a known sampling effort

As already mentioned, combining the two datasets also allows to estimate $\widetilde{A}_{ij}$ for some species i that are not monitored in the dataset 0, but are monitored in the opportunistic dataset 1. This situation formally corresponds to the case where $P_{i0} = 0$. For $E_{j1} \gg 1$, the limit variance of the estimator $\widehat{A}_{ij}$ is (see Formula (6) in the web appendix A)

$$\mathrm{var}(\widehat{A}_{ij}) \overset{E_j\gg 1}{\sim} \frac{\widetilde{A}^2_{ij}}{\sum_l \widetilde{P}_{l0}\widetilde{A}_{lj}\widetilde{E}_{j0}}.$$

Because the species i is not monitored in dataset 0, the density $\widetilde{A}_{ij}$ cannot be estimated with the sole dataset 0. Thus, there is an obvious improvement to be made by using our estimation scheme that combines the two datasets. To reveal the power of our approach, let us compare the variance $\mathrm{var}(\widehat{A}_{ij})$ of our density estimator with the variance of the virtual estimator $\widehat{A}^{0,\mathrm{virtual}}_{ij}$ based on a virtual dataset 0 where the species i would have been monitored with some (virtual) probability detection $P^{\mathrm{virtual}}_{i0}$. The variance of the maximum likelihood estimation $\widehat{A}^{0,\mathrm{virtual}}_{ij}$ of $\widetilde{A}_{ij}$ with this virtual dataset 0 would be $\widetilde{A}_{ij}/(\widetilde{E}_{j0}\widetilde{P}^{\mathrm{virtual}}_{i0})$ so that

$$\mathrm{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\sim} \mathrm{var}(\widehat{A}^{0,\mathrm{virtual}}_{ij}) \times \frac{P^{\mathrm{virtual}}_{i0}A_{ij}}{\sum_l P_{l0}A_{lj}}.$$

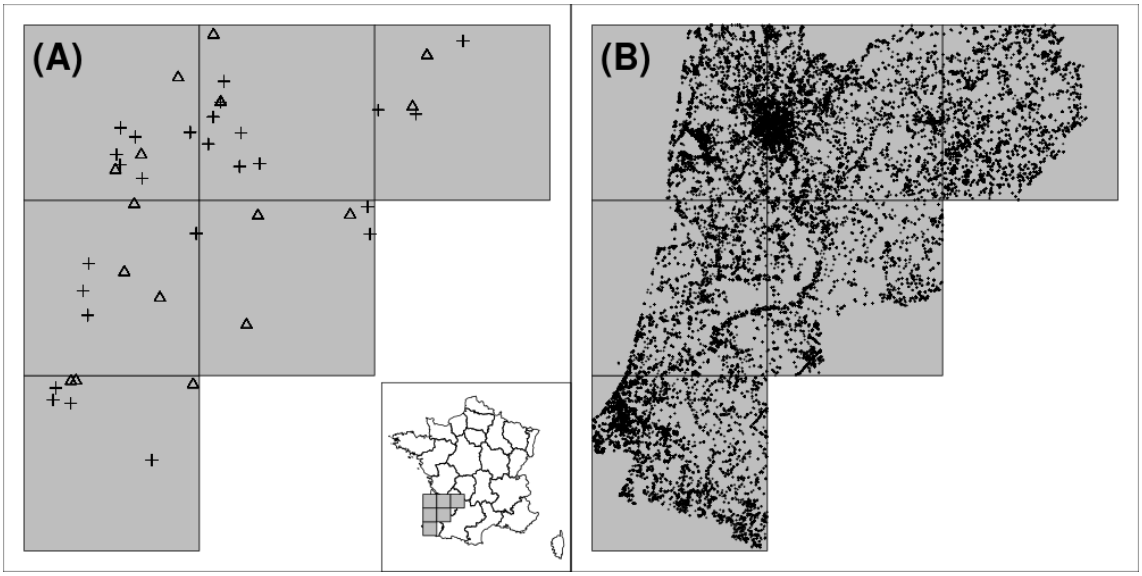

Figure 1: The datasets used to illustrate our statistical framework. The location of the Aquitaine region in France is displayed in the insert. (A) distribution of the FBBS (French Bird Breeding Survey) squares in the region. Triangle-shaped points define the reference dataset, and cross-shaped points define the dataset used for the model fit. (B) distribution of the LPO records (opportunistic dataset) in the region. Grey squares are 100×100 km.

In particular, the estimation provided by $\widehat{A}_{ij}$ can significantly outperform the virtual estimation we would have obtained with the sole virtual dataset 0 (where the species i would have been monitored). Moreover, if we compare the estimator $\widehat{A}_{ij}$ with the virtual estimator $\widehat{A}_{ij}^{\text{virtual}}$ based on both the virtual dataset $k = 0$ and the dataset $k = 1$, we observe that the ratio of their variance

$$\frac{\text{var}(\widehat{A}_{ij})}{\text{var}(\widehat{A}_{ij}^{\text{virtual}})} = \frac{P_{i0}^{\text{virtual}} A_{ij} + \sum_l P_{l0} A_{lj}}{\sum_l P_{l0} A_{lj}}$$

remains close to one when $P_{i0}^{\text{virtual}} A_{ij} \ll \sum_l P_{l0} A_{lj}$. This means that with our estimation scheme, there is not much difference between the estimation based on a dataset collected with known effort where a species i is rare and the estimation based on a dataset collected with known effort where a species i is not monitored. In other words, there is no instability on the estimation of the density of a species when it is not present in the dataset collected with known effort.

4 Illustration

4.1 Datasets

Our aim in this section is to test whether combining an opportunistic dataset collected with unknown effort, with a dataset collected by a monitoring scheme with known effort significantly improves accuracy in the prediction of relative density variations. We explore this issue with two typical datasets having the same spatial and temporal coverage. We demonstrate the ability of our statistical approach to provide better predictions of spatial variation of species density than either of the two datasets alone.

We first describe the opportunistic dataset. We used the recent online database developed by the Ligue de Protection des Oiseaux (LPO, Bird Life representative in France, largest French bird watcher NGO, with regional delegations). This online system was launched successively by the different regional LPO groups, and we acquired data from one of the first groups to start, Aquitaine, South-Western France, with data collection starting in 2007 (www.fauneaquitaine.org). Any citizen who can identify bird species can register on this website and record any bird observation s/he wishes, noting the species, date, and location (to the nearest 500 m). Hundreds of observers thus record hundreds of thousands observations.

