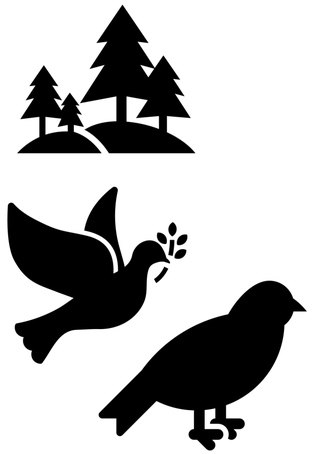


EXTREME-VALUE MODELLING OF MIGRATORY BIRD ARRIVAL DATES: INSIGHTS FROM CITIZEN SCIENCE DATA

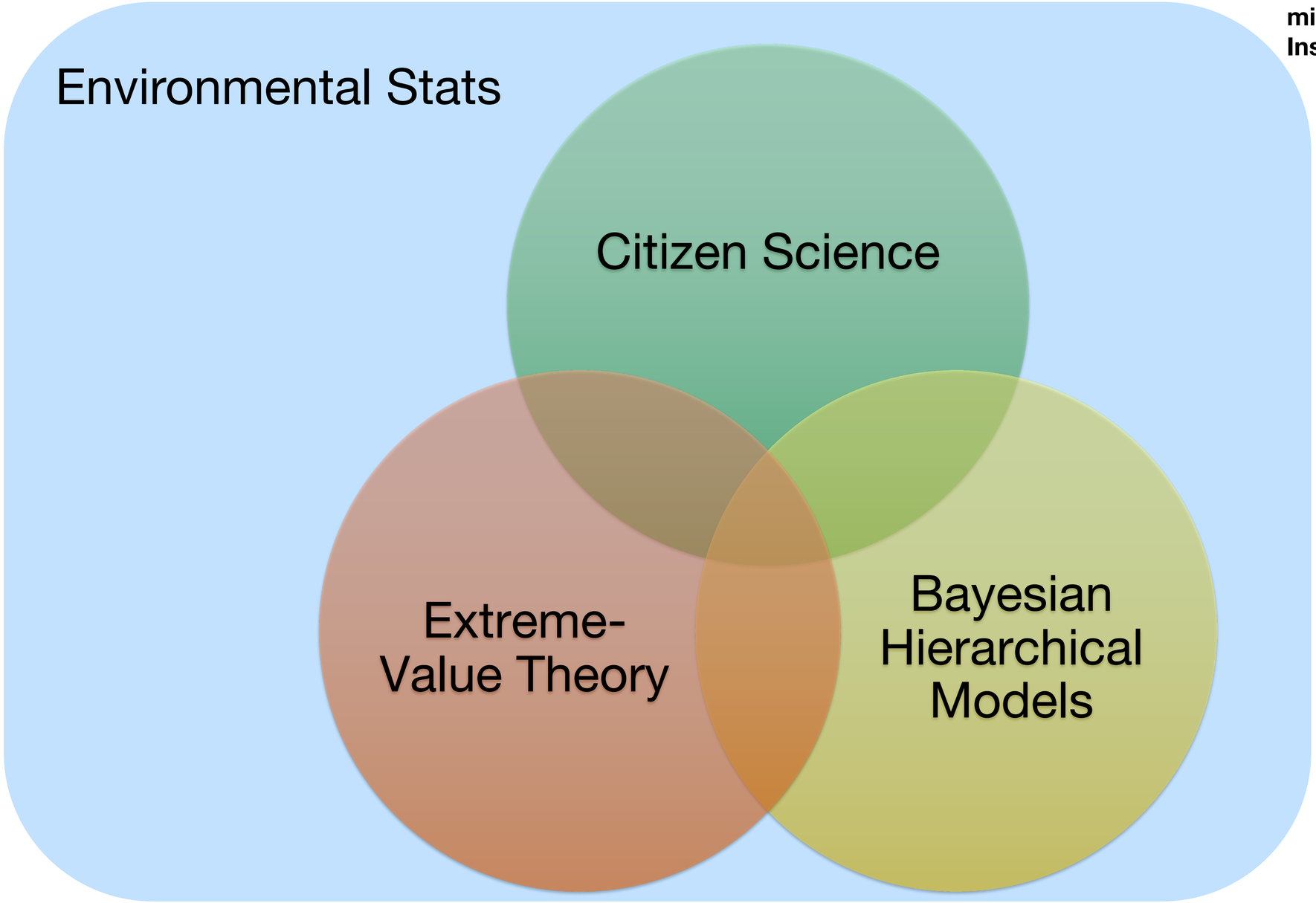
Jonathan Koh
Joint work with Thomas Opitz

