ANALIZA PODATKOV OBČANSKE ZNANOSTI V STATISTIKI IN PODATKOVNI ZNANOSTI

26. NOVEMBER 2025 OB 17. URI
CENTRALNA TEHNIŠKA KNJIŽNICA

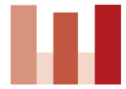
Analysis of citizen science data in statistics and data science

Andrej Srakar, Asst. Prof., Jožef Stefan Institute (Dept. of AI) and School of Economics and Business, University of Ljubljana; Coordinator, YoungStatS project and Young Statisticians Europe (YSE), FENStatS association; Section Officer, Royal Statistical Society, UK

Elizabeth (Betsy) Bersson, Postdoctoral Fellow, Prof. Tamara Broderick's Research Group, Massachusetts Institute of Technology (MIT), United States

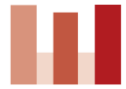
Outline of the presentation

- Shortly about citizen science
- Analysis of citizen science data
- First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets
- Presentation by Betsy
- Third example paper: Dennis et al. 2025
- Fourth example paper: Koh and Opitz 2025
- Discussion, reflections and conclusion



Shortly about citizen science

- Citizen science has emerged as an interesting novel paradigm in scientific process.
- Emergence of citizen science is quite recent and the term citizen science was used for the first time in the late 1980s. The first book that theorized the concepts to understand and develop it was published in the mid-1990s.
- Since then interest in and acknowledgments of citizen science have grown. However, despite the large volume of work, it is impossible to define citizen science for all contexts and there is still not an international consensus about its definition.



Shortly about citizen science

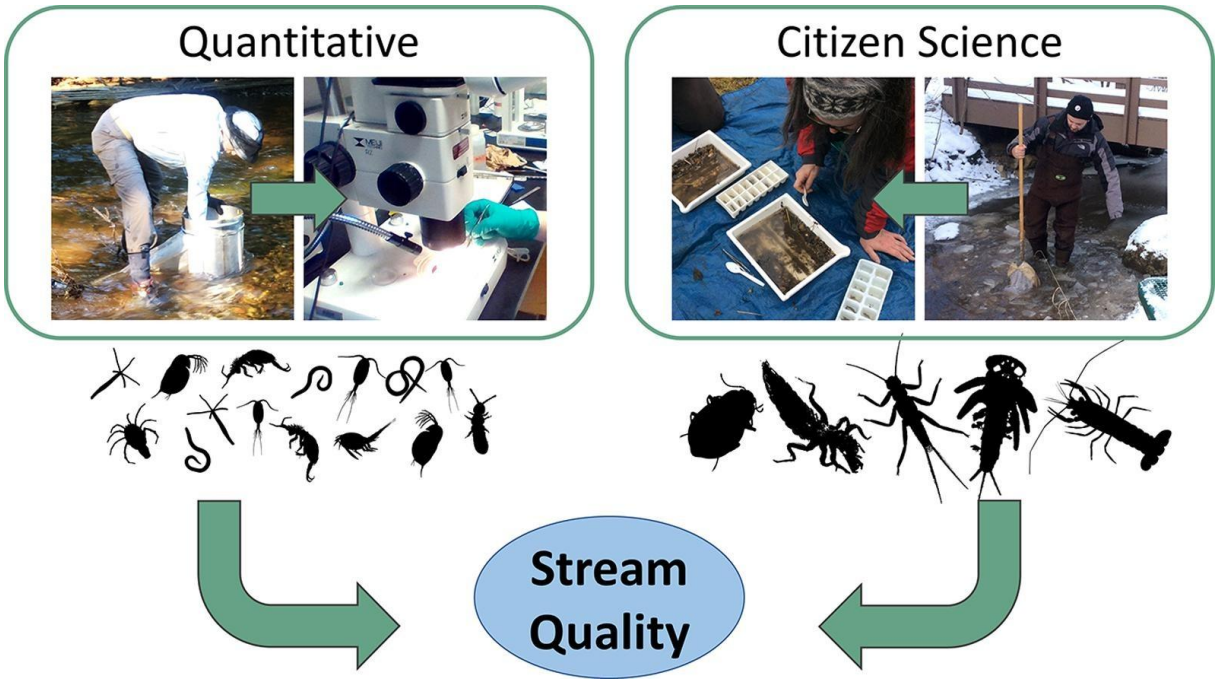
- In the United States citizen science was defined at the national level (by J. Holdren) as a process where the public participates voluntarily in the scientific process, addressing real-world problems in ways that may include formulating research questions, conducting scientific experiments, collecting and analyzing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems.
- The European Commission has used the definition from the Oxford English Dictionary, defining citizen science as scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions. Citizen science is often linked with outreach activities, science education or various forms of public engagement with science.

Shortly about citizen science

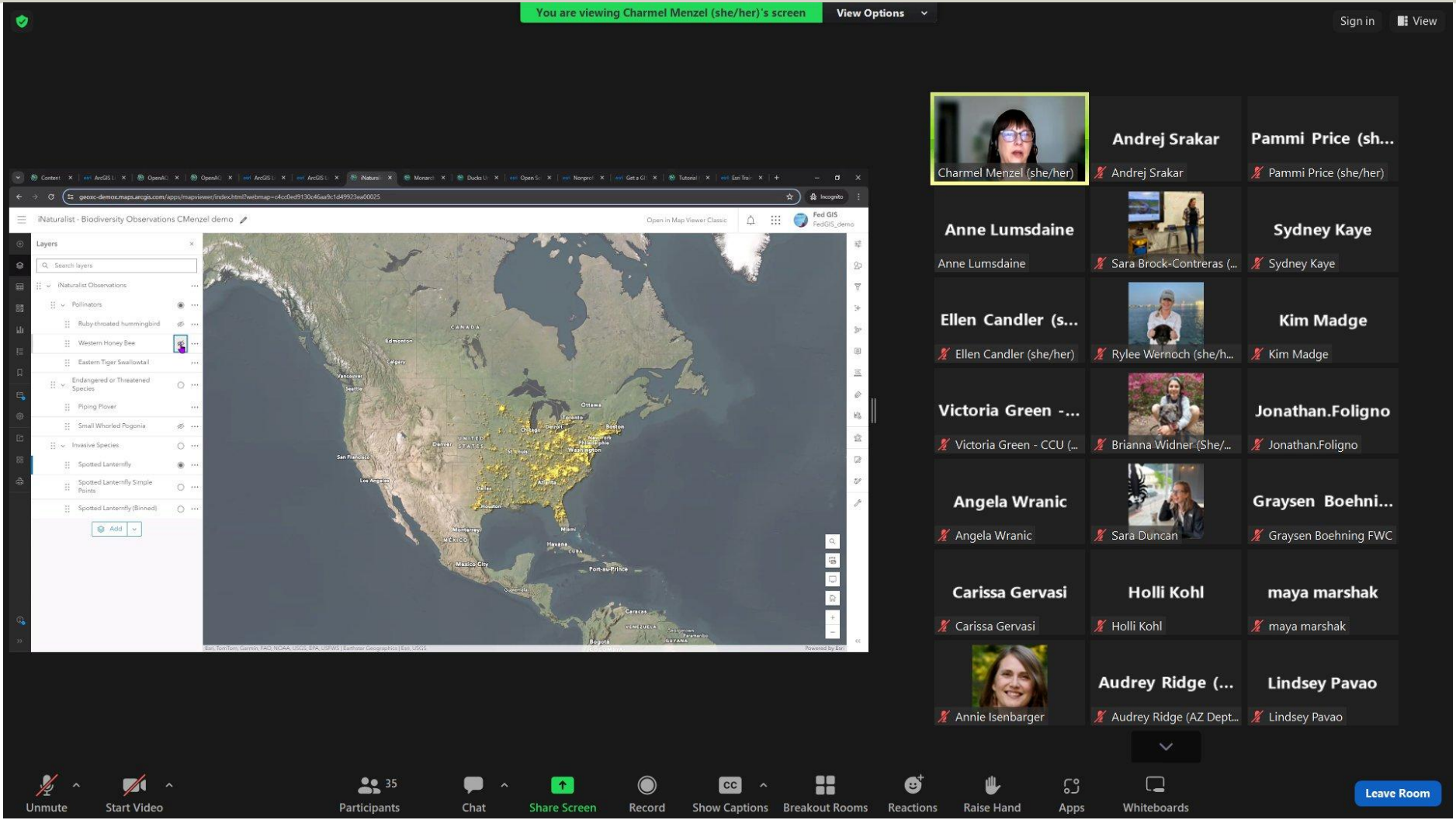
- As citizen science includes data gathering processes and data analysis it is important for a statistician to ask if using common statistical approaches such as regression, classification, dimension reduction or correlation analysis can be used with citizen science data.
- Traditionally a way for amateur experts to participate in a structured and complex survey, many citizen science projects are now characterised by large-scale unstructured surveys. Increasingly accessible technology and devices, along with artificial intelligence and machine learning, are being used to complement the process of gathering data by anyone.
- This has many positive implications for engaging people with nature but has also led to a more complex landscape of projects and biases.
- Still a few analyses comprehensively address sources of bias in the analysis of citizen science data, typically focussing on observer-based sources, while there is inconsistent use of terms such as detectability or availability which hinders cross-study comparisons.