We typically ignore why these observations were made, e.g., the motivation of the observer, the reasoning for choosing to report these observations over others, whether they report all the species they have seen at a given place and time, the underlying observation effort, etc. We selected all such opportunistic records between April and mid-June 2007-2011. For each record, we only considered the occurrence of a species, and not the number of animals detected by the observer. Data were pooled over a year, because we will focus here only on spatial variation in density. Over 153 000 species records were considered in this study (see Fig. 1B).

We now describe the dataset collected with known effort. We used the data from the French Breeding Bird Survey (FBBS) for the same region and the same years. The FBBS (Jiguet et al., 2012) is based on a stratified random sampling, with each volunteer observer being assigned a 2×2 km square randomly chosen within 10 km of his house. The observer then homogeneously distributed 10 points within the square. Each point was visited twice between April and mid-June (before and after May 8th, with at least 4 weeks between visits) for exactly 5 minutes within 4 hours after sunrise in appropriate weather conditions (no rain or strong winds). Every bird heard or seen was recorded, and for each point and each species, the maximum count among the two visits was retained. These counts were then summed for a given square, year and species. Fifty-one such squares have been surveyed, most of them for several years. For the purpose of analysis, we used one year of record per square, randomly sampled among those available.

Our aim was to test our model's ability to provide a better estimation of the spatial variation in species' density than either of the two datasets alone. The "sites" of our model were defined as cells (much larger than a FBBS square) from a regular grid superimposed onto the Aquitaine region. We chose a grid size and position to maximise the number of cells that contained at least 6 FBBS squares. The best compromise for a regular square grid was a 100×100 km grid with 6 cells containing 6 to 11 FBBS squares (see Fig. 1A); 3 out of 51 squares were located outside the grid and were thus excluded from this analysis. Therefore, our study focused on $J = 6$ sites. In each site, the known observational effort corresponded to the number of FBBS squares. We stress that our aim here is only to illustrate our main point, i.e. that combining the two datasets lead to more precise estimates than the analysis of the datasets considered separately. If our aim was only to estimate the relative density of bird species to achieve a biological conservation purpose, we would define much smaller sites (presently, such sites are too large for our inference to be practically useful). However, the definition of such large sites was required to allow the definition of a "reference dataset" used to evaluate the predictive capacity of our model.

Thus, we split the FBBS dataset into two parts: one dataset M was used for the model fit (jointly with the LPO dataset), and one dataset R was used as a reference to evaluate the predictive capabilities of our model. The dataset R was built by randomly sampling 3 FBBS squares in each site. Therefore, the observational effort was constant across sites in the dataset R. The dataset M consisted of the remaining FBBS squares. Finally, we note L denote the LPO dataset. We modelled the relative density of the species that were present in all the sites, in all the datasets. We focused on $I = 22$ bird species (see Table 1 for a list of the species considered).

4.2 Comparison of the predictive efficiency

We first estimated the spatial variation of the density of each species among sites for each dataset separately. For each species i, and each pair of sites j, j' (with $j \neq j'$), we calculated the following ratios:

Table 1: Estimates and standard errors of the relative detection probability $\widetilde{P}_{i0}$ for all species i, relative to the European greenfinch *Carduelis chloris*: The larger the estimate, the smaller the attention of the observers in the opportunistic dataset. The species are ordered by decreasing relative attention in the opportunistic dataset. The more abundant and widespread species tend to be relatively less reported than the more localised, solitary and territorial ones.

Latin name	Species	Estimate	Standard Error
Phoenicurus ochruros	black redstar	-0.50	0.38
Hippolais polyglotta	melodious warbler	-0.32	0.42
Sitta europaea	Eurasian nuthatch	-0.29	0.44
Garrulus glandarius	Eurasian jay	-0.29	0.39
Dendrocopos major	great spotted woodpecker	-0.25	0.41
Certhia brachydactyla	short-toed treecreeper	-0.12	0.40
Phylloscopus collybita	common chiffchaff	-0.05	0.33
Carduelis chloris	European greenfinch	0	—
Cianistes caeruleus	blue tit	0.00	0.34
Erithacus rubecula	European robin	0.07	0.34
Hirundo rustica	barn swallow	0.10	0.31
Parus major	great tit	0.10	0.31
Turdus philomelos	song thrush	0.12	0.35
Columba palumbus	common wood pigeon	0.12	0.34
Turdus merula	common blackbird	0.23	0.29
Corvus corone	carrion crow	0.23	0.30
Fringilla coelebs	common chaffinch	0.24	0.30
Sylvia atricapilla	Eurasian blackcap	0.27	0.30
Troglodytes troglodytes	Eurasian wren	0.40	0.31
Streptopelia decaocto	Eurasian collared dove	0.53	0.30
Sturnus vulgaris	common starling	1.02	0.28
Passer domesticus	house sparrow	1.19	0.27

$$
\begin{aligned}
R_R(i,j,j') &= \frac{X^R_{ij}}{X^R_{ij'}} \\
R_M(i,j,j') &= \frac{X^M_{ij}/E_j}{X^M_{ij'}/E_{j'}} \\
R_{L1}(i,j,j') &= \frac{X^L_{ij}/a_j}{X^L_{ij'}/a_{j'}} \\
R_{L2}(i,j,j') &= \frac{X^L_{ij}/\sum_i X^L_{ij}}{X^L_{ij'}/\sum_i X^L_{ij'}}
\end{aligned}
$$

where E_j is the number of FBBS squares present in the site j in dataset M, a_j is the area of the site (determined by intersecting each cell of the grid with the Aquitaine region), and X^D_{ij} represents the number of individuals of the species i in the site j, in the dataset D. For the LPO dataset, we had to account for the site-specific unknown effort. We estimated this effort with two proxies that are commonly used in such cases. First, we assumed that effort was spatially uniform so that it varied only with cell area a_j (the resulting ratio is labelled L1). Another proxy considered that the effort within a site was proportional to the total number of records across the sites (pooled over all species; the resulting ratio is labelled L2).

