Finding Inter-species Associations on Large Citizen Science Datasets

Jacob Deutsch
Terrabyte Lab, Santa Cruz, California, USA
jacobadeutschwork@gmail.com

August 2025

Abstract

Determining associations among different species from citizen science databases is challenging due to observer behavior and intrinsic density variations that give rise to correlations that do not imply species associations. This paper introduces a method that can efficiently analyze large datasets to extract likely species associations. It tiles space into small blocks chosen to be of the accuracy of the data coordinates, and reduces observations to presence/absence per tile, in order to compute pairwise overlaps. It compares these overlaps with a spatial Poisson process that serves as a null model. For each species i, an expected overlap μ_i is estimated by averaging normalized overlaps over other species in the same vicinity. This gives a z-score for significance of a species-species association and a correlation index for the strength of this association. This was tested on $874,263$ iNaturalist observations spanning $15,975$ non-avian taxa in the Santa Cruz, California region ($\approx 4.68 \times 10^6$ tiles). The method recovers well-known insect host-plant obligate relationships, particularly many host-gall relationships, as well as the relationship between Yerba Santa Beetles and California Yerba Santa. This approach efficiently finds associations on $\sim 10^8$ species pairs on modest hardware, filtering correlations arising from heterogeneous spatial prevalence and user artifacts. It produces a ranked shortlist of ecological interactions that can be further pursued. Extensions to this method are possible, such as investigating the effects of time and elevation. It could also be useful in the determination of microhabitats and biomes.

Keywords: citizen science, inter-species interactions, iNaturalist, Poisson

1 Introduction

Mapping inter-species interactions is fundamental to understanding how ecosystems function [1]. Interactions have usually been studied one species pair at a time. The aim of this paper is to demonstrate a new method for uncovering inter-species interactions with large data sets. For example, iNaturalist's

dataset consists of more than 200 million observations [2]. Although some of these data points are annotated with respect to species interactions, the majority of iNaturalist data only contain three columns of relevant information. A taxon name, coordinate, and time. The overwhelmingly larger amount of this simpler information and what can be attained from it is the main subject of this research.

Many ecological-niche modeling (ENM) techniques have been developed that use co-occurrence and environmental factors to make predictions [3–6]. An earlier approach, Deep Multi-Species Embedding (DMSE), used neural-network models on eBird datasets to predict inter-species co-occurrence [7]. In addition, the analysis of photographs from citizen science projects have been used to extract species interactions [8]. Notably many pollinator plant interactions have been revealed this way [9, 10]. Studies continue to fill in ecological knowledge gaps using citizen science data from websites like iNaturalist [11, 12]. Despite the insight that has been gained by ecological-niche modeling, even the best ENMs can misestimate interaction potential between species [13]. One reason for this could be the tendency to make lots of assumptions within their models. With such a complex system, minimizing the number of assumptions made should help improve accuracy.

Citizen science data is usually far less reliable than data taken in the course of scientific studies and it also has a significant user bias [14, 15]. People tend to stay on trails, upload more iconic and obvious species, and mainly observe during daylight hours. Despite the lower quality of individual observations, there are many orders of magnitude more of them. The aim of this work is to come up with statistical techniques to handle data of this kind in order to extract useful biological information.

2 Methods

The first and simplest way to quantify inter-species correlations is to check if two species' data points tend to be closer together than expected by chance. If that is the case, this is an indication that there may be some correlation between them, which may suggest a potential for an inter-species interaction.

We start by considering two species i and j. Then we take an individual data point from species i and draw a circle around it. We ask if the number of data points from species j in that same circle is to be expected assuming that there is no interaction between species i and j. I discuss this approach in the results section. However, it is much less computationally efficient than simpler methods. The primary goal of this research is to extract information about species associations from very large datasets, computing correlations for millions of species-species pairs over large regions. I have focused on a more efficient method that gives us very similar information by tiling the entire area containing the dataset and looking at the presence or absence of each species in each tile.