Shortly about citizen science

Image: Citizen science data are a reliable complement to quantitative ecological assessments in urban rivers (source: Krabbenhoft and Kashian, 2020).



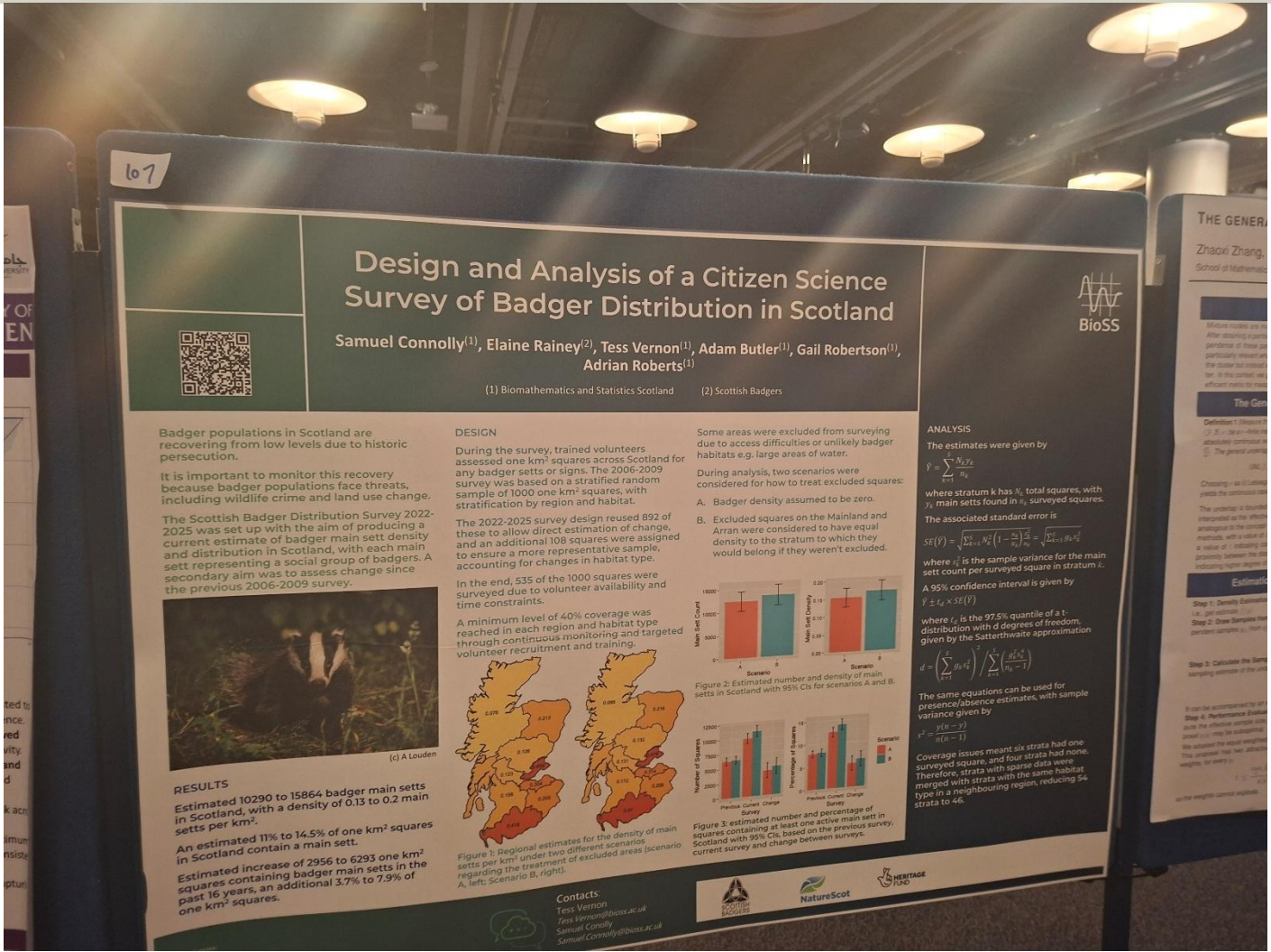
Shortly about citizen science



Analysis of citizen science data

- When is citizen science data a good way to study a statistical problem? As one example, recently awarded at the RSS 2025 conference in Edinburgh as best poster, volunteers were engaged in a study of Scottish badgers.
- They were engaged in a stratified survey, adding to previous ones conducted by professional interviewers. Volunteers were assigned their territories to gather the required badger data. Number of volunteers and their easier decisions to gather data and cross the country allowed more broad access to data.
- When comparing results with the baseline survey carried out in 2006-2009 it was estimated that there has been an increase of between 2956 and 6293 1-km squares containing badger main setts over the past 16 years. This is equivalent to an additional 3.7% to 7.9% of 1-km squares in Scotland containing badger main setts since the previous survey.
- However, it isn't advisable to estimate the increase in badger population based on main sett estimates, which relates to common biases in results of analyses using citizen science data.
- We can not control the exactness of the work of a volunteer and his or her work can be biased.