Finally, we fitted the model described in the previous sections, using the dataset M as the dataset collected with known effort ($k = 0$), and the dataset L as the opportunistic dataset ($k = 1$). Note that we supposed a quasi-Poisson distribution, to account for moderate overdispersion in our dataset. We were therefore able to estimate the ratio between the estimated densities:

$$
R_S(i,j,j') = \frac{\widehat{A}_{ij}}{\widehat{A}_{ij'}}
$$

We hypothesised that the estimates obtained by our model would be closer to the true densities than any of the estimates that could be obtained from the individual datasets. In other words, the ratio R_S should be closer to the reference ratio R_R than to any other ratio. To verify this hypothesis, we calculated the correlation coefficient between the logarithm of each ratio $\{R_M, R_{L1}, R_{L2}, R_S\}$ and the logarithm of the ratio R_R, for all species i and all pairs j, j'. We also calculated the variance, over all species i and all pairs j, j' of the quantities:

$$
D_D(i,j,j') = \log R_D(i,j,j') - \log R_R(i,j,j')
$$

When the relative densities estimated with a given dataset D are identical to the relative densities estimated with the reference dataset R, the variance of D_D is equal to zero. The online supplementary material contains the code for the R software (R Core Team, 2013) that will allow the reader to reproduce our calculations.

4.3 Results

Our statistical framework was more efficient than any other dataset (either the "known-effort" or the "unknown-effort" scheme alone) in estimating the relative density of the 22 bird species in the 6 sites of interest, as indicated by both the correlation coefficients and the variance of D_D (Table 2). This finding confirms that our statistical framework provided a better estimation of the spatial variation of species density than either of the two datasets alone. It is noteworthy that the hypothesis that the unknown effort for the opportunistic dataset L would be constant was clearly not verified: the correlation between R_{L1} and the reference ratio was essentially zero. The common strategy consisting to standardise the number of detections of a given species by the total number of detections (all species pooled, i.e., dataset L2) provided slightly closer estimates of the true values than the estimates that relied on the "known-effort" dataset M, but still more different from the true values than our approach combining the two types of data.

Table 2: Predictive capabilities of the various possible approaches to estimate the relative densities of 22 bird species in 6 sites in the Aquitaine region. For each possible dataset D, we present the correlation coefficient (r) between the ratio R_D and the "reference" ratio R_R, as well as the variance of the quantity D_D

Ratio	r	$\mathrm{Var}(D_D)$
R_S	0.52	0.67
R_M	0.29	1.45
R_{L1}	-0.09	1.59
R_{L2}	0.30	0.83

We noted in Section 3.5 that our framework enabled the estimation of the relative density of a species monitored in the opportunistic scheme with unknown effort but not in the scheme with known effort. We verified this point using the following procedure: for each species i in turn, we built a dataset $M2$ by removing the counts of individuals of that species from the "known effort" dataset M. We then fitted our model combining the complete LPO dataset (opportunistic dataset, including records of the species i) and the dataset $M2$ ("known effort" dataset, excluding records of the species i). This model provided estimates of the relative densities $\widehat{A}_{ij}$ of the species i in all sites j. We were thus able to calculate the ratio $R_{S'}$ for the species i removed from M and all pairs of sites j, j'. This ratio was calculated for each species. The correlation between the ratio $R_{S'}$ and the ratio R_R was almost identical to the correlation between the ratio R_S and the reference R_R ($r = 0.51$). Similarly, the variance of $D_{S'}$ was almost identical to the variance of D_S ($\mathrm{Var}(D_{S'}) = 0.67$). These results illustrate that our approach can be used to estimate the spatial variation of species not monitored in the scheme characterized by an unknown effort.

4.4 Discussion of the results

The opportunistic dataset alone was a rather poor predictor of the reference dataset. In particular, this was the case when the total number of records per site was used as a proxy for site-specific effort. This suggests that there was a large variation in both the species-specific detectability (as is indeed suggested by the Table 1) and the species- and site-specific density. According to local bird-watchers, the estimates we obtained with our model are credible, indicating that the opportunistic dataset, even considering the plethora of data, should not be used for monitoring purposes without attempts to correct for variations in the detection process. In addition, we found the dataset collected with a known effort to be a poor predictor of the variation in density within the reference dataset. This may be due to the small number of sampled data points (3 to 5 2×2 km for each 100×100 km sites). Although these data points were drawn randomly, such a low number was unlikely to efficiently capture the density variation on the larger scale.

One of the most striking results from our model is that the combination of the two datasets was much better at predicting the variation of the reference dataset than expected from the predictive power of each dataset alone. This suggests that our model enables one to capitalise on the complementary strengths of each dataset. The very fine-grained distribution of observations contained in the opportunistic dataset can more efficiently predict site-specific variation in density than can the "known-effort" dataset. On the other hand, the opportunistic dataset only provides acceptable estimates when the variation in effort and species detectability is estimated with the help of the "known-effort" dataset. The last series of models confirmed this interpretation. Specifically, when one species was removed from the "known-effort" dataset, the model was equally good at predicting that species' relative density, suggesting that the information for estimating species density had entirely come from the opportunistic dataset.

5 Discussion

5.1 Discussion of the model

We have modelled the number of detections X_{ijk} as a result of a sampling process of intensity $A_{ij}E_{jk}P_{ik}$. This consisted of a generalised linear model fit involving only first-order interactions and assuming no second-order interaction between variable species i, sites j and schemes k. Let us observe the implications of this fundamental assumption. For this, we note that the above is a combination of a biological component A_{ij} and a component describing the observation process $O_{ijk} = E_{jk}P_{ik}$. We consider the two components in turn.

Recall that A_{ij} is the true density of species i in site j. The model only gives access to the relative density $\widetilde{A}_{ij}$, that is, A_{ij} multiplied by an unknown constant depending only on species i. The first important implication of our fundamental assumption is that the two schemes aim to measure the same quantity proportional to A_{ij}. This will be met if we can assume that the two schemes have the same spatial coverage (and temporal coverage when density also varies in time). Within a site j, a species may not be distributed homogeneously in the landscape depending on habitat preference. To estimate $\widetilde{A}_{ij}$ without bias, it is thus necessary that the dataset with known effort samples habitats proportional to their availability. We will see below that the constraint on habitat sampling is less stringent for the opportunistic data.

The second significant implication of our fundamental assumption is that for each scheme, the intensity of the observation process O_{ijk} can be broken down into two components: a site-dependent effort E_{jk} and a species-dependent detection probability P_{ik}. In practice, the observation process O_{ijk} is influenced for a given site j and scheme k by the following: the total observational effort; how such effort is distributed among habitats; the species detectability *per se* (detection probability given the joint presence of an individual member of the species and the observer); and the reporting rate of the detected species. What then are the critical assumptions for these influences to fit in the model components? Basically, none of these elements should vary *simultaneously* across the site, species and scheme. For the dataset collected with a known effort, we already noted that habitats need to be sampled proportionally to their availability. Yet, observers from the opportunistic scheme are unlikely to sample habitats in available proportion (e.g., farmland habitats are generally undersampled). In such cases, this biased attention will affect the P_{ik} (farmland birds detectability will appear systematically lower for scheme $k = 1$). However, such corrections will be efficient only if land use does not vary significantly across sites j or if the attention of the observers is not too biased toward certain habitats. Otherwise, the estimation of A_{ij} will be biased toward the density in the oversampled habitats in the opportunistic dataset. This is a clear limitation of our current model, and we plan to explore how within-site stratification (e.g., habitats) could be incorporated.