The data itself is restricted to where users typically visit (see Fig. 1). The

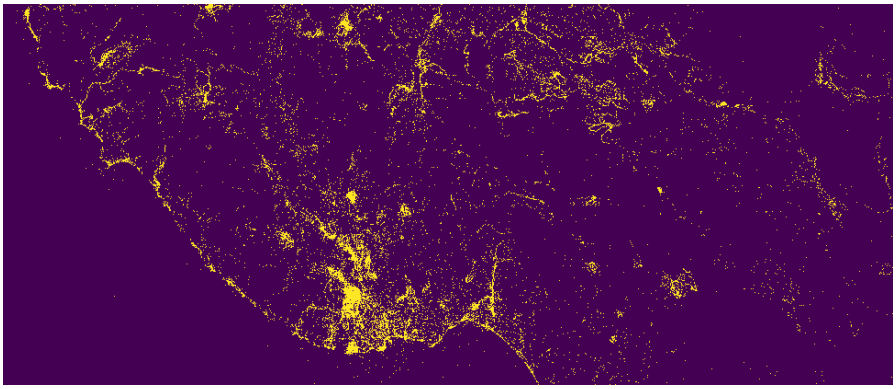

Figure 1: Tiled map of the Santa Cruz Mountains. The yellow tiles contain at least one observation. The purple tiles are empty.

clumping that is seen in the data is predominantly an artifact of the observers, and not an intrinsic ecological property. Take a wetland trail for example: Right in the same spot people might upload birds seen hundreds of meters away and might also upload small beetles crossing the trail. This is because the location data of the photos is predominantly confined to the trail. Most people will not try to fine tune the locations of all the birds they spot. This could easily lead to a seemingly significant but spurious relation between herons and marsh beetles, for example. An apparent correlation due to this sort of observational bias carries no useful ecological information. Of course they both inhabit the same ecosystem, but the goal here is to gain deeper insight into inter-species interactions. The primary focus of this method is determining a baseline to check whether the population of species j in the vicinity of species i is to be expected assuming there is no interaction between the two species.

The first step of this method is to tile the data into square areas. For this paper I chose tile sizes of 33 m as an estimate of average location data accuracy [16, 17].

Having multiple data points for the same animal leads to problems with the analysis: spuriously high correlations, and makes the data non-Poissonian. Only using the presence or absence of a species within a tile mitigates both of these issues. Therefore we employ a tiling method in which we mark present or absent per tile for each species rather than working with the raw number of data points.

The inter-species comparisons are then computed by going through each species pair in the dataset (if two species share at least one tile then they are compared). We define the overlap, which we denote as O_{ij}, as the number of tiles in which the two species both are present together.

Our null hypothesis is that the data for a single species is drawn from a spatial Poisson process [18]. This would mean that correlations in the data

that we see are due to spatial heterogeneity in observations and random noise, rather than a real association. More specifically, we consider two species i and j. Given the data points for i, we ask if the data points for j are well described as a spatial Poisson process. In other words, the quantities derived from the data, such as the overlap, O_{ij}, between i and j, are what are expected typically for species in geographical proximity.

In order to determine if species j has an atypically large number of tile overlaps with species i, one needs to first determine the typical overlap number. We can get a typical overlap number for a species i, which we will denote as μ_i, by considering other species that share geographical proximity with species i.

To get an estimate of a typical overlap with a given species i, we can make use of the fact that there are almost always many other species that overlap with it. We can use these other species to get an estimate for a typical overlap. To be more specific, we are considering all tiles where species i is present. For each of these we list the other species in the same tiles. Then we will obtain an average using these other species.

There is a problem caused by the varying tile count of species in the dataset. The total number of tiles occupied by a species varies greatly [19]. This leads to artifacts when performing averages that are discussed further in section 4.1. To account for this, we normalize by the tile count of each species j. Dividing the overlaps of i and j by the total tiles of j, which we denote as T_j, will remove this bias. We define the overlap density of species i with j as

$$\theta_{ij} \equiv \frac{O_{ij}}{T_j} \tag{1}$$

Then we can compute the average overlap density.

$$\overline{\theta_i} = \frac{\sum_{j=0}^{N_i} \theta_{ij}}{N_i} \tag{2}$$

If the null hypothesis is correct, we expect the overlap density for species i to be the average overlap over all the other species that we are comparing species i with. Therefore

$$\mu_i / T_i = \overline{\theta_i} \tag{3}$$

Therefore

$$\mu_i = T_i \frac{\sum_{j=0}^{N_i} (\frac{O_{ij}}{T_j})}{N_i} \tag{4}$$

if the data for species i was drawn from the same spatial Poisson process.

