Leveraging citizen science to assess richness, diversity, and abundance

Tim M. Szewczyk[1] and Cleo Bertelsmeier[1]

[1]University of Lausanne

1 Introduction

What structures ant species richness, diversity, and community structure at different spatial scales? We know that at a coarse scale, climate is generally important, with some support for phylogenetically conserved temperature preferences. At a local scale, richness typically decreases with canopy cover. In general, abundance seems to be more idiosyncratic and variable, both temporally and spatially. This is particularly true for small species. In ants, abundance can be measured as either the number of colonies in a particular area (i.e., colony density) or as the number of workers (i.e., worker density).

Increasingly, ecologists have access to occurrence data collected in various and haphazard ways, typically in the form of online databases or citizen science projects. These data are commonly used for species distribution models, but their use in predicting richness or diversity directly has been somewhat more limited. There are several reasons for this. First, the data do not generally come from communities or assemblages, but rather an aggregation of detections from many different collectors across a variety of time spans. We like to think of diversity and richness as properties of communities, and these are decidedly not samples of communities, obscuring the ability to detect or account for interactions among species. Second, the collections for each species may have differing spatial biases, rendering any simple aggregation methods erroneous. Third, there are biases in the species that are more likely to be detected, such that any estimate of richness or diversity will necessarily be of a subset of the community biased toward larger, more active, or more interesting species.

However, that doesn't mean these data can't be useful. Instead, the geographic breadth and rather indiscriminate collection methods can capture occurrences in unexpected locations or species that may be missed in alternative, more structured sampling methods. Here, we combine species occurrences of ants collected in a citizen science project in western Switzerland with a concurrent structured sampling effort. In a hierarchical Bayesian framework, we use the citizen science data to help inform species' responses to regional variables, and the structured samples to inform responses to both regional and local variables. We then estimate species diversity in each community, accounting for species that may not have been detected, and predicting the posterior alpha, beta, and gamma diversities, capturing the uncertainty in the community composition. Based on the species composition in each sample from the posterior distribution, we calculate taxonomic and phylogenetic diversity across spatial scales. For comparison, we run the model with each dataset individually to assess the impact of including the citizen science data.

Model goals:

1. Predict site-level diversity, including species undetected in structured sampling
2. Leverage benefits of each sampling type to better predict diversity across elevations

2 Methods

2.1 Study area

This is a blurb about Vaud, including some about the climate, topography, as well as current and past human land use. Also a little bit about what is known about the ant fauna here, perhaps.

2.2 Datasets

Ants were collected during the summer of 2019 in the canton Vaud, Switzerland in two simultaneous but distinct collection efforts. First, a citizen science project distributed labelled vials of ethanol to interested residents of the canton, who collected approximately 10 ant workers per colony for each vial. Participants were encouraged to explore under rocks, on bark, inside twigs, and in downed wood, and an online map was updated throughout the summer to highlight data-sparse areas. Collectors returned the vials along with the locality of the sample (latitude and longitude, an address, or an annotated map), and often basic habitat information. Nearly 7,000 samples were returned between April and November. We discretized the landscape into a $1km^2$ grid to generate the number of counts for each species in each cell.

Second, structured samples of ants were collected within 44 $1km^2$ sites. Thirty-nine of the sites are a part of long-term biodiversity monitoring efforts by the federal government, and as such are arranged on a regular grid with 5-7 km between adjacent sites. Five additional sites are established monitoring sites by the University of Lausanne. The ants at each site were characterized by 25 plots, distributed among habitat types in approximate proportion to the abundance of each habitat, where each habitat type present within the site was represented by at least one plot. Inaccessible areas (e.g., cliffs, water, property beyond Vaud) were excluded, resulting in several sites with areas less than $1km^2$, and the number of plots was reduced proportionally (Appendix). Plots consisted of a 2m radius circle, with soil temperature recorded in the center, and vegetation characterized according to coverage classes within the plot (grass, forb, shrub, litter, bare, moss). Six flags were evenly spaced around the circumference, and within 20cm of each (total surface area $0.75m^2$), we searched for ant colonies within any downed wood, under large rocks, and in 2L of soil, litter, and small rocks using Hori Hori gardening knives. We haphazardly collected 10 workers from each colony. All trees within the plot were also inspected for ant workers which were collected regardless of whether or not a colony was identifiable. Additionally, transects were mapped *a priori*, distributed proportionally among habitat types and totalling 2km. Workers were collected from all permanent mounds within 2m of the transect line.