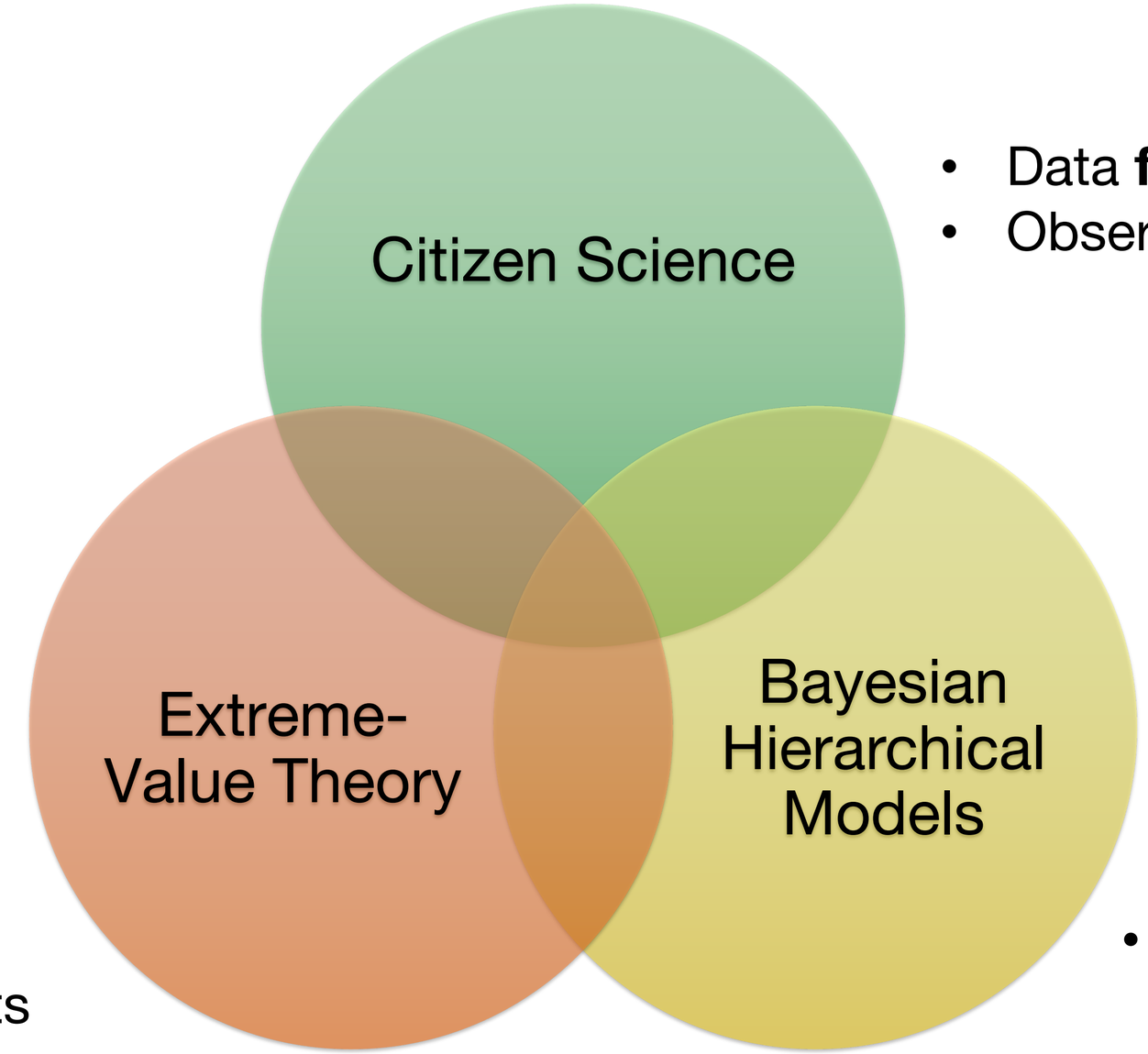
ETH zürich

CMStat 2025
London, 13/12/2025

Environmental Stats

(includes Ecological Stats)