To assess whether the observed overlap O_{ij} is to be expected typically, we use a z-score. To get the z-score we need to get the standard deviation σ. Assuming the data is drawn from a spatial Poisson process, the variance σ^2 equals the mean μ_i. Therefore the z-score goes as follows

$$z = \frac{O_{ij} - \mu_i}{\sqrt{\mu_i}} \tag{5}$$

This measures how far O_{ij} is from its expected value μ_i in standard-deviation units. If z is sufficiently large, we reject the null hypothesis.

Given this equation to ascertain the z value, we only need to compute μ_i and O_{ij}. But in making the comparison between species i and j, we must choose which of the species is i and which is j, as the mathematical meaning of z is altered if we exchange i and j. In order to choose one we make a reasonable assumption that the rarer species inhabiting less tiles has a greater chance of relying on the more common species with more tiles. Therefore we choose the i species as the species with less tiles, $T_i < T_j$, as it should yield more accurate results.

In addition to a z-score which indicates significance, a level of correlation can also be quantified. In this case it is the overlap over the mean [20].

$$\rho_{ij} = \frac{O_{ij}}{\mu_i} \tag{6}$$

Fine tuning may be needed depending on the dataset. The subsequent results are based on a dataset where I removed species that inhabited 5 or less tiles from the dataset. The reason being that a species with such few data points will not be able to yield useful results and it may very well skew the means in an unfavorable way. Another adjustment made was omitting avian observations from the dataset. This was done because birds tend to have less obligate relations with other species. For avian species, this method is less likely to yield as many interesting interactions as are found with invertebrate species.

3 Results

This method was run on a dataset centered around Santa Cruz California (see Fig. 1). It was bound by the latitudes 37.2551 and 36.9209 and the longitudes -121.464 and -122.446. The study area contains $4,684,582$ tiles, arranged 3313×1414. This corresponds to 109.329 km (east–west) by 46.662 km (north–south).

This dataset was downloaded in June of 2025 from iNaturalist. It includes all recorded non-avian species and genus-level taxa from the given area that inhabit at least six or more tiles. Observations with obscured or private locations were omitted from this dataset as well.

The data analysis on this dataset was performed on an 8-GB MacBook Air (M1). The quality and insight gained should increase with larger data sets and this will be explored in future work.

3.1 Radial distribution function

A quantitative way of understanding the correlations between different species is to use the radial distribution function (RDF) [21]. The RDF for two species, i and j, gives the probability per unit area that for any observation of i, an observation of j is at a radius r from it. To compute the RDF over a maximum distance of R, one determines all distances between observations for i and observations for j and bins them according to this distance, normalizing by the total number of pairs of distance less than R and the area spanned by the corresponding bin. This gives the probability per pair, per unit area.

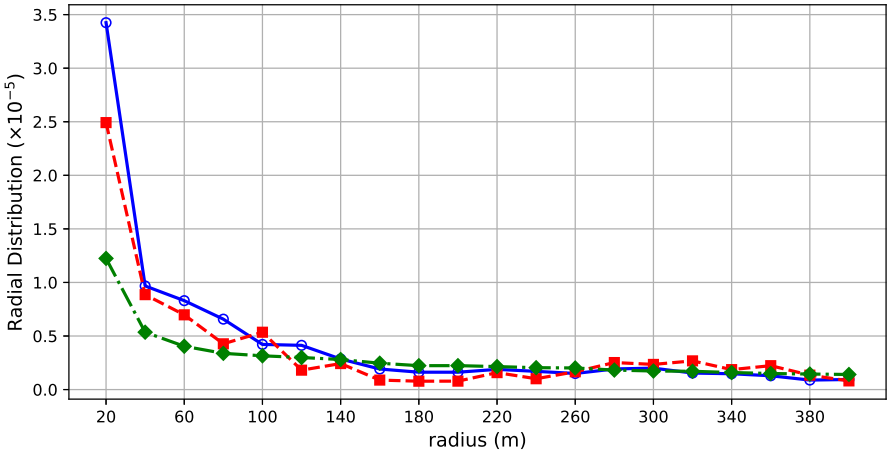

Figure 2: Normalized Radial Distribution functions for three species pairs.
Blue: California Yerba Santa × Yerba Santa Beetle
Red: Pacific Poison Oak × Yerba Santa Beetle
Green: Coast Redwood × Slender Banana Slug