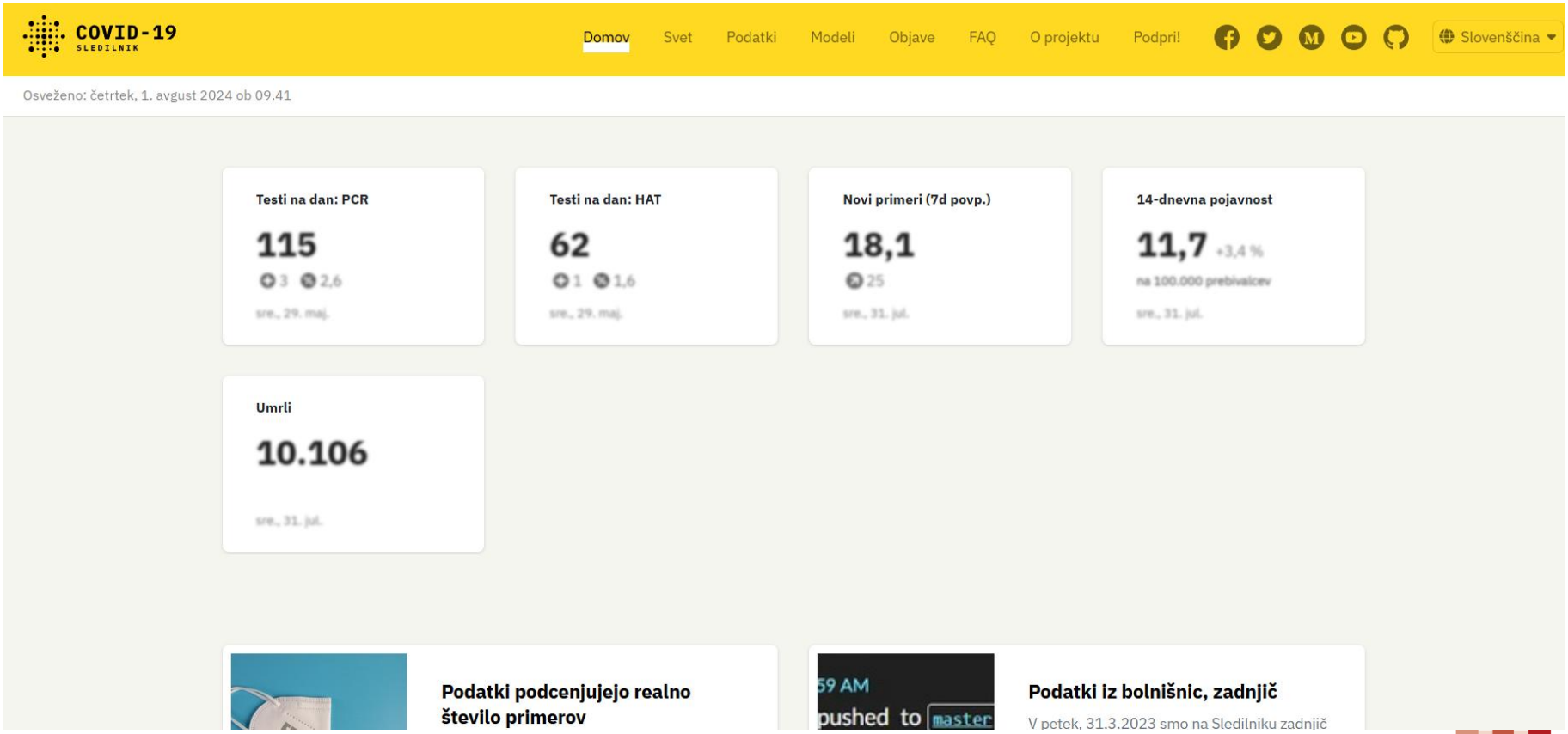
Discussion, reflections and conclusion



Analysis of citizen science data

- Second example: pandemic-time Slovenian multiply awarded online platform volunteer-based initiative COVID-Sledilnik.
- At the start of the pandemic it turned out that limitations in access to necessary data to monitor the pandemic and adopt appropriate measures in Slovenia were severe. Many scientists have joined the starting initiative of Slovenian entrepreneur and computer expert Luka Renko who collected simple and at the time very short data on pandemic measures and included it in an openly accessible Excel table.
- This has lead to a larger scale scientific movement which lasted throughout the pandemic time and allowed all Slovenian citizens to get access to large-scale datasets, combined, monitored and computationally processed by volunteers (frequently leading younger scientists of various professions).
- Never though have the analysis of COVID-Sledilnik data been verified and tested to robustness against citizen science biases. But it shows the possible scope and importance of citizen science-collected and accessible data.

Analysis of citizen science data



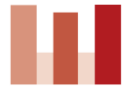
Analysis of citizen science data

- At the 2024 Royal Statistical Society annual conference in Brighton, UK, we conducted a multi-paper discussion meeting on the Analysis of Citizen Science Data with three papers presented and an intense follow-up debate.
- Emily B. Dennis and coauthors presented illustrations of computationally efficient methods for fitting new models applied to three major citizen science data sets for butterflies.
- Elizabeth Bersson and Peter D. Hoff presented a nonparametric framework to obtain precise prediction sets for citizen science data sets based on „indirect“ information.
- Jonathan Koh and Thomas Opitz developed a spatiotemporal Bayesian hierarchical model for bias-corrected estimation of arrival dates of the first migratory bird individuals at their breeding sites.

Analysis of citizen science data

Analysis of citizen science data

- We continued these debates in another special session in the framework of the recent RSS annual conference in Edinburgh.
- If you would want to use such data in your study or article a statistician might advise you to be careful.
- At present, understudied biases that such analyses bring can turn the results wrong.
- It is urgent to provide solid methodological responses ahead of trusting the statistics in such results.



First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

Received: 5 July 2021 | Accepted: 6 December 2022

DOI: 10.1111/biom.13816

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC PRACTICE

Fast Bayesian inference for large occupancy datasets

Alex Diana[1] | **Emily Beth Dennis**[2,1] | **Eleni Matechou**[1] | **Byron John Treharne Morgan**[1]

[1]School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK

[2]Butterfly Conservation, Manor Yard, East Lulworth, Wareham, Dorset, UK

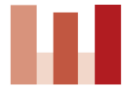
Correspondence
Alex Diana, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7FS, UK.
Email: ad625@kent.ac.uk

Abstract

In recent years, the study of species' occurrence has benefited from the increased availability of large-scale citizen-science data. While abundance data from standardized monitoring schemes are biased toward well-studied taxa and locations, opportunistic data are available for many taxonomic groups, from a large number of locations and across long timescales. Hence, these data provide opportunities to measure species' changes in occurrence, particularly through the use of occupancy models, which account for imperfect detection. These opportunistic datasets can be substantially large, numbering hundreds of thousands of sites, and hence present a challenge from a computational perspective, especially within a Bayesian framework. In this paper, we develop a unifying framework for

First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

- Robust measures of biodiversity change are vital for monitoring the varying state of species' populations and evaluating progress of conservation actions, for example, toward national and international targets (Butchart et al., 2010).
- Data from standardized, long-running monitoring schemes are used to produce estimates of species' status and trends, particularly in terms of changes in abundance.
- However, such data sources are limited taxonomically and geographically.
- By their nature of intensive, formal sampling they may be limited in spatial coverage and therefore cannot always be used to appropriately measure changes in species' distributions over time.