Environmental Stats
Citizen Science
Extreme-Value Theory
Bayesian Hierarchical Models



- Data **fusion**
- Observational **bias**

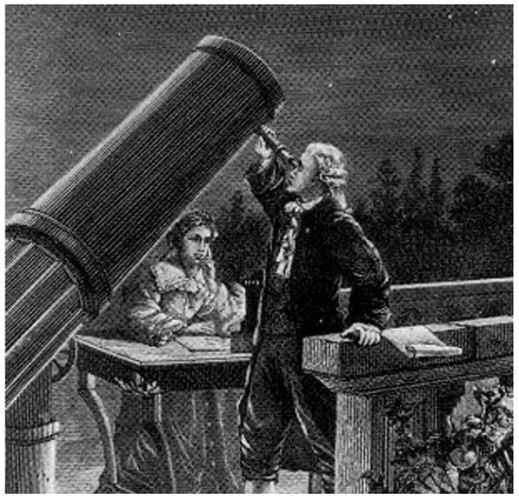
- Extremes of **phenological** events

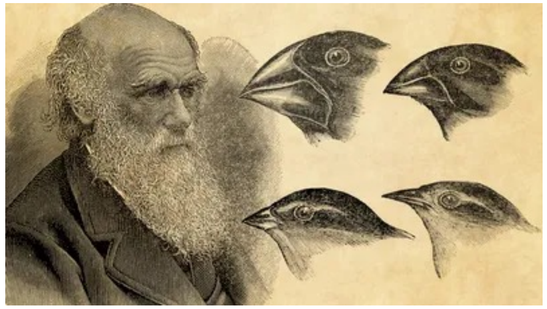
- Inference on important **latent processes**

What is citizen science?

Citizen Science

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new…





What is citizen science?

Citizen Science

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new…

What is citizen science?

Citizen Science

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new…
3. Form has changed, and significant **traction** over the last two decades

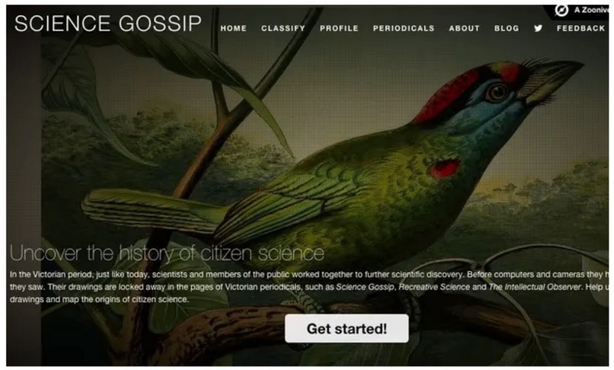
The H word
Science

People power: how citizen science could change historical research

Crowdsourcing research by 'non-specialists' could help historians investigate big-data archives, and in the process make everyone an expert

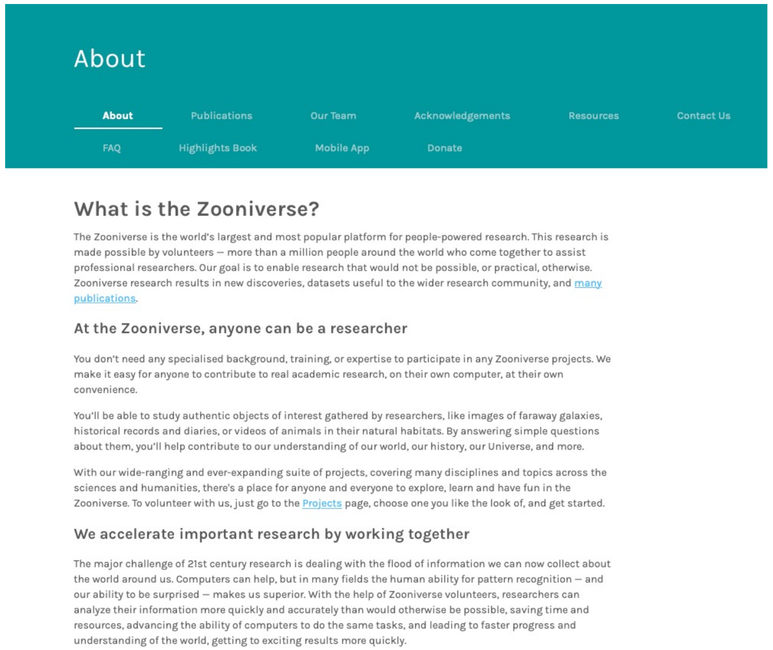
Geoffrey Belknap
Tue 26 Apr 2016 08.45 CEST