Here I have highlighted three different species pairs to illustrate typical and atypical correlations. One of the strongest associations, that will be further discussed below, is between California Yerba Santa with the Yerba Santa Beetle, (solid line) in Fig. 2. The other two pairs, Pacific Poison Oak with Yerba Santa Beetle (dashed line) and Coast Redwood with Slender Banana Slug (dot dash) have weaker correlations that are still quite substantial. These distributions are typical of what one expects for species correlations and do not imply any biological association. The heightened probability density seen for small radii are due to the aforementioned factors, that species observations occur along trails, paths, and in more biologically favorable areas.

Table 1: The top eleven known inter-species correlations within our dataset sorted by significance value (Sig). The results are displayed in order of highest significance descending. The percentile column indicates the rank within the entire output dataset.

Species (i)	Overlapping Species (j)	Sig. (z)	Corr. (ρ)	(%)
Tamalia glaucensis	Big Berry Manzanita	37.52	21.24	99.98
Coyote Brush Bud Gall Midge	Coyote Brush	30.42	9.18	99.96
Coffeeberry Midrib Gall Moth	Coffeeberry	29.92	26.84	99.92
Coyote Brush Stem Gall Moth	Coyote Brush	28.94	12.26	99.90
Sagebrush Woolly Stem Gall Midge	California Sagebrush	28.64	13.65	99.88
White Sage Leaf Gall Midge	Black Sage	28.28	44.07	99.86
Ceanothus Bud Gall Midge	Wartleaf Ceanothus	28.28	44.07	99.84
Pumpkin Gall Wasp	Coast Live Oak	25.25	8.98	99.75
Red Cone Gall Wasp	Valley Oak	24.66	7.5	99.71
Yerba Santa Beetle	California Yerba Santa	24.62	21.5	99.69
California Gall Wasp	Valley Oak	24.31	5.46	99.65

3.2 Tables

Table 1 lists the top eleven known inter-species correlations within our dataset sorted by significance value, z. The rows are ordered from highest significance descending. It lists species i, species j, significance z, correlation ρ, and percentile ranking %. Significance, correlation, and percentile are rounded to two decimal places.

The majority of the highest scoring significances are known obligate relations. In fact all but one of these from Table 1 are cecidogenous (gall producing) species. This means whenever these are observed it is unlikely for them to be found in a spot without the host plant.

The Yerba Santa Beetle stands out for being the one species that is not cecidogenous. It represents many non-cecidogenous insect species that fully rely on a single host species / genus as a part of their life cycle. The Yerba Santa Beetle, *Trirhabda eriodictyonis*, relies fully on species from the genus *Eriodictyon* such as *Eriodictyon californicum*, California Yerba Santa. It feeds exclusively on *Eriodictyon* leaves as a larva and as an adult [22].

Table 2 lists the top six inter-species correlations between two species that

Table 2: The top six inter-species correlations between two species that share the same host plant. The results are displayed in order of highest significance descending. The percentile column indicates the rank within the entire output dataset.

Species (i)	Overlapping Species (j)	Sig. (z)	Corr. (ρ)	(%)
Convoluted Gall Wasp	Red Cone Gall Wasp	30.22	15.36	99.94
Spined Turban Gall Wasp	Red Cone Gall Wasp	27.69	9.19	99.82
Club Gall Wasp	Disc Gall Wasp	26.56	17.61	99.80
Yellow Wig Gall Wasp	Red Cone Gall Wasp	26.05	13.83	99.78
Honeydew Gall Wasp	Red Cone Gall Wasp	25.38	13.63	99.77

Table 3: The top three inter-species correlations between two species that do not appear to have any known inter-species interactions documented. The results are displayed in order of highest significance descending. The percentile column indicates the rank within the entire output dataset.