First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

- Conversely, opportunistic records of occurrence are often and increasingly available in large quantities for extensive geographic areas and time periods, and for a wide variety of taxa.
- However, opportunistic data are inherently biased (Isaac & Pocock, 2015). Data are typically presence-only, where records only indicate where and when a species is seen rather than including information on non-detection, unless complete lists are recorded.
- Data recording the distribution of animals and plants are frequently analyzed using occupancy models (MacKenzie et al., 2018), as they allow for imperfect detection. Applying such models to presence-only data requires non-detections to be inferred from the observations of other species (Kéry et al., 2010).
- Data of this nature are not standardized, and result from the submission and collation of records by citizen scientists who choose where, when, and what to record, but are often available in large quantities.

First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

- In addition to imperfect detection, modeling approaches for occurrence data of this type also need to account for spatial and temporal autocorrelation (Guélat & Kéry, 2018; Strebel et al., 2022).
- In this case, Bayesian hierarchical models are an appropriate choice, thanks to the available tools for accounting for and inferring the effects of site and time-specific random effects.
- However, Bayesian inference is computationally demanding, in particular when model-fitting involves large numbers of latent variables.
- Efficient model-fitting is increasingly important with the ongoing growth in the volume of biological recording data, partly due to increasing participation through new technologies and platforms for data submission (August et al., 2015).
- Fast inference is also motivated by the increasing desire to update species' trend estimates frequently, in order to inform the measuring and reporting of biodiversity change.

First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

- In this paper authors model spatial and temporal autocorrelation using a Gaussian process (GP) approach (Rasmussen & Williams, 2006).
- The advantage of using a GP is that it allows to naturally model spatial autocorrelation between sites sampled at continuous locations, which is typically the case for opportunistic data, and allows for a different degree of correlation between sites according to their distance, even if they are neighboring. They also model temporal autocorrelation within a GP framework.
- They cast the occupancy and detection process within a logistic regression framework, and take advantage of the *efficient Polya-Gamma augmentation scheme* (Polson et al., 2013) for inference.
- In addition they describe and compare different (low-rank and sparse) approximations for the GP, subset of regressors (SoR) (Smola & Bartlett, 2001) and nearest neighbor GPs (NNGPs) (Datta et al., 2016), and demonstrate how they can be used within a Polya-Gamma framework.
- This responds to the need for a computationally efficient approach to analyze presence-absence data arising from a large number of sites, while accounting for spatial and temporal autocorrelation.

First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

- Hierarchical representation of the Bayes model:

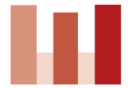
$$\begin{cases} y_i \sim Be(p_i z_{k_i}) & z_j \sim Be(\psi_j) \\ l(\psi_j) = \mu^{\psi} + b_{t_j} + a_{s_j} + X_j^C \beta^{\psi} + \epsilon_{s_j} & \mu^{\psi} \sim N\left(\mu_0^{\psi}, \sigma_0^{\psi}\right), \beta^{\psi} \sim N(0, \phi^{\psi} I), \epsilon_s \sim N(0, \sigma_\epsilon^2), \sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon) \\ l(p_i) = u_{t_{b_{k_i}}} + X_i \beta^p, & u_t \sim N(\mu_0^p, \sigma_0^p), \beta^p \sim N(0, \phi^p I) \\ (b_1, \dots, b_Y) \sim N\left(0, K_{l_T, \sigma_T}(\omega_1, \dots, \omega_Y)\right), & \sigma_T \sim IG(a_{\sigma_b}, b_{\sigma_b}), l_T \sim Gamma(a_{l_T}, b_{l_T}) \\ (a_1, \dots, a_S) \sim N\left(0, K_{l_S, \sigma_S}(x_1, \dots, x_S)\right), & \sigma_S \sim IG(a_{\sigma_S}, b_{\sigma_S}), l_S \sim Gamma(a_{l_S}, b_{l_S}) \end{cases}$$

- A random variable ω has a Polya-Gamma distribution, i.e. $\omega \sim PG(d, c)$ if $\omega = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-\frac{1}{2})^2 + \frac{c^2}{4\pi^2}}$, where $g_k \sim Gamma(d, 1)$.
- Full conditional distributions used for the Gibbs sampler are:

$$(\omega_i | \beta) \sim PG(d_i, X_i \beta) \qquad i = 1, \dots, n$$

$$(\beta | y, \omega) \sim N\left(\left(X^T \Omega X + B^{-1}\right)^{-1} \left(X^T k + B^{-1} b\right), \left(X^T \Omega X + B^{-1}\right)^{-1}\right),$$

- where $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$ and $k = (y_1 - \frac{d_1}{2}, \dots, y_n - \frac{d_n}{2})$.



First example paper: Diana et al. 2022 Fast Bayesian inference for large occupancy datasets