6



The front page of the Science Gossip website, a Zooniverse Citizen Humanities project.

Citizen science is a digital method, which has been applied to a range of big-data scientific problems. The Zooniverse is a key player in this; having first sought the help of the crowd in classifying galaxies almost a decade ago, it now boasts 47 different projects with well over a million users. The projects hosted on their site have been bringing to the forefront concerns over who



What is citizen science?

Citizen Science

1. Involves a collaborative approach to scientific inquiry that engages **volunteers** and **non-professionals** in the collection, analysis, and interpretation of data.
2. Not new…
3. Form has changed, and significant **traction** over the last two decades
4. **Government** agencies are doing it

How does eBird work?

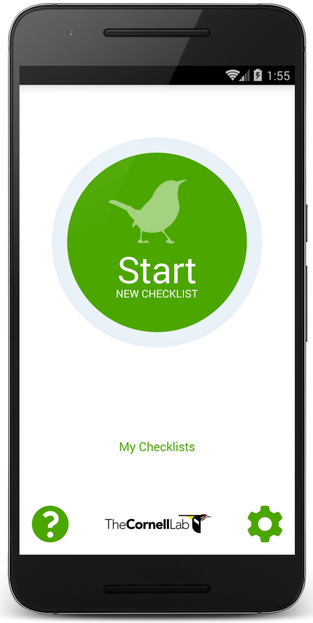
Citizen Science

1. Simone is a **birder**
2. While hiking, he opens the eBird app and starts a `**checklist**'. The app notes the **date** and **time** he starts birding, **where** he has travelled during the checklist and **how long** he has been birding

How does eBird work?

Citizen Science

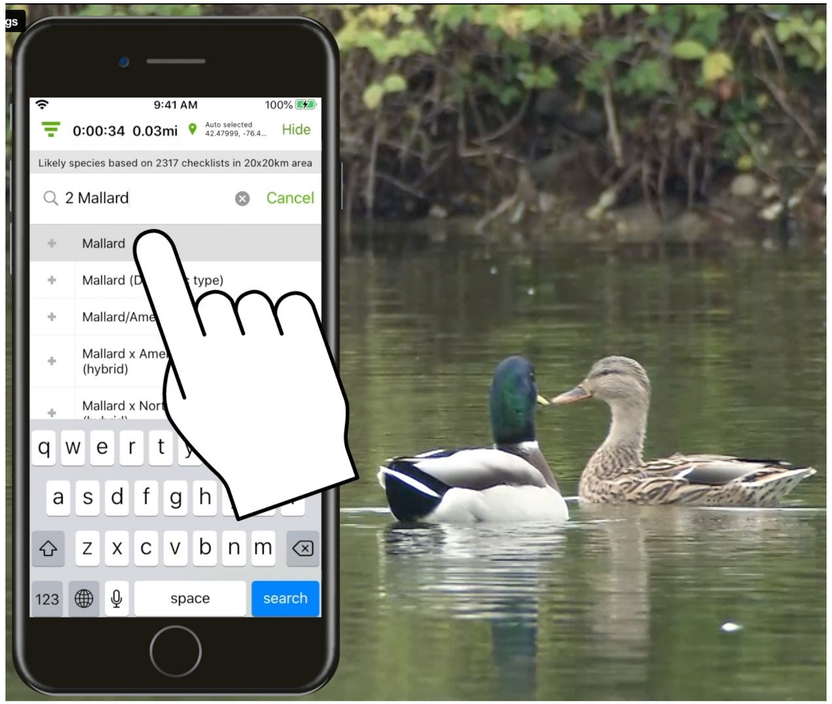
1. Simone is a **birder**
2. While hiking, he opens the eBird app and starts a `**checklist**'. The app notes the **date** and **time** he starts birding, **where** he has travelled during the checklist and **how long** he has been birding



How does eBird work?