Regarding the species detectability and reporting rate, we know that observers vary considerably in their species-specific attention relative to the scheme characterized by a known effort in the opportunistic dataset, and this is shown by the estimation of the relative detection probability (Table 1). However, as long as these species detectabilities and reporting rates do not vary across sites, these discrepancies will be taken into account in the P_{ik}. This will arise if all observers behave similarly or if the sites are large enough to include many observers (ensuring an averaging effect in the observing behaviour).

The hypothesis of constant detectability across sites j for a given species i for a given scheme may not be met, as the heterogeneity of probability detection is documented in the literature (Link and Sauer, 1997; MacKenzie and Kendall, 2002). Many statistical frameworks based on a particular sampling design have been suggested to estimate detectability, such as using mixture models based on repeated counts (Royle, 2004). Further work is required to adapt such methods to our proposed framework. Meanwhile, our method should only be used if the hypothesis of constant species detectability across sites j for the scheme with known effort is reasonable.

In our estimation framework, we did not take into account any variable affecting the distribution of the relative densities in the different sites. However, it is well-known that there might be a spatial (if the "sites" are spatial units) or temporal (if the "sites" are time units) autocorrelation in the densities. For example, it is frequent that if the density of a given species is high in a given spatial unit, it will also be high in neighboring units. Moreover, spatial units with a similar environmental composition will often

be characterized by similar densities. Explicitly accounting for these patterns in the estimation process could lead to an increased precision of the estimation (by reducing the effective number of parameters). This could be done by modelling the relative densities $\widetilde{A}_{ij}$ as a function of environmental variables, or as a function of spatial effects (e.g. using conditional autoregression effects in a hierarchical model, see Banerjee et al., 2004). Alternatively, it is possible to maximize a regularized log-likelihood, i.e. to maximize for example:

$$\log \mathcal{L} - \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{m=1}^{J} \nu \pi_{jm} (a_{ij} - a_{im})^2$$

where $\mathcal{L}$ is the likelihood of the model, π_{jm} is a measure of "environmental and spatial proximities" between the unit j and the unit m, and ν is a positive parameter that determines the strength of the penalty. The proximities could be of any sort (e.g. taking the value 1 if the two spatial units are neighbours, and 0 otherwise; inverse Euclidean distances between the units in the space defined by the environmental variables, etc.). This kind of regularization would reduce the number of effective parameters in the model and thereby increase the precision of the estimation (for example, see Malbasa and Vucetic, 2011). We stress that the model described here is a very general approach that will need to be precised and adapted, depending on the specifics of each case study.

We end this discussion of the model by highlighting two promising applications of our framework. Although we illustrated our framework while focusing on the estimation of spatial variation in density, recall that it can be readily used to estimate temporal variation. In such cases, E_{j1} would represent the parameters describing the unknown effort at time j for the opportunistic dataset. We stress that for temporal variation, biased attention for some habitats in the opportunistic dataset will not bias the estimation of A_{ij} as long as this biased attention is constant over time. Such biases will be entirely captured in the estimation of the P_{ik}. For example, the precision of bird population trends for France will be considerably improved by the addition of opportunistic data to the current Breeding Bird Survey.

Another very interesting feature of our framework is its ability to estimate the relative abundance of very rare species, even if this species is not part of a scheme with known sampling effort. This has important practical implications. For example, Guisan et al. (2006) noted "in a sample of 550 plots surveyed in a random-stratified way based on the elevation, slope, and aspect of the plot during two consecutive summers in the Swiss Alps (704.2 km^2), not one occurrence of the rare and endangered plant species *Eryngium alpinum* L. was recorded. This was despite the species being easily detectable if present and independent records of the species existing in the area within similar vegetation types". Our framework would be very useful in this context. In particular, if a citizen science program allowed the collection of opportunistic data on this species along with other more common species, then the relative abundance of the rare species could be estimated by combining these data with a standardised scheme with these same common species.

5.2 Recommendations for opportunistic schemes

We stress again that the proposed method does not require the assumption of equal detectability among species or equal species-detectability among schemes. What matters is some spatial homogeneity in observer behaviour. This can be achieved through large numbers of observers or through appropriate animation of the opportunistic scheme. Given the internet's ability to foster strong social interaction among observers, there could be some calls for attention (or avoidance) for some species (e.g., "we need to pay attention to ...", "we now have enough data for..."). If such calls are spatially biased, this would impair the quality of the database for monitoring purposes. Coordinators of such databases should discourage such variation in attention. In fact, in the proposed statistical framework, the dataset has its maximal value if one can assume observers belong to the same statistical population with respect to species-specific attention with homogenous mean and variance. Paradoxically, any recommendations for more standardisation may harm more than help. A better message to deliver may be "Keep on recording what you want, where you want, when you want; the accumulation of data will be useful".

Acknowledgements

We warmly thank the members of the CiSStats group for stimulating and fruitful discussions on opportunistic data and related statistical issues. We also thank Laurent Couzy and Ondine Filippi-Codaccioni for facilitating access to the LPO-Aquitaine database. This work was partially funded by the Mastodon program from CNRS, by the CiSStats program from INRA and by the Chaire de Modélisation Mathématiques et Biodiversité from VEOLIA-Ecole Polytechnique-MNHN.

References

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data.* Boca Raton, Florida: Chapman and Hall/CRC.

Bishop, J., Venables, W. N., and Wang, Y.-G. (2004). Analysing commercial catch and effort data from a penaeid trawl fishery: a comparison of linear models, mixed models, and generalised estimating equations approaches. *Fisheries Research* **70,** 179–193.

Buckland, S., Anderson, D., Burnham, K., and Laake, J. (1993). *Distance Sampling: Estimating Abundance of Biological Populations.* Boca Raton, Florida: Chapman and Hall.

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T., and Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment* **10,** 291–297.

Dickinson, J. L., Zuckerberg, B., and Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics* **41,** 149–172.

Gibbons, D. W., Donald, P. F., Bauer, H.-G., Fornasari, L., and Dawson, I. K. (2007). Mapping avian distributions: the evolution of bird atlases. *Bird Study* **54,** 324–334.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N. G., Lehmann, A., and Zimmermann, N. E. (2006). Using niche-based models to improve the sampling of rare species. *Conservation Biology* **20,** 501–511.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., and Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution* **27,** 130–137.