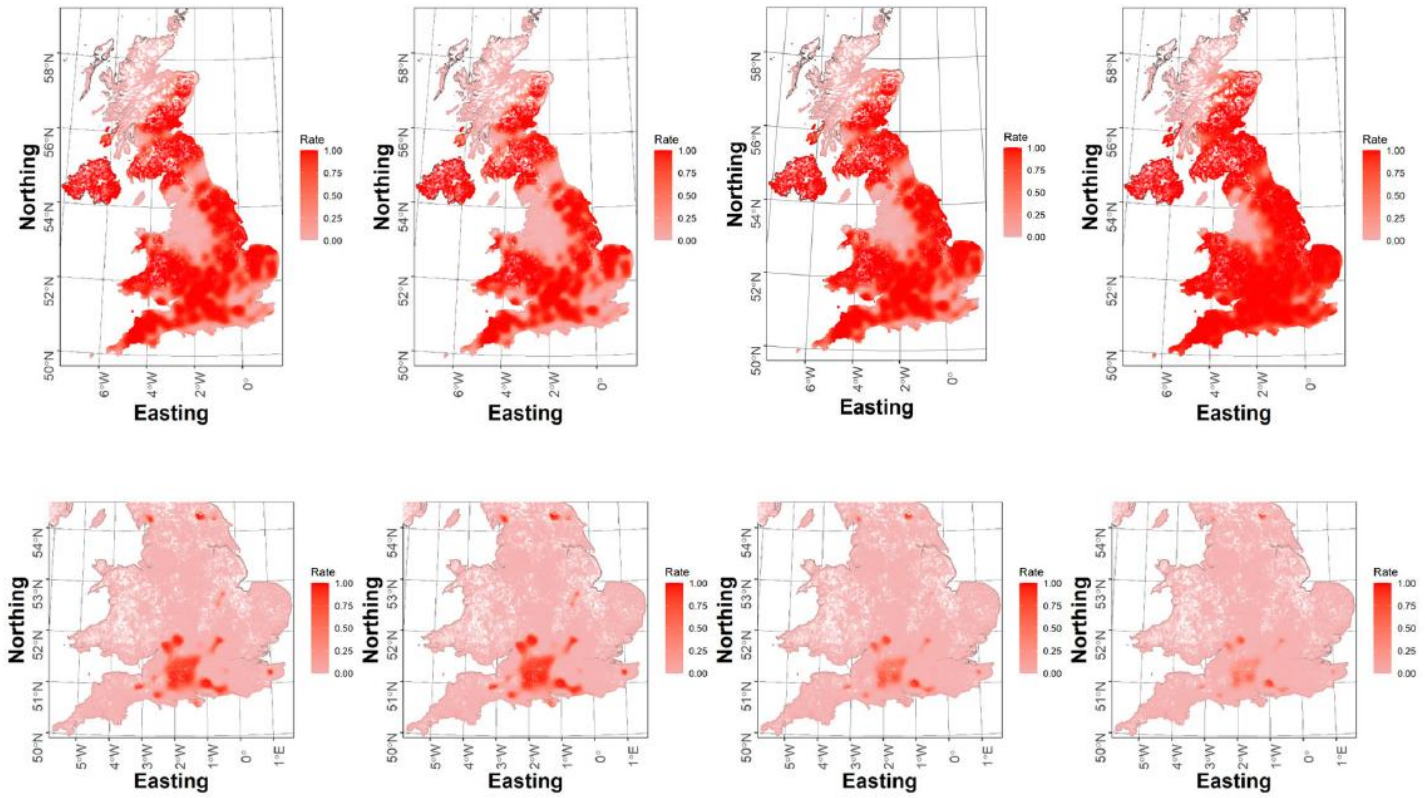


FIGURE 3 Posterior medians of the site-specific occupancy probabilities for Ringlet (top row) and Duke of Burgundy (bottom row) for 1970 (first column), 1985 (second column), 2000 (third column), and 2014 (fourth column). White areas represent parts of the country with no records of any butterfly species. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

Presentation by Betsy

Elizabeth (Betsy) Bersson

I am a postdoctoral fellow in Tamara Broderick's group at MIT. In May 2024, I received my PhD from the Department of Statistical Science at Duke University under the supervision of Peter Hoff. My work is related to hierarchical modeling, covariance estimation, small area estimation, and conformal prediction.

Outside of statistics, I enjoy reading, going on walks, particularly through gardens, and ~~vegetable gardening~~ (no yard in Boston).

News

- September 2025: Check out an invited session I organized on citizen science data analysis at RSS 2025 in Edinburgh.
- June 2025: Talk on Covariance Meta Regression at BNP in LA.
- April 2025: Invited talk on Covariance Meta Regression at the Department of Statistics seminar at U Mass Amherst.

Third example paper: Dennis et al. 2025

Journal of the Royal Statistical Society Series A, 2024, 1–17
doi: DOI HERE

Efficient statistical inference methods for assessing changes in species' populations using citizen science data

Emily B. Dennis,[1,2] Alex Diana,[3] Eleni Matechou[2] and Byron J.T. Morgan[2,*]

[1] Butterfly Conservation, Manor Yard, East Lulworth, BH20 5QP, Dorset, UK , [2] University of Kent, School of Mathematics, Statistics and Actuarial Science, Canterbury, CT2 7NF, Kent, UK and [3] University of Essex, School of Mathematics, Statistics and Actuarial Science, Wivenhoe Park, Colchester, CO4 3SQ. Essex, UK

*Corresponding author. B.J.T.Morgan@kent.ac.uk

Abstract

The global decline of biodiversity, driven by habitat degradation and climate breakdown, is a significant concern. Accurate measures of change are crucial to provide reliable evidence of species' population changes. Meanwhile citizen science data have witnessed a remarkable expansion in both quantity and sources and serve as the foundation for assessing species' status. The growing data reservoir presents opportunities for novel and improved inference but often comes with computational costs: computational efficiency is paramount, especially as regular

Third example paper: Dennis et al. 2025

- Citizen science data from systematic, designed surveys and monitoring schemes, collected by skilled and committed volunteers, have long been used to produce estimates of species' status, whereas observations of species from less structured, opportunistic or mass-participation sampling, attracting contributions from wider society, are increasingly gathered (Pocock et al., 2017), broadening the scope to measure changes in populations from extensive geographic areas and for a variety of taxa.
- The many sources of CS data provide vast opportunities for biodiversity monitoring, but require suitable analytical approaches, for example to deal with sources of bias (Isaac and Pocock, 2015).
- Johnston et al. (2023) outlined a diversity of challenges for biodiversity monitoring using CS data, relating to dealing with observer behaviour, data structures, statistical models, and communication.
- This paper focuses upon one such key challenge, which is in the computational cost of analysing CS data, particularly with increasing data and complex models.

Third example paper: Dennis et al. 2025

- For a given species, in any year we suppose that counts are obtained at S sites, each visited on at most V occasions. Each count, $y_{s,v}$, for site s and visit v is regarded as the realization of a random variables, such as Poisson, with expectation $\lambda_{s,v} = N_s a_{s,v}$, where the likelihood then takes the form:

$$L(N, \theta; y) \propto \prod_{s=1}^{S} \prod_{v=1}^{V} \exp(-N_s a_{s,v}) \left(N_s a_{s,v}\right)^{y_{s,v}}.$$

- Authors expand the Poisson likelihood above to counts $y_{s,v,r}$ where r denotes one of Y successive years, with a corresponding expansion to $a_{s,v,r}$. Then, expression for the annual model $N_{s,r} = e^{\alpha_s + \beta_r}$ is incorporated which results in the following epression for the log-likelihood, ignoring an additive constant:

$$l(\alpha, \beta, \theta; y) = \sum_{r} \sum_{s} \sum_{v} \{ - e^{(\alpha_s + \beta_r)} a_{s,v,r} + y_{s,v,r}(\alpha_s + \beta_r) + y_{s,v,r} \log\left(a_{s,v,r}\right) \}$$

Third example paper: Dennis et al. 2025

- Authors use a concentrated likelihood approach to form maximum-likelihood parameter estimates efficiently, by concentrating out the parameters α, resulting in the log-likelihood:

$$l(\beta, \theta; y) = \sum_r \sum_s \sum_v \left[-\frac{y_{s,.,.} e^{\beta_r} a_{s,v,r}}{\sum_j e^{\beta_r} a_{s,v,r}} + y_{s,v,r} \left(\beta_r - \log(\sum_j e^{\beta_r} a_{s,.,j}) \right) \right]$$

- Computational efficiency is addressed using variational inference (VI) which has earlier and in many studies been proposed as an alternative tool to overcome the computational issues of MCMC (Jordan et al., 1999).
- VI is traditionally faster than MCMC-based approaches because it transforms the problem from sampling from a posterior distribution to optimization. Therefore, VI combines the speed of classical inference, with the interpretability of Bayesian inference.
- However, while MCMC-based inference always recovers the true posterior (given enough MCMC iterations), in VI the true posterior distribution is approximated using an appropriate flexible family of distributions, which is called the variational family.