Species (i)	Overlapping Species (j)	Sig. (z)	Corr. (ρ)	(%)
Oregon Gumplant	Common Yarrow	24.81	9.13	99.73
Strigamia	Pacific Newts	24.55	10.77	99.67
Rockweed	Ochre Sea Star	24.16	6.8	99.63

share the same *Quercus* host plant within our dataset sorted by significance value, z. The rows are ordered from highest significance descending. It lists species i, species j, significance z, and correlation ρ, and percentile %. Significance, correlation, and percentile are rounded to two decimal places.

Some species end up correlating with other species because they have the same host plant. This was mainly seen with various oak associated gall wasps from the Tribe Cynipini. While strong user bias could be a reason for these results, there is a possibility that certain individual oaks have a greater chance of being infected by multiple cecidogenous wasp species [23].

Table 3 lists the top three inter-species correlations within our dataset between two species that do not appear to have any known inter-species interactions documented, sorted by significance value, z. The rows are ordered from highest significance descending. It lists species i, species j, significance z, and correlation ρ. Both significance and correlation are rounded to two decimal places.

With complex ecosystems and observational biases in the data, the reason for some of these strong correlations is sometimes difficult to interpret. Because the coast is such a thin span of area, the coastal data from this dataset is too small

to yield useful results. Therefore coastal relations should be ignored from these results. In the greater Santa Cruz area, Oregon Gumplant and Common Yarrow both generally inhabit coastal scrub ecosystems. It is likely the Rockweed and Ochre Sea Star association exists for a similar reason. If coastal ecosystems were to be explored more with this method, a much larger dataset would be needed.

Strigamia, also found in Table 3, is a genus of Soil Centipedes (Order Geophilomorpha) that can be found in damp areas. It is not clear why it would correlate so strongly with Pacific Newts. One likely explanation is that in the Santa Cruz Mountains, next to the Lexington Reservoir, there is the Alma Bridge Road Newt Passage Project which has been documenting Pacific Newts and Pacific Newt roadkill instances along Alma Bridge Road [24]. Throughout this 5.4 km section of road, numerous unidentifiable Pacific Newt roadkill instances have been documented and marked as "Pacific Newts" on iNaturalist. Within this same section many *Strigamia* roadkill have been observed as well. Because of the large scale of this citizen science project, the quantity of roadkill has caused a significant amount of co-occurrence for these species. It seems that because of this data anomaly, a strong association between *Strigamia* and Pacific Newts has shown up in the results. A larger dataset should eliminate this artifact, unless there is a true unknown relation, which is possible but unlikely.

4 Discussion

This method is different from previous methods in the way it is able to filter correlations that are an artifact of the data. There are two kinds of reasons for why things appear to correlate when they do not. The first reason is user bias [25]. The second reason is that species will be more prevalent in certain regions rather than others, but this does not mean to say that they interact [26, 27]. In both cases a baseline is needed for what one would typically expect. Other work has not properly addressed this problem and will therefore yield inter-species interactions that are not as reliable. It is for this reason that this method is an important addition to modeling ecological data.

Additions to this method to characterize microhabitats for each species would be interesting. While a species of beetle might not correlate very highly with any single plant species, there is a chance it may find a particular combination of plants most habitable. That information could be revealed by the data given the right methodology. Incorporating time, elevation, and weather into a larger method could help yield further insight into ecosystem functionality.

There is also the potential to use this method to help in the delineation of ecoregions alongside algorithms that involve data clustering [28].

4.1 Exploring different estimates for the mean

I have considered multiple ways of obtaining an average, μ_i. Because this average is a crucial part of the methodology, it is important to explain why seemingly simpler alternatives fail. We can take a species i and average every O_{ij} for each

overlapping species j. Let N_i denote the number of j species we compare to species i. The formula for the average overlap $\mu_i^{(O)}$ is

$$\mu_i^{(O)} = \frac{\sum_{j=0}^{N_i} O_{ij}}{N_i} \tag{7}$$

The reason for not using this average is because of the very broad distribution of species abundance that is best described by a log series distribution [19]. The degree to which a species' tiles overlap will correlate strongly with their abundance. The very broad form of this abundance distribution will make the average of the mean abundance much larger than other measures, such as the median. This means that species with high abundance will dominate the mean in Eq. 7 because of high value outliers. This leads to a greatly reduced mean statistical power compared to the alternative definition Eq. 4 that mitigates these outlier effects. In that formula the averaging of O_{ij} employed a division by T_i inside the computation of the average, which suppresses the effects of these high abundance outliers.