There are many differences between W and Y, but they can be reduced down to just three key points:

1. W is collected very broadly spatially, with no constraints within Vaud, while Y was collected in discrete, representative plots within pre-determined squares

2. Effort varies across space for W, but is standardized for Y

3. Species collections are biased in W toward larger, more active species, while Y provides a less biased survey of ground-nesting ants

The model described below accounts for each of these three differences, leveraging the information contained in W to improve estimates for the local diversity estimated by Y.

2.3 Model overview

To leverage the information contained in each dataset, we use a community-level hierarchical inhomogenous Poisson point process model (PPM). Inhomogenous Poisson PPMs assume that the distribution of occurrences is dependent on the variation in local intensity, $\lambda(s)$, across space, s, which may be observed imperfectly resulting in a thinned point process. One key benefit of PPMs is that the predicted point location intensity can be integrated to arbitrary spatial resolutions. Following the structure of our sampling design, we modelled the expected abundance of each species at two resolutions (coarse: $1km^2$, fine: $0.75m^2$), representing the area of the sampling sites and the area of the sampling plots respectively.

2.4 Model structure

We use two datasets: haphazard citizen science collections ($\mathbf{W}$, res: $1km^2$), and structured samples ($\mathbf{Y}$, $0.75m^2$ plots within $1km^2$ sites). Note that for the citizen science dataset, we use only the coarse resolution ($1km^2$), while the structured samples are incorporated at both resolutions. We assume that the colony density for each species is a function of local and regional environmental conditions, with potential phylogenetic conservatism among species-specific responses. We estimate the effects of regional conditions using both $\mathbf{Y}$ and $\mathbf{W}$, and the effects of local conditions using $\mathbf{Y}$.

At a local scale, Y_{is} represents the number of colonies of species s detected at plot i of site j, which is an estimate of the true number of colonies λ_{is}, which in turn depends on the local environment and the overall abundance of species s at site j:

$$Y_{is} \sim Poisson(\lambda_{is}) \tag{1}$$

$$log(\lambda_{is}) = a_s V_i + log(h\Lambda_{js}) \tag{2}$$

where V_i is a vector of local environmental covariates, a_s is a vector of species-specific responses, h is a constant representing the proportion of site j sampled by plot i ($0.75m^2/1km^2 = 7.5e-7$), and Λ_{js} is the total expected abundance of species s at site j. Thus, $log(h\Lambda_{js})$ functions as an intercept, determining the baseline expected abundance at each plot within each site, which are then further refined according to the effects of the local environment.

At the site scale, the total abundance within a $1km^2$ cell is observed indirectly by Y_{is} as describe above, and directly by W_{ks}. The species counts from the unstructured survey in cell k are thinned such that:

$$W_{ks} \sim Poisson(\Lambda_{ks} E_k D_s) \tag{3}$$

where E_k is the sampling effort in that cell, and D_s is the detection bias for each species, which accounts for bias in the community composition based on species that are more readily observed. The effort E_k represents the expected proportion of colonies sampled, and is modeled as $logit(E_k) = \eta U_k$, where η is a vector of slopes and $\mathbf{U}$ is a set of covariates including the total number of vials collected, the population density, the length of roads, and the length of hiking trails. Thus, for each species s in cell k, the observed count W_{ks} is expected to be higher if the species is abundant in cell k (Λ_{ks} is larger), more samples were collected in k (E_k is larger), or if species s is likely to be overrepresented in haphazard collections by non-experts (D_s is larger).

Thus, both Λ_{js} and Λ_{ks} represent the expected abundance of species s per km^2, but the exact cells are not required to overlap. The ecological processes driving $\Lambda_{(jk)s}$ are assumed to be the same, however, such that:

$$log(\Lambda_{(j,k)s}) = b_s X_{(j,k)} \tag{4}$$

where b_s is a vector of species-specific slopes, and $\mathbf{X}$ is a matrix of environmental covariates.

The slopes a_s and b_s are species-specific responses at $0.75m^2$ and $1km^2$ resolutions, respectively, and are distributed about genus-level means A_g and B_g with standard deviations σ^a and σ^b. The genus-level means are in turn distributed about aggregate means, α and β with covariance matrices Σ^A and Σ^B, which reflect the overall responses of the ant community to environmental variables at each resolution while accounting for phylogenetic relatedness at the genus level.