Third example paper: Dennis et al. 2025

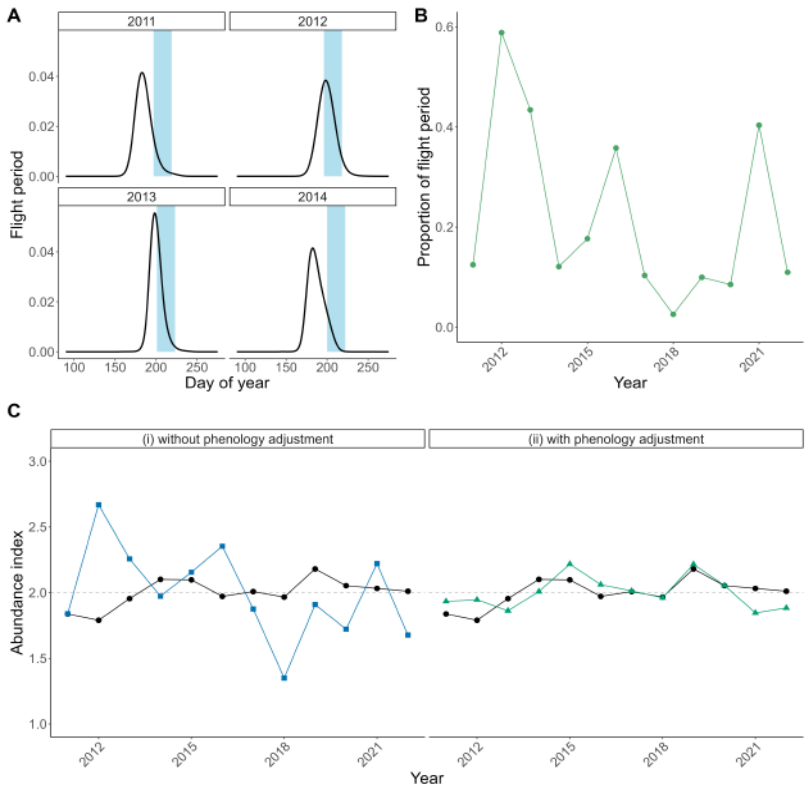


Fig. 2: Demonstration of the phenology adjustment approach for the Marbled White butterfly: A) flight period curves estimated from UKBMS data for four years. The blue shaded areas represents the BBC sampling period each year. B) the proportion of the Marbled White flight period covered by the BBC sampling period each year. C) relative abundance indices produced from the GAI applied to UKBMS data (black), from BBC data without phenology adjustment (i, blue squares), and from BBC data with phenology adjustment (ii, green triangles). Indices are on the $\log_{10}$ scale with a mean value of 2 (indicated by the horizontal dashed lines).

Fourth example paper: Koh and Opitz 2025

Journal of the Royal Statistical Society, Series A (Statistics in Society), 2024, 1–17
Preprint article

Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data

Jonathan Koh[1,*] **and Thomas Opitz**[2]

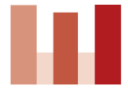
[1]Institute of Mathematical Statistics and Actuarial Science, Oeschger Centre for Climate Change Research, University of Bern, Switzerland and [2]Biostatistics and Spatial Processes (UR546), INRAE, Avignon, France

*Corresponding author. jonathan.koh@unibe.ch

[To be read before The Royal Statistical Society at the Discussion Meeting on the 'Analysis of citizen science data' held at the Society's 2024 annual conference in Brighton on Tuesday, 3 September 2024, the President, Dr Andrew Garrett, in the Chair]

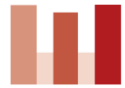
Abstract

Citizen science mobilises many observers and gathers huge datasets but often without strict sampling protocols, resulting in observation biases due to heterogeneous sampling effort, which can lead to biased statistical inferences. We develop a spatiotemporal Bayesian hierarchical model for bias-corrected estimation of arrival dates of the first migratory bird individuals at a breeding site. Higher sampling effort could be correlated with earlier observed dates. We implement data fusion of two citizen-science datasets with fundamentally different protocols (BBS, eBird) and map posterior distributions of the latent process, which contains four spatial components with Gaussian process priors: species niche; sampling effort; position and scale parameters of annual first arrival date. The data layer includes four response variables: counts of observed eBird locations (Poisson); presence-absence at observed eBird locations (Binomial); BBS occurrence counts (Poisson); first arrival dates (Generalised Extreme-Value). We devise a Markov Chain Monte Carlo scheme and check by simulation that the latent process components are identifiable. We apply our model to several migratory bird species in the northeastern United States for 2001–2021 and find that the sampling



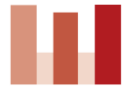
Fourth example paper: Koh and Opitz 2025

- Many observation data in ecology are opportunistic, i.e., they were collected incidentally or without a predefined research question in mind, so information on the criteria applied by the observer for sampling and reporting observations is limited.
- However, such data often allow for a spatiotemporal coverage of species monitoring that would not be attainable with protocol-based data collected by professionals at relatively sparse and deterministically chosen sampling locations.
- Citizen science data is particularly valuable for statistical inference on low-probability events (e.g., occurrences of rare species, or unusual or extreme phenological events) and on occurrence times of events, such as the arrival of migratory birds at their breeding site in spring.



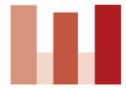
Fourth example paper: Koh and Opitz 2025

- The general data quality from citizen-science initiatives is high (Kosmala et al., 2016), though the need for statistical methods to account for different data biases prevails.
- Isaac et al. (2014) discuss bias-correction approaches for ecological trend estimates, and conclude that opportunistic data would be further enhanced if information on the sampling effort at the data collection points can be captured.
- They further identify four key aspects of biases induced by volunteer sampling, due to uneven sampling in time, space, effort per visit, and uneven detectability of species individuals, which can vary by observer, species and land cover.



Fourth example paper: Koh and Opitz 2025

- Another problem arises when only the presence of species is reported but not their absence (presence-only data), so areas and times without any reported occurrences could correspond either to the true absence of the species or to the presence of the species but the absence of any observer.
- Authors adopt the common approach of considering species not reported within a checklist as being absent from the spacetime location of the checklist.
- By systematically adding such (pseudo-)absence information to the dataset, they obtain a binary observation for each combination of checklist and species and then estimate the species presence probability using available predictors with a binomial likelihood.



Fourth example paper: Koh and Opitz 2025

- Previous studies often used rather complex approaches to define and estimate a date that can be viewed as representative of the arrival of birds at their breeding site.
- Authors focus only on the first arrival for each pixel of a spatial mesh covering the study area, i.e., they model the minimum of all dates of occurrence of the species during the year.
- They use the Generalised Extreme-Value (GEV) distribution motivated by extreme value theory to model this sample extreme of all observed occurrence times in the year, similar to Wijeyakulasuriya et al. (2023).
- They consider dates as a continuous variable (since the distribution of observed first arrivals spans over a long enough period, discretisation effects have a weak influence).
- Their modelling approach is the first that aims to appropriately capture the interplay of the niche, the sampling effort and the observed first arrivals to provide bias-corrected predictions of the true first arrivals.