Jiguet, F., Devictor, V., Julliard, R., and Couvet, D. (2012). French citizens monitoring ordinary birds provide tools for conservation and ecological sciences. *Acta Oecologica* **44,** 58–66.

Julliard, R., Jiguet, F., and Couvet, D. (2004). Common birds facing global changes: what makes a species at risk? *Global Change Biology* **10,** 148–154.

Kéry, M., Dorazio, R., Soldaat, L., van Strien, A., Zuiderwijk, A., and Royle, J. (2009). Trend estimation in populations with imperfect detection. *Journal of Applied Ecology* **46,** 1163–1172.

Link, W. and Sauer, J. (1997). Estimation of population trajectories from count data. *Biometrics* **53,** 488–497.

MacKenzie, D. I. and Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology* **83,** 2387–2393.

MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., and Bailey, L. L. (2005). Improving inferences in population studies of rare species that are detected imperfectly. *Ecology* **86,** 1101–1113.

Malbasa, V. and Vucetic, S. (2011). Spatially regularized logistic regression for disease mapping on large moving populations. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1352–1360. ACM.

Maunder, M. and Punt, A. (2004). Standardizing catch and effort data: a review of recent approaches. *Fisheries Research* **70,** 141–159.

Phillips, S., Dudík, M., Elith, J., Graham, C., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19,** 181–197.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60,** 108–115.

Seber, G. (1982). *The estimation of animal abundance and related parameters*. Charles Griffin & company ltd.

van Strien, A. and Pannekoek, J. (2001). Indexing european bird population trends using results of national monitoring schemes: a trial of a new method. *Bird Study* **48,** 200–213.

van Swaay, C. A., Nowicki, P., Settele, J., and van Strien, A. J. (2008). Butterfly monitoring in europe: methods, applications and perspectives. *Biodiversity and Conservation* **17,** 3455–3469.

Williams, B., Nichols, J., and Conroy, M. (2002). *Analysis and management of animal populations: modeling, estimation, and decision making*. San Diego, California: Academic Press.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., et al. (2007). How global is the global biodiversity information facility? *PLoS One* **2,** e1124.

Yoccoz, N., Nichols, J., and Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution* **16,** 446–453.

Appendix: Introduction

In this web appendix, we provide:

- several additional mathematical proofs used in the paper to develop our modeling framework;
- The R code used for the calculations illustrating the modeling framework.

We have written this appendix with the R package knitr (Xie, 2013), which allows to combine LaTeX and R code, ensuring the reproducibility of the calculations.

A Mathematical proofs

A.1 From count to binary data

We have developed a framework to model a sample of counts, but the proposed method can also be used with other distributions. For example, it can be used to model relative abundance based on detection / non-detection data (also called use/availability data, corresponding to design I data according to Thomas and Taylor, 1993). In this situation, one simply changes the family distribution from Poisson to Binomial. In this case, the outcome of the observation is a Bernoulli random variable with probability equal to $A_{ij}E_jP_i$. Our framework developed in Section 3 of the paper can then be applied in a similar way, using a logarithm as the link between the response and the linear predictor.

A.2 Identifiability conditions

With the notations $a_{ij} = \log(A_{ij})$, $e_{jk} = \log(E_{jk})$ and $p_{ik} = \log(P_{ik})$, the model (1) described in our paper can be recast as a classical generalised linear model

$$X_{ijk} \sim \text{Poisson}(\lambda_{ijk}), \quad \text{with } \log(\lambda_{ijk}) = a_{ij} + e_{jk} + p_{ik}.$$

The kernel of the design matrix associated with this linear regression has a dimension equal to $I + J + 1$. Therefore, we need $I + J + 1$ constraints to ensure the identifiability of the model.

A.3 Properties of the estimators

The negative log-likelihood of the parameters $(\widetilde{A}_{ij}, \widetilde{E}_{jk}, \widetilde{P}_{ik})$ is

$$\mathcal{L} = \sum_{i\in I}\sum_{j\in J}\sum_{k\in\{0,1\}} \Big(\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik} - X_{ijk}\log(\widetilde{A}_{ij}\widetilde{E}_{jk}\widetilde{P}_{ik}) + \log(X_{ijk}!)\Big)$$

where the parameters $\{\widetilde{E}_{j0},\ j \in J\}$ and $\{\widetilde{P}_{i0},\ i \in I\}$ are known, $\widetilde{P}_{10} = 1$ and $\widetilde{P}_{i1} = 1$ for all $i \in I$.

To keep the mathematical analysis of the maximum likelihood estimators comprehensible, we focus below on the case where the $\widetilde{P}_{i0}$ are known. The maximum likelihood estimators of $\widetilde{A}_{ij}$ and $\widetilde{E}_{j1}$ are then the solutions of

$$\widehat{A}_{ij} = \frac{X_{ij0} + X_{ij1}}{\widetilde{P}_{i0}\widetilde{E}_{j0} + \widehat{E}_{j1}} \quad \text{and} \quad \widehat{E}_{j1} = \frac{X_{\#j1}}{\widehat{A}_{\#j}}, \tag{5}$$

where $X_{\#jk} = \sum_i X_{ijk}$ and $\widehat{A}_{\#j} = \sum_i \widehat{A}_{ij}$.

We first treat the simplest case where the $\widetilde{P}_{i0}$ are all equal.

A.3.1 Case of constant ratios P_{i0}/P_{i1}

We consider in this paragraph the case where $\widetilde{P}_{i0} = \widetilde{P}_{10}$ for all $i \in I$. This corresponds to the case where for all the species i, the probability detection ratios P_{i0}/P_{i1} are the same and equal to P_{10}/P_{11}. We derive from (5)

$$\widehat{A}_{\#j} = \frac{X_{\#j0} + X_{\#j1}}{\widetilde{E}_{j0} + \widehat{E}_{j1}}$$

and inserting this expression in the formula for $\widehat{E}_{j1}$ we obtain $\widehat{E}_{j1} = \widetilde{E}_{j0}X_{\#j1}/X_{\#j0}$. As a consequence, we obtain the closed-form expression for $\widehat{A}_{ij}$

$$\widehat{A}_{ij} = \frac{X_{ij0} + X_{ij1}}{X_{\#j0} + X_{\#j1}} \times \frac{X_{\#j0}}{\widetilde{E}_{j0}}. \tag{6}$$