Citizen Science

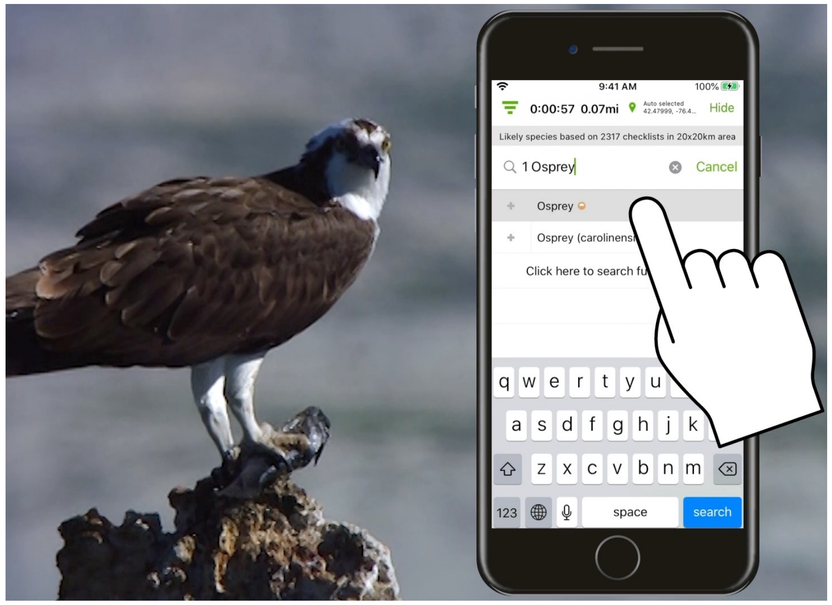
3. He spots a Mallard, and can easily **record** it in the app (based on a pack of recommended bird species)

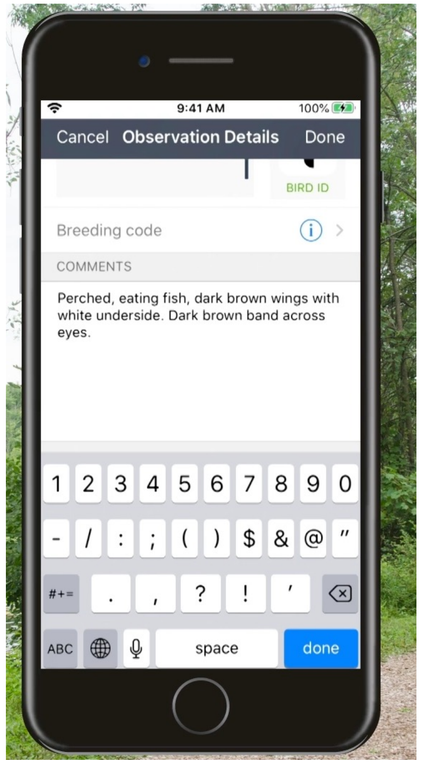


How does eBird work?

Citizen Science

4. He spots an Osprey, and records it in the app. He can also supply more information (including media files)

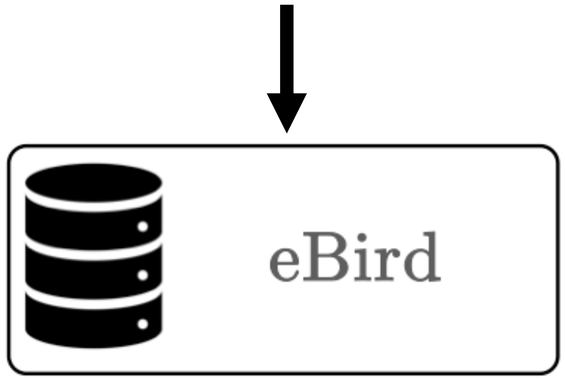




How does eBird work?