Fourth example paper: Koh and Opitz 2025

- **Generalized Extreme Value (GEV) Distribution**:

$$\Pr(M_n \leq z) \approx \begin{cases} \exp\left\{-\exp\left(-\frac{z-\mu}{\sigma}\right)\right\}, & \xi = 0 \\ \exp\left\{-\left[1+\xi\left(-\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, & \xi \neq 0 \end{cases}$$

- Hierarchical system of the four regression equations:

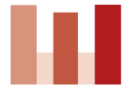
$$N_j^{BBS} | \lambda^{BBS}, \theta_{bbs} \sim Pois\left\{\sum_{k \in route_j} \omega_k \, \lambda^{BBS}(s_k; \theta_{bbs})\right\}$$

$$N_i^{ckl} | \lambda^{ckl}, \theta_{ckl} \sim Pois\{\lambda^{ckl}(s_i, t_i; \theta_{ckl})\}$$

$$N_i^{spc} | N_i^{ckl}, p^{spc}, \theta_{spc} \sim Bin\{N_i^{ckl}, p^{spc}(s_i, t_i; \theta_{spc})\}$$

$$Z_i | \mu, \theta_\mu, \sigma, \theta_\sigma \sim GEV\{\mu(s_i, t_i; \theta_\mu), \sigma(s_i; \theta_\sigma), \xi\}$$

- where $\theta_{bbs}, \theta_{ckl}, \theta_\mu, \theta_\sigma \sim Hyperpriors$



Fourth example paper: Koh and Opitz 2025

- A **Gaussian process** is specified by a mean function $\mu: \mathcal{X} \to \mathbb{R}$, such that $\mu(x)$ is the mean of $f(x)$ and a covariance/kernel function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x_i, x_j)$ is the covariance between $f(x_i)$ and $f(x_j)$. We say $f \sim GP(\mu, k)$ if for any $x_1, x_2, \ldots, x_n \in \mathcal{X}$, $[f(x_1), f(x_2), \ldots, f(x_n)]^T$ is Gaussian distributed with mean $[\mu(x_1), \mu(x_2), \ldots, \mu(x_n)]^T$ and $n \times n$ covariance/kernel matrix K_{xx}:

$$K_{xx} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

- Four **Gaussian spatial process priors** included in our model:

$$X^{pref}(\cdot) \sim GP(\omega_1), X^{niche}(\cdot) \sim GP(\omega_2),$$
$$X^{GEV-\mu}(\cdot) \sim GP(\omega_3), X^{GEV-\sigma}(\cdot) \sim GP(\omega_4)$$

- where $\omega_1, \omega_2, \omega_3 \; and \; \omega_4$ contain individual range and standard deviation parameters that control the spatial dependence of each process. These parameters are assigned joint Penalized Complexity Priors (Fuglstad et al., 2018).

Fourth example paper: Koh and Opitz 2025

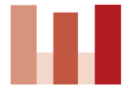
- For a chosen permutation m of our observations, the exact joint density of observations can be written as the product of conditional densities, i.e.:

$$f(x_1, \dots, x_D) = f(x_{m(1)}) \prod_{i=2}^{D} f\left(x_{m(1)} \middle| x_{H(i;m)}\right)$$

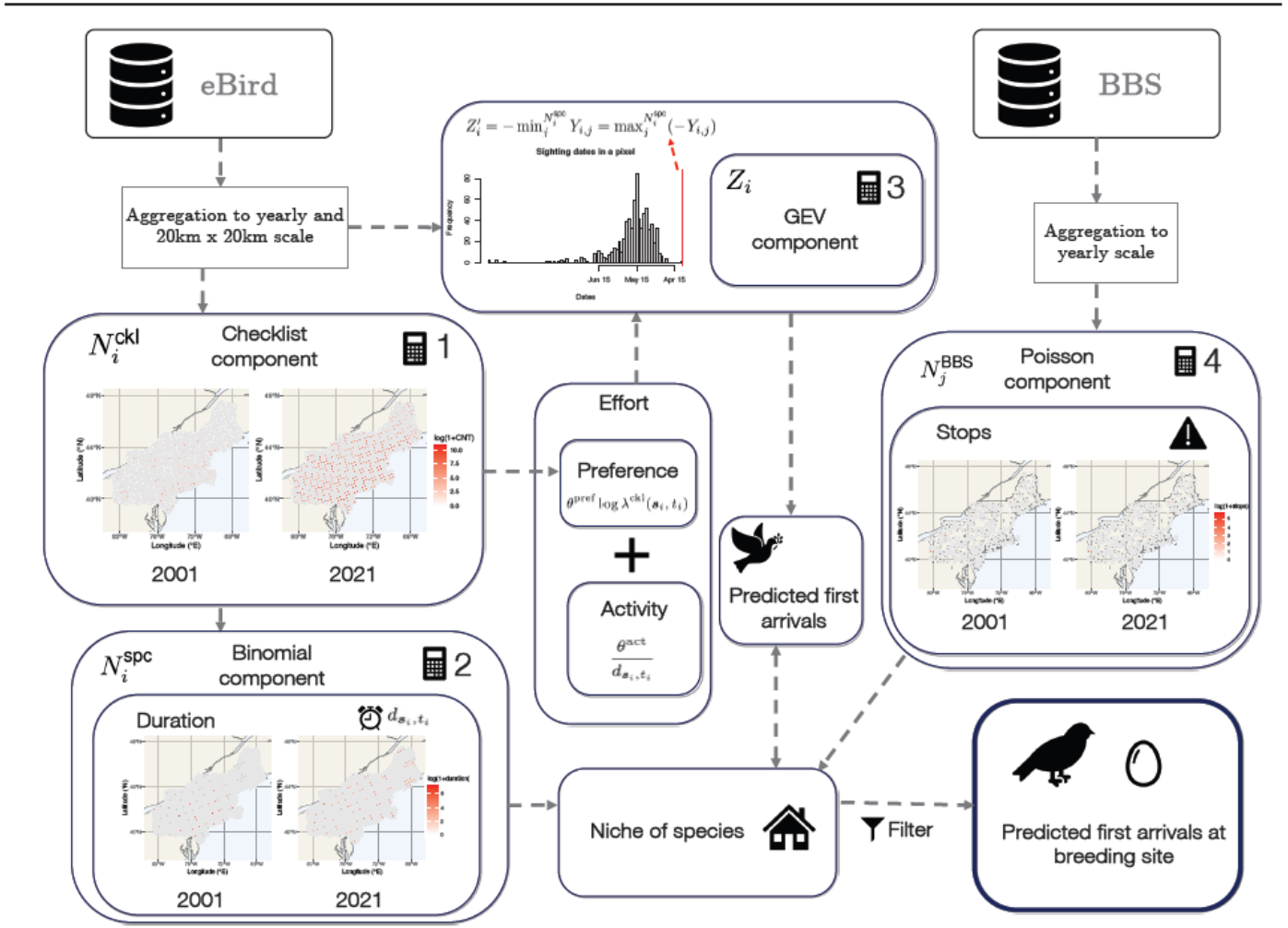
- The **Vecchia approximation** still represents a valid Gaussian process, but assumes that:

$$f(x_1, \dots, x_D) \approx \hat{f}(x_1, \dots, x_D) = f(x_{m(1)}) \prod_{i=2}^{D} f\left(x_{m(1)} \middle| x_{S(i;m)}\right)$$

- where $S(i;m) \subseteq H(i;m)$ and $|S(i;m)| = k$.