According to the strong law of large numbers for Poisson processes, we have

$$\widehat{A}_{ij} \overset{E_{j1}\gg 1}{\to} \frac{\widetilde{A}_{ij}}{\widetilde{A}_{\#j}} \times \frac{X_{\#j0}}{\widetilde{E}_{j0}} \tag{7}$$

and

$$\text{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\to} \left(\frac{\widetilde{A}_{ij}}{\widetilde{A}_{\#j}}\right)^2 \times \frac{\widetilde{A}_{\#j}}{\widetilde{E}_{j0}} = \frac{\widetilde{A}_{ij}}{\widetilde{E}_{j0}} \times \frac{A_{ij}P_{i0}}{\sum_l A_{lj}P_{l0}}.$$

If we estimate $\widetilde{A}_{ij}$ with the sole "known-effort" data X_{ij0}, the maximum likelihood estimator is given by $\widehat{A}^0_{ij} = X_{ij0}/\widetilde{E}_{j0}$ and its variance equals $\text{var}(\widehat{A}^0_{ij}) = \widetilde{A}_{ij}/\widetilde{E}_{j0}$. We can then compare the variance of $\widehat{A}_{ij}$ and $\widehat{A}^0_{ij}$

$$\text{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\sim} \text{var}(\widehat{A}^0_{ij}) \times \frac{A_{ij}P_{i0}}{\sum_l A_{lj}P_{l0}}. \tag{8}$$

A.3.2 Case of arbitrary ratios P_{i0}/P_{i1}

We no longer assume that the $\widetilde{P}_{i0}$ are all equal. In this case, we have no closed-form formula for $\widehat{A}_{ij}$ but we can compute a first-order expansion of $\widehat{A}_{ij}$ in terms of the inverse of $X_{\#j1}$.

The first step is to check that $\widehat{A}_{\#j}$ is upper-bounded independently of the X_{ij1}. When $P_{i0} > 0$ for all i (which means that the same species are monitored in the datasets 0 and 1), we have from (5)

$$\widehat{A}_{ij} \leq \frac{X_{ij0} + X_{ij1}}{\min_i(\widetilde{P}_{i0}\widetilde{E}_{i0}) + X_{\#j1}/\widehat{A}_{\#j}}\,.$$

Summing these inequalities we obtain the upper-bound

$$\widehat{A}_{\#j} \leq X_{\#j0}/\min_i(\widetilde{P}_{i0}\widetilde{E}_{j0})$$

which does not depend on X_{ij1}. The case where $P_{i0} = 0$ for some i can be treated similarly: splitting apart the indices in $I_0 = \{i \in I : P_{i0} = 0\}$ and those out of I_0, we get from (5)

$$\widehat{A}_{\#j} \leq \frac{\sum_{i\in I_0}(X_{ij0} + X_{ij1})}{X_{\#j1}/\widehat{A}_{\#j}} + \frac{\sum_{i\notin I_0}(X_{ij0} + X_{ij1})}{\min_{i\notin I_0}(\widetilde{P}_{i0}\widetilde{E}_{i0}) + X_{\#j1}/\widehat{A}_{\#j}}.$$

This inequality is equivalent to

$$\widehat{A}_{\#j}\left(1 - \frac{\sum_{i\in I_0}(X_{ij0} + X_{ij1})}{X_{\#j1}}\right) \leq X_{\#j0}/\min_{i\notin I_0}(\widetilde{P}_{i0}\widetilde{E}_{j0}).$$

In the asymptotic $E_{j1} \gg 1$ we obtain the asymptotic upper-bound

$$\widehat{A}_{\#j} \leq \frac{X_{\#j0}}{\min_{i\notin I_0}(\widetilde{P}_{i0}\widetilde{E}_{j0})} \times \frac{\sum_{i\in I}\widetilde{A}_{ij}}{\sum_{i\in I\setminus I_0}\widetilde{A}_{ij}}.$$

Now that we have checked that $\widehat{A}_{\#j}$ is (asymptotically) upper-bounded independently of the X_{ij1}, we can write a first-order expansion of the formula (5)

$$\widehat{A}_{ij} = \frac{(X_{ij0} + X_{ij1})\widehat{A}_{\#j}}{X_{\#j1}} - \frac{(X_{ij0} + X_{ij1})\widehat{A}^2_{\#j}\widetilde{P}_{i0}\widetilde{E}_{j0}}{X^2_{\#j1}} + O\left(\frac{X_{ij1}}{X^3_{\#j1}}\right). \tag{9}$$

Summing these expansions over $i \in I$ and simplifying the expression gives

$$\widehat{A}_{\#j} = \frac{X_{\#j0}X_{\#j1}}{\widetilde{E}_{j0}\sum_l \widetilde{P}_{l0}(X_{lj0} + X_{lj1})}\left(1 + O\left(\frac{1}{X_{\#j1}}\right)\right).$$

Plugging this formula in (9) gives

$$\begin{aligned}\widehat{A}_{ij} \quad &= \quad \frac{X_{ij0} + X_{ij1}}{\sum_l \widetilde{P}_{l0}(X_{lj0} + X_{lj1})} \times \frac{X_{\#j0}}{\widetilde{E}_{j0}} \times \left(1 + O\left(\frac{1}{X_{\#j1}}\right)\right) \\ &\overset{E_{j1}\gg 1}{\to} \quad \frac{\widetilde{A}_{ij}}{\sum_l \widetilde{P}_{l0}\widetilde{A}_{lj}} \times \frac{X_{\#j0}}{\widetilde{E}_{j0}},\end{aligned}$$

where the last limit follows again from the law of large numbers for Poisson processes. Computing the asymptotic variance when $E_{j1} \gg 1$, we find after simplification

$$\mathrm{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\to} \frac{\widetilde{A}^2_{ij}}{\sum_l \widetilde{P}_{l0}\widetilde{A}_{lj}\widetilde{E}_{j0}} = \frac{\widetilde{A}_{ij}}{\widetilde{P}_{i0}\widetilde{E}_{j0}} \times \frac{P_{i0}A_{ij}}{\sum_l P_{l0}A_{lj}}. \tag{10}$$

As in the previous case, we can compare this variance to the variance of the maximum likelihood

estimator $\widehat{A}^0_{ij} = X_{ij0}/(\widetilde{P}_{i0}\widetilde{E}_{j0})$ obtained by estimating $\widetilde{A}_{ij}$ with the sole values X_{ij0}. The variance of $\widehat{A}^0_{ij}$ being $\text{var}(\widehat{A}^0_{ij}) = \widetilde{A}_{ij}/(\widetilde{P}_{i0}\widetilde{E}_{j0})$, we obtain the reduction of variance

$$\text{var}(\widehat{A}_{ij}) \overset{E_{j1}\gg 1}{\sim} \text{var}(\widehat{A}^0_{ij}) \times \frac{P_{i0}A_{ij}}{\sum_l P_{l0}A_{lj}}. \tag{11}$$

B Calculations with the R software

In this section, we illustrate how to fit the model presented in the paper with the R software.