5 Conclusion

This paper has introduced a highly efficient method of getting likely candidates for strong inter-species interactions that can be extracted from citizen science datasets. This was done by constructing a method for filtering out specious correlations and devising a computational method that is very efficient. This method was run on a dataset of $15,975$ species, and $874,263$ data points, which has potentially $127,592,325$ interactions, using only modest computer resources. The results obtained find known interactions and some other ones that are likely artifactual but may actually have some ecological significance.

Future work could include: looking at spatial-temporal correlations, more comprehensive analysis of the entire database, using similar methods to better delineate biomes, and the analysis of species microhabitat.

6 Acknowledgments

I thank Prof. James Helfield and Prof. Merrill Peterson for useful discussions, and J.M. Deutsch for useful discussions and a critical reading of the manuscript.

Code availability

All analysis code is available at `https://github.com/Jacob-Deutsch-Work/Finding-Inter-species-Associations-on-Large-Citizen-Science-Datasets`

References

[1] Jordi Bascompte and Pedro Jordano. *Mutualistic Networks*. Number 53 in Monographs in Population Biology. Princeton University Press, Princeton, 2014. ISBN 978-0691131269. doi: 10.1515/9781400848720.

[2] Brittany M. Mason, Thomas Mesaglio, Jackson Barratt Heitmann, Mark Chandler, Shawan Chowdhury, Simon B. Z. Gorta, Florencia Grattarola, Quentin Groom, Colleen Hitchcock, Levi Hoskins, Samantha K. Lowe, Marina Marquis, Nadja Pernat, Vaughn Shirey, Shukherdorj Baasanmunkh, and Corey T. Callaghan. inaturalist accelerates biodiversity research. *BioScience*, 2025. doi: 10.1093/biosci/biaf104. URL `https://doi.org/10.1093/biosci/biaf104`. Advance article; published 28 July 2025.

[3] A. Márcia Barbosa, Neftalí Sillero, Fernando Martínez-Freiría, and Raimundo Real. Ecological niche models in mediterranean herpetology: Past, present and future. *Ecological Niche Models in Mediterranean Herpetology*, pages 173–202, 2012. doi: 10.13140/2.1.3746.6560.

[4] Wilfried Thuiller. Ecological niche modelling. *Current Biology*, 34:R217–R236, 2024. doi: 10.1016/j.cub.2024.02.018.

[5] Jane Elith and John R. Leathwick. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009. doi: 10.1146/annurev.ecolsys.110308.120159.

[6] Steven J. Phillips, Robert P. Anderson, and William Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006. doi: 10.1016/j.ecolmodel.2005.03.026.

[7] Di Chen, Yexiang Xue, Shuo Chen, Daniel Fink, and Carla P. Gomes. Deep multi-species embedding. *arXiv preprint arXiv:1609.09353*, 2016.

[8] Quentin Groom, Nadja Pernat, Tim Adriaens, Maarten de Groot, Sven D. Jelaska, Diana Marčiulynienė, Angeliki F. Martinou, Jiří Skuhrovec, Elena Tricarico, Ernst C. Wit, and Helen Roy. Species interactions: Next level citizen science. *Ecography*, 44(12):1781–1789, 2021. doi: 10.1111/ecog.05790.

[9] Camila Bosenbecker, Pedro Amaral Anselmo, Roberta Zuba Andreoli, Gustavo Hiroaki Shimizu, Paulo Eugênio Oliveira, and Pietro Kiyoshi Maruyama. Contrasting nation-wide citizen science and expert collected data on hummingbird–plant interactions. *Perspectives in Ecology and Conservation*, 21(2):164–171, 2023. doi: 10.1016/j.pecon.2023.03.004.

[10] Milan Gazdic and Quentin Groom. inaturalist is an unexploited source of plant–insect interaction data. *Biodiversity Information Science and Standards*, 3:e37303, 2019. doi: 10.3897/biss.3.37303.

[11] Breanna J. Putman, Riley Williams, Enjie Li, and Gregory B. Pauly. The power of community science to quantify ecological interactions in cities. *Scientific Reports*, 11:3069, 2021. doi: 10.1038/s41598-021-82491-y.