Fourth example paper: Koh and Opitz 2025



Fourth example paper: Koh and Opitz 2025

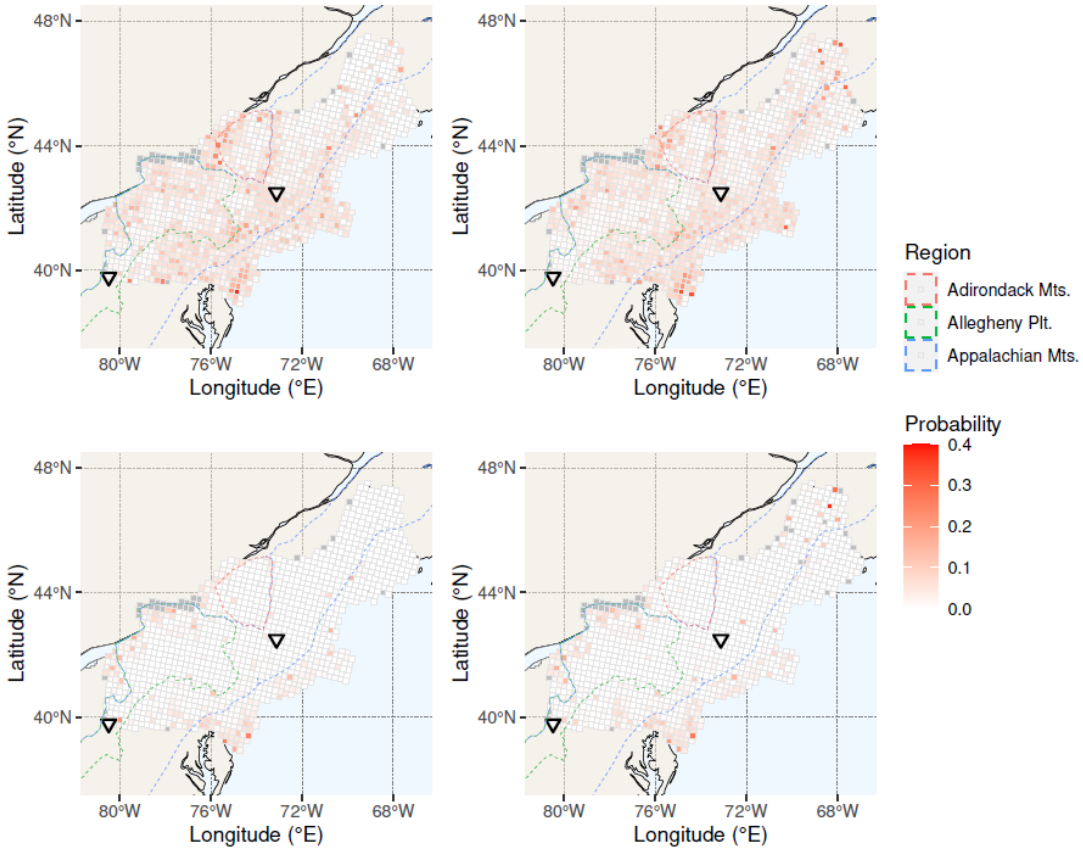
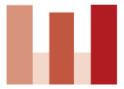


Fig. 8: Empirical mean probability (left displays) of observing the species *Great Crested Flycatcher* (top) and *Purple Martin* (bottom) in checklists aggregated over all years; posterior mean probability (right displays) of observing the species from the presence model by using the mean checklist duration aggregated over all years as covariate. The two triangles are the two pixels used in the predictions of first arrivals detailed in Section 5.3.



Fourth example paper: Koh and Opitz 2025

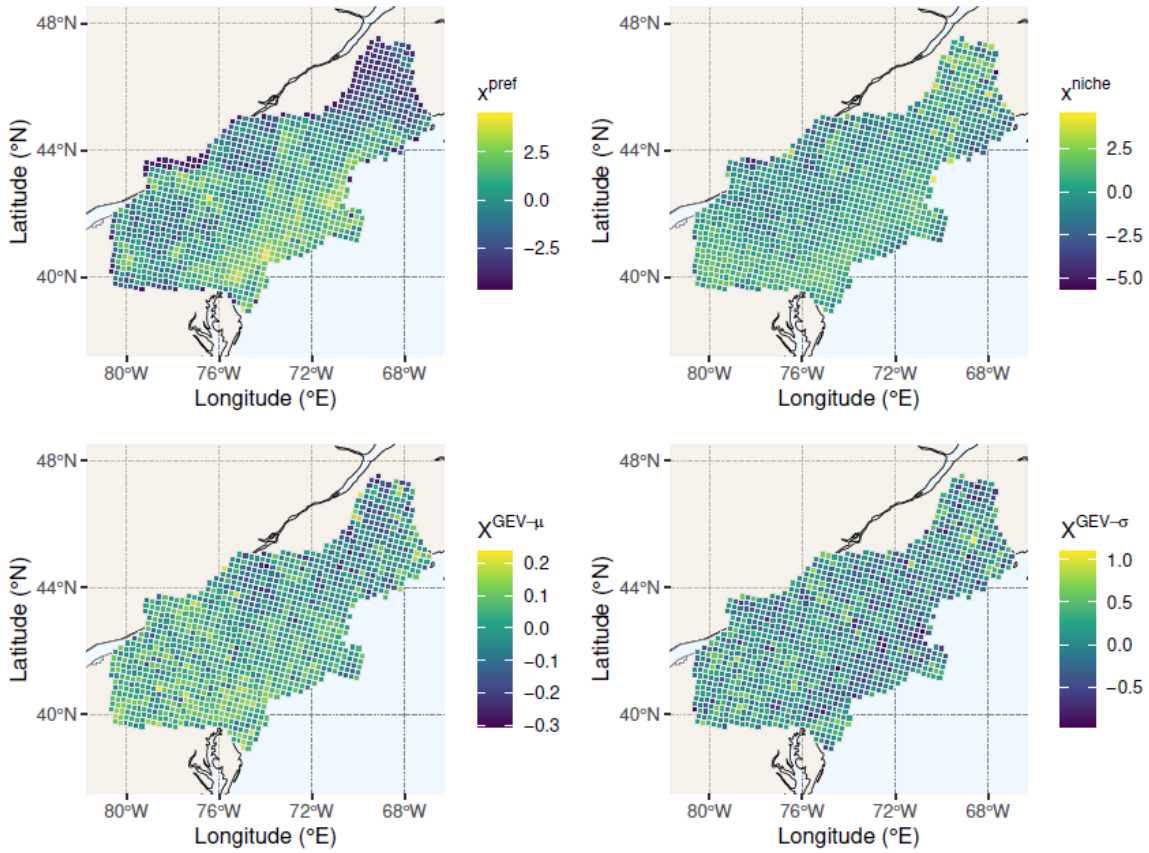


Fig. 13: Posterior means plots of the four spatial random effects in (2) for the model fitted to the species *Chimney Swift*.

Discussion, reflections and conclusion

- Citizen science data is a novel dataset type
- While being connected to „liberative“ and „empowerment“ societal aspects, such data are prone to statistical issues and problems, and *provide biased (possibly heavily biased, i.e. wrong) results* if used uncarefully
- Methods to address this are urgent to develop
- Several biases appear as predominant in the study of citizen science data: sampling biases of various kind, spatial bias, computational cost, observer biases
- All presented articles used Bayesian approaches, which provide a need to even additionally consider computational efficiency aspects
- Combinations with machine learning approaches (such as conformal prediction, as used in Betsy's paper) would be great to study&consider
- K. Mengersen noted an even more „wild CS studies“: this line of statistics research promises interesting strands and extensiosn for the future!

Discussion, reflections and conclusion

Thank you for listening!

Q&A

andrej.srakar@ef.uni-lj.si

ebersson@mit.edu