Finally, we calculated several quantities at the site-scale and plot-scale. Following Hefley & Hooten (2016), we calculated the probability of presence as $\psi_i s = 1 - e^{-\lambda_{is}}$ and $\Psi_j s = 1 - e^{-\Lambda_{js}}$, as well as Fisher's α at each plot, β diversity among plots within each site, and among sites, and overall γ diversity at each site.

Parameter	**Description**	**Type**
i	number of $0.75m^2$ structured sampling plots	index
j	number of $1km^2$ structured sampling cells	index
k	number of $1km^2$ citizen science cells	index
s	number of species	index
g	number of genera	index
l	number of local covariates ($0.75m^2$)	index
r	number of regional covariates ($1km^2$)	index
q	number of effort covariates ($1km^2$)	index
$\mathbf{Y}_{is}$	structured sampling counts ($0.75m^2$)	data
$\mathbf{W}_{ks}$	citizen science counts ($1km^2$)	data
$\mathbf{V}_{il}$	local covariates ($0.75m^2$)	data
$\mathbf{X}_{(jk)r}$	regional covariates ($1km^2$)	data
$\mathbf{U}_{kq}$	citizen science effort covariates ($1km^2$)	data
h	structured sampling proportional effort	data
λ_{is}	true density ($0.75m^2$)	latent
$\mathbf{\Lambda}_{(jk)s}$	true density ($1km^2$)	latent
E_K	citizen science effort proportion ($1km^2$)	latent
α_l	aggregate ant responses ($0.75m^2$)	slopes
$\mathbf{A}_{lg}$	genus-level ant responses ($0.75m^2$)	slopes
$\mathbf{a}_{ls}$	species-level ant responses ($0.75m^2$)	slopes
σ^a_l	response sd among congeners	sd
$\mathbf{\Sigma^A}_{gg}$	genus-level covariance matrix	cov mx
β_r	aggregate ant responses ($1km^2$)	slopes
$\mathbf{B}_{rg}$	genus-level ant responses ($1km^2$)	slopes
$\mathbf{b}_{rs}$	species-level ant responses ($1km^2$)	slopes
σ^b_r	response sd among congeners	sd
$\mathbf{\Sigma^B}_{gg}$	genus-level covariance matrix	cov mx
η_q	citizen science effort	slopes
D_s	citizen science species bias (proportional)	random effect

2.5 Validation and model selection

To evaluate the effect of including $\mathbf{W}$ and to perform variable selection, we compared a version of the model with and without $\mathbf{W}$. We randomly divided the 44 sites with data for $\mathbf{Y}$ into training (75%: $J = 33$) and testing (25%: $J = 1$) subsets (**Appendix**). For each combination of covariates, we parameterized the models with the training subset and evaluated the ability to predict λ with the testing subset using the log predictive density, a metric that uses the full posterior predictive distribution.

3 Results

We used simulated data to validate the model, testing versions of the model using only **W**, only **Y**, and both **W** and **Y**. The model using both datasets performs best at predicting both site-level and plot-level abundances for testing subsets. I don't have any statistics yet, and will need to do full runs to be more confident. It is, of course, unsurprising that the site-level abundance Λ is predicted more accurately, as the use of both datasets effectively increases the sample size for estimating the site-level slopes. The more accurate prediction of λ is more exciting, since **W** doesn't directly include any information about it, but rather helps indirectly by improving predictions of Λ at each site, which helps to constrain the intercept in the plot-level equation.

4 Discussion

5 Appendixes

Appendix 1. Supplementary methods.

Appendix 2. Model code.

Full equations

$$Y_{is} \sim Poisson(\lambda_{is}) \tag{5}$$

$$log(\lambda_{is}) = a_s V_i + log(h\Lambda_{js}) \tag{6}$$

$$W_{ks} \sim Poisson(\Lambda_{ks} E_k D_s) \tag{7}$$

$$log(\Lambda_{(j,k)s}) = b_s X_{(j,k)} \tag{8}$$

$$logit(E_k) = \eta U_k \tag{9}$$

$$\begin{aligned} a_s &\sim Norm(A_g, \sigma^a) \\ A_g &\sim Norm_G(\alpha, \Sigma^A) \end{aligned} \tag{10}$$

$$\begin{aligned} b_s &\sim Norm(B_g, \sigma^b) \\ B_g &\sim Norm_G(\beta, \Sigma^B) \end{aligned} \tag{11}$$