The data are stored in the file "DataSTOCLPO.txt" available in the supplementary material of the paper. We load these data in R, and display the first rows of the data:

```
detections <- read.table("DataSTOCLPO.txt",
    sep = "\t", header = TRUE, stringsAsFactors = TRUE)
head(detections)

##   Species Site Effort_M AreaSite
## 1  CARCHL   S1        4     5787
## 2  CARCHL   S2        4     5534
## 3  CARCHL   S3        6     3992
## 4  CARCHL   S4        3     5949
## 5  CARCHL   S5       10     7419
## 6  CARCHL   S6        3     5562
##   Total_X_LPO X_M  X_L X_R
## 1       32633  22  799   6
## 2       16393  10  520   5
## 3       30113  16 1146   2
## 4       12497  15  413  25
## 5       44965  38 1839   7
## 6       17068  11  674  22
```

The data are stored in a data frame with 5 columns:

- The species (column `Species`) is coded as a factor with 22 levels. The latin name corresponding to each species code is available in the table 2 of the main text;
- The site (column `Site`) is coded as factor with 6 levels (S1 to S6);
- The numbers of detections in the datasets M, L and R are stored respectively in the columns `X_M`, `X_L` and `X_R`.
- The log-effort for the dataset M is stored in the column `Effort_M`: it corresponds to the number of FBBS squares in each site;
- The area of each site is stored in the column `AreaSite`
- The total number of detections (all species pooled) for the dataset L in each site is stored in the column `Total_X_LPO`.

Each row of this dataframe is a combination of species and site. Therefore, this dataframe contains $I \times J = 132$ rows.

C Implementation of the model

The R function below can be used to fit our model:

```
fitmodel <- function(species_site, X_standardized,
    X_opportunistic, Effort_standardized) {

    ## First check the data: Check that
    ## species_site is a data.frame
    if (!is.data.frame(species_site))
        stop("species_site should be a data.frame")

    ## Check that the columns are factors
    if (!all(sapply(species_site, is.factor)))
        stop("The columns of species_site should be of class factor")

    ## Checks for missing values
    if (any(is.na(unlist(species_site))))
        stop("No missing values are allowed in species_site")
    if (any(is.na(X_standardized)))
        stop("No missing values are allowed in X_standardized")
    if (any(is.na(X_opportunistic)))
        stop("No missing values are allowed in X_opportunistic")
    if (any(is.na(Effort_standardized)))
        stop("No missing values are allowed in Effort_standardized")

    ## Check that All species are represented
    ## for each site
    if (!all(sapply(split(species_site[,
        1], species_site[, 2]), length) ==
        length(unique(species_site[, 1]))))
        stop("At least one site does not contain information for some species")

    ## Check the length of the other elements
    if (length(X_standardized) != nrow(species_site))
        stop("species_site and X_standardized do not match")
    if (length(X_opportunistic) != nrow(species_site))
        stop("species_site and X_opportunistic do not match")
    if (length(Effort_standardized) != nrow(species_site))
        stop("species_site and Effort_standardized do not match")

    ## Prepare the design matrix

    ijk <- cbind(rbind(species_site, species_site),
        c(rep("Standardized", nrow(species_site)),
            rep("Opportunistic", nrow(species_site))))
    X <- c(X_standardized, X_opportunistic)
    Species.sites_ <- factor(paste(ijk[,
        1], ijk[, 2], sep = "."))
    Sites.data_ <- factor(paste(ijk[, 2],
        ijk[, 3], sep = "."))
    Species.data_ <- factor(paste(ijk[, 1],
        ijk[, 3], sep = "."))

    matrix1 <- model.matrix(glm(X ~ Species.sites_ -
        1))
    matrix2 <- model.matrix(glm(X ~ Sites.data_ -
```

```
        1))
    matrix2 <- matrix2[, -grep("Standardized",
        colnames(matrix2))]
    matrix3 <- model.matrix(glm(X ~ Species.data_ -
        1))
    matrix3 <- matrix3[, -grep("Opportunistic",
        colnames(matrix3))]
    matrix3 <- matrix3[, -1]
    model_matrix <- as.data.frame(cbind(matrix1,
        matrix2, matrix3))

    ## Prepare the log-effort

    logEffort <- log(c(Effort_standardized,
        rep(1, length(Effort_standardized))))

    ## Fit the model
    mod <- glm(X ~ . - 1, data = model_matrix,
        offset = logEffort, family = quasipoisson())

    ## Extract the coefficients of interest
    coe <- summary(mod)$coefficients[, 1:2]

    ## Log-Relative abundance
    LogAij <- cbind(species_site, do.call("rbind",
        lapply(1:nrow(species_site), function(i) {
            coe[row.names(coe) == paste("Species.sites_",
                species_site[i, 1], ".",
                species_site[i, 2], sep = ""),
                ]
        })))
    ## Log-Effort for the opportunistic
    ## dataset
    so <- sort(unique(species_site[, 2]))
    LogEjk <- data.frame(Site = so, do.call("rbind",
        lapply(1:length(so), function(i) {
            coe[row.names(coe) == paste("Sites.data_",
                so[i], ".", "Opportunistic",
                sep = ""), ]
        })))
    ## Log-Relative detection probability
    so <- sort(unique(species_site[, 1]))[-1]
    LogPik <- data.frame(Site = so, do.call("rbind",
        lapply(1:length(so), function(i) {
            coe[row.names(coe) == paste("Species.data_",
                so[i], ".", "Standardized",
                sep = ""), ]
        })))

    ## the results
    results <- list(LogAij = LogAij, LogEjk = LogEjk,
        LogPik = LogPik, model = mod)

    ## Return
    return(results)
}
```

This function takes the following arguments:

- `species_site`: a data.frame with two columns and $I \times J$ rows containing all the possible combinations of Species (first column) and sites (second column);
- `X_standardized`: a vector with $I \times J$ values, each value corresponding to the number of detections in the standardized dataset, for the corresponding combination of species and site in the data frame `species_site`
- `X_opportunistic`: a vector with $I \times J$ values, each value corresponding to the number of detections in the opportunistic dataset, for the corresponding combination of species and site in the data frame `species_site`
- `Effort_standardized`: a vector with $I \times J$ values, each value corresponding to the value of the search effort in the standardized dataset, for the corresponding site $\times$ species combination in the data frame `species_site`.