[12] Fang-Shuo Hu, Yun Hsiao, and Alexey Solodovnikov. A global citizen science effort via inaturalist reveals food webs of large predatory rove beetles. *Food Webs*, 43:e00399, 2025. doi: 10.1016/j.fooweb.2025.e00399.

[13] Xuanye Wen, Guofei Fang, Shouquan Chai, Chuanjie He, Shouhui Sun, Guanghua Zhao, and Xiao Lin. Can ecological niche models be used to accurately predict the distribution of invasive insects? a case study of *Hyphantria cunea* in china. *Ecology and Evolution*, 14:e11159, 2024. doi: 10.1002/ece3.11159.

[14] Erin D. Brown and Byron K. Williams. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology*, 33(3):561–575, 2019. doi: 10.1111/cobi.13223.

[15] Nick J. B. Isaac, Arco J. van Strien, Tom A. August, Martijn P. de Zeeuw, and David B. Roy. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, 5(10):1052–1060, 2014. doi: 10.1111/2041-210X.12254.

[16] Paul A. Zandbergen and Sean J. Barbeau. Positional accuracy of assisted gps data from high-sensitivity gps-enabled mobile phones. *Journal of Navigation*, 64(3):381–399, 2011. doi: 10.1017/S0373463311000051.

[17] Krista Merry and Pete Bettinger. Smartphone gps accuracy study in an urban environment. *PLOS ONE*, 14(7):e0219890, 2019. doi: 10.1371/journal.pone.0219890.

[18] Adrian Baddeley, Imre B'ar'any, and Rolf Schneider. *Stochastic geometry: lectures given at the CIME summer school held in Martina Franca, Italy, September 13–18, 2004*. Springer, 2007.

[19] Elita Baldridge, David J. Harris, Xiao Xiao, and Ethan P. White. An extensive comparison of species-abundance distribution models. *PeerJ*, 4: e2823, December 2016. doi: 10.7717/peerj.2823. URL `https://doi.org/10.7717/peerj.2823`.

[20] Joseph A. Veech. A probabilistic model for analyzing species co-occurrence. *Global Ecology and Biogeography*, 22(2):252–260, 2013. doi: 10.1111/j.1466-8238.2012.00789.x.

[21] Jean-Louis Burgot. Radial distribution function. In *The notion of activity in chemistry*, pages 329–335. Springer, 2016.

[22] Katherine Gould and Paul Wilson. Lack of evolution in a leaf beetle that lives on two contrasting host plants. *Ecology and Evolution*, 5(18):3905–3913, 2015. doi: 10.1002/ece3.1658. URL `https://doi.org/10.1002/ece3.1658`.

[23] Ramón Perea, Rodolfo Dirzo, Stephanie Bieler, and Geraldo Wilson Fernandes. Incidence of galls on sympatric california oaks: Ecological and physiological perspectives. *Diversity*, 13(1), January 2021. ISSN 1424-2818. doi: 10.3390/d13010020. URL https://www.mdpi.com/1424-2818/13/1/20.

[24] Midpeninsula Regional Open Space District. Alma Bridge Road Newt Passage Project, 2023. URL https://www.openspace.org/what-we-do/projects/newt-passage. Project page.

[25] Ellyne M. Geurts, John D. Reynolds, and Brian M. Starzomski. Turning observations into biodiversity data: Broadscale spatial biases in community science. *Ecosphere*, 14(6):e4582, June 2023. doi: 10.1002/ecs2.4582. URL https://doi.org/10.1002/ecs2.4582.

[26] F. Guillaume Blanchet, Kevin Cazelles, and Dominique Gravel. Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23 (7):1050–1063, July 2020. doi: 10.1111/ele.13525. URL https://doi.org/10.1111/ele.13525. Epub 2020 May 19.

[27] Marta Goberna and Miguel Verdú. Cautionary notes on the use of co-occurrence networks in soil ecology. *Soil Biology and Biochemistry*, 166: 108534, 2022. doi: 10.1016/j.soilbio.2021.108534. URL https://doi.org/10.1016/j.soilbio.2021.108534.

[28] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205.