We use this function in this appendix to carry out our computation, but the reader can use it on its own data.

D Model fit

We now carry out the calculations described in the paper. First, we fit the model, using the dataset M as the standardized dataset and the dataset L as the opportunistic dataset:

```
mod <- fitmodel(detections[, 1:2], X_standardized = detections$X_M,
    X_opportunistic = detections$X_L, Effort_standardized = detections$Effort_M)
## Displays a short summary of the
## resulting object
str(mod, 1)

## List of 4
##  $ LogAij:'data.frame': 132 obs. of  4 variables:
##  $ LogEjk:'data.frame': 6 obs. of  3 variables:
##  $ LogPik:'data.frame': 21 obs. of  3 variables:
##  $ model :List of 30
##   ..- attr(*, "class")= chr [1:2] "glm" "lm"
```

This function returns a list with the following elements:

- `LogAij`: a data.frame containing the logarithm of the relative abundance for each species in each site, as well as the standard error of this estimate
- `LogEjk`: a data.frame containing the estimate of the logarithm of the search effort estimated by the model for the opportunistic data.set, as well as the standard error of this estimate;
- `LogPik`: a data.frame containing the estimate of the logarithm of the relative detection probability for all species in the standardized dataset, as well as the standard error of this estimate.
- `model`: the object returned internally by the function `glm()`, containing all the details of the fit.

E Predictive capacities of the model

Based on the fit carried out in the previous section, we calculate the various ratios described in the main text:

```
R_R <- lapply(unique(detections$Species),
    function(i) outer(detections$X_R[detections$Species ==
        i], detections$X_R[detections$Species ==
        i], FUN = "/"))

R_M <- lapply(unique(detections$Species),
    function(i) outer(detections$X_M[detections$Species ==
        i]/detections$Effort_M[detections$Species ==
        i], detections$X_M[detections$Species ==
        i]/detections$Effort_M[detections$Species ==
        i], FUN = "/"))

R_L1 <- lapply(unique(detections$Species),
    function(i) outer(detections$X_L[detections$Species ==
        i]/detections$AreaSite[detections$Species ==
        i], detections$X_L[detections$Species ==
        i]/detections$AreaSite[detections$Species ==
        i], FUN = "/"))

R_L2 <- lapply(unique(detections$Species),
    function(i) outer(detections$X_L[detections$Species ==
        i]/detections$Total_X_LPO[detections$Species ==
        i], detections$X_L[detections$Species ==
        i]/detections$Total_X_LPO[detections$Species ==
        i], FUN = "/"))

R_S <- lapply(unique(detections$Species),
    function(i) outer(mod$LogAij$Estimate[mod$LogAij$Species ==
        i], mod$LogAij$Estimate[mod$LogAij$Species ==
        i], FUN = "/"))
```

All these objects are lists with one element per species, each element i corresponding to a square $J \times J$ matrix (containing at the intersection of the row j and of the column j' the ratio $R_D(i, j, j')$, where D is one of the datasets). We calculate the correlation coefficient between each log-ratio R_D and the log-ratio calculated on the independent dataset R_R:

```
r_M <- cor(log(unlist(lapply(1:length(R_R),
    function(i) R_M[[i]][lower.tri(R_R[[i]])]))),
    log(unlist(lapply(1:length(R_R), function(i) R_R[[i]][lower.tri(R_R[[i]])]))))

r_L1 <- cor(log(unlist(lapply(1:length(R_R),
    function(i) R_L1[[i]][lower.tri(R_R[[i]])]))),
    log(unlist(lapply(1:length(R_R), function(i) R_R[[i]][lower.tri(R_R[[i]])]))))

r_L2 <- cor(log(unlist(lapply(1:length(R_R),
    function(i) R_L2[[i]][lower.tri(R_R[[i]])]))),
    log(unlist(lapply(1:length(R_R), function(i) R_R[[i]][lower.tri(R_R[[i]])]))))

r_S <- cor(log(unlist(lapply(1:length(R_R),
    function(i) R_S[[i]][lower.tri(R_R[[i]])]))),
    log(unlist(lapply(1:length(R_R), function(i) R_R[[i]][lower.tri(R_R[[i]])]))))
```

```
data.frame(ratio = c("R_S", "R_M", "R_L1",
    "R_L2"), value = c(r_S, r_M, r_L1, r_L2))

##   ratio    value
## 1   R_S  0.50978
## 2   R_M  0.28784
## 3  R_L1 -0.09043
## 4  R_L2  0.29676
```

These values correspond to the values displayed in table 1 of the main text.

We now calculate the variance of the log-differences D_D:

```
varD_M <- var(unlist(lapply(1:length(R_R),
    function(i) {
        aa <- log(R_M[[i]]) - log(R_R[[i]])
        ## keep only the lower triangular part of
        ## this matrix
        aa[lower.tri(aa)]
    })))
varD_L1 <- var(unlist(lapply(1:length(R_R),
    function(i) {
        aa <- log(R_L1[[i]]) - log(R_R[[i]])
        ## keep only the lower triangular part of
        ## this matrix
        aa[lower.tri(aa)]
    })))
varD_L2 <- var(unlist(lapply(1:length(R_R),
    function(i) {
        aa <- log(R_L2[[i]]) - log(R_R[[i]])
        ## keep only the lower triangular part of
        ## this matrix
        aa[lower.tri(aa)]
    })))
varD_S <- var(unlist(lapply(1:length(R_R),
    function(i) {
        aa <- log(R_S[[i]]) - log(R_R[[i]])
        ## keep only the lower triangular part of
        ## this matrix
        aa[lower.tri(aa)]
    })))

data.frame(ratio = c("R_S", "R_M", "R_L1",
    "R_L2"), value = c(varD_S, varD_M, varD_L1,
    varD_L2))

##   ratio  value
## 1   R_S 0.6764
## 2   R_M 1.4538
## 3  R_L1 1.5928
## 4  R_L2 0.8307
```

This table present the values displayed in table 1 of the main text.

References

Xie, Y. (2013). knitr: A general-purpose package for dynamic report generation in R. *R package version* **1**.

Thomas, D.L. and Taylor, E.J. (1990). Study designs and tests for comparing resource use and availability. *Journal of Wildlife Management*, **54**, 322–330.