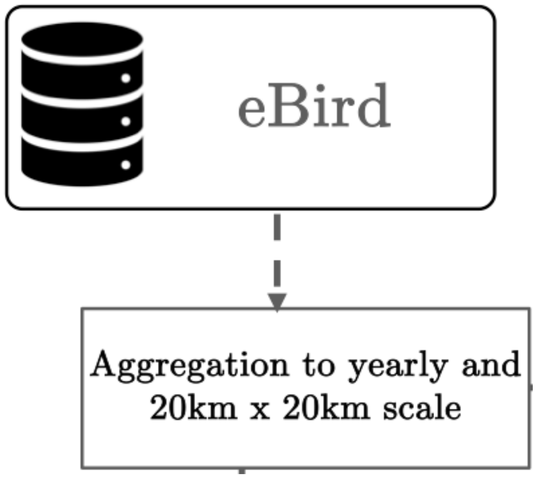
Citizen Science

5. He finishes his checklist and submits his data to eBird
6. eBird internally verifies it, and it goes into their database

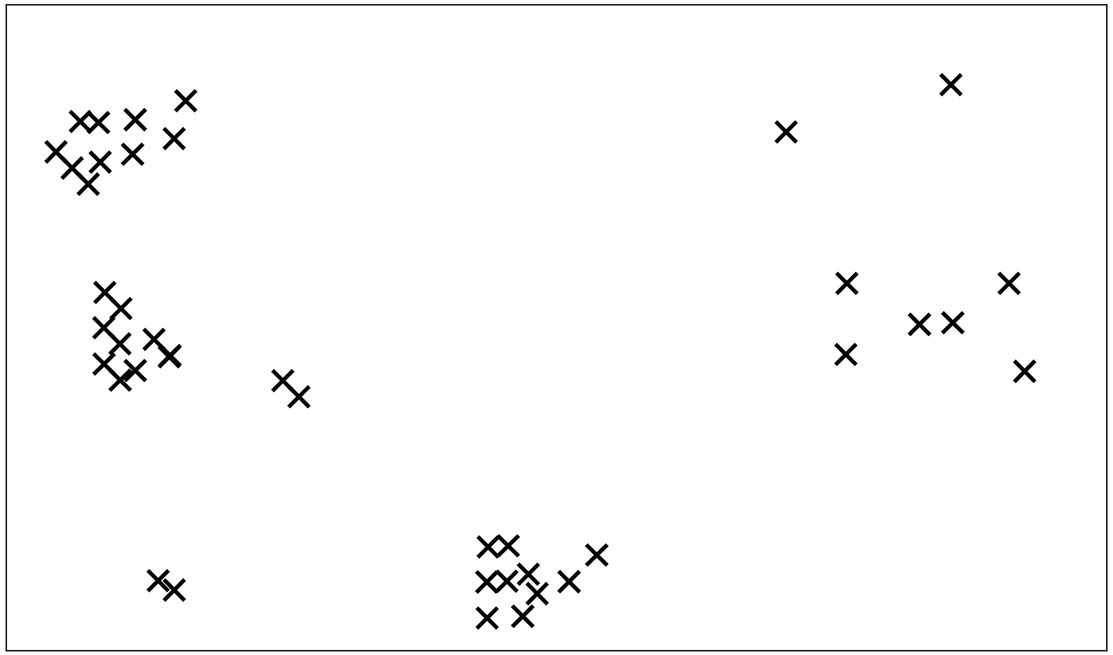


eBird data processing

Citizen Science

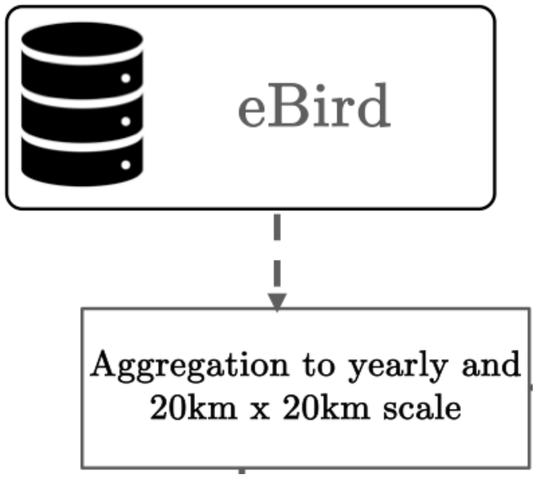


Year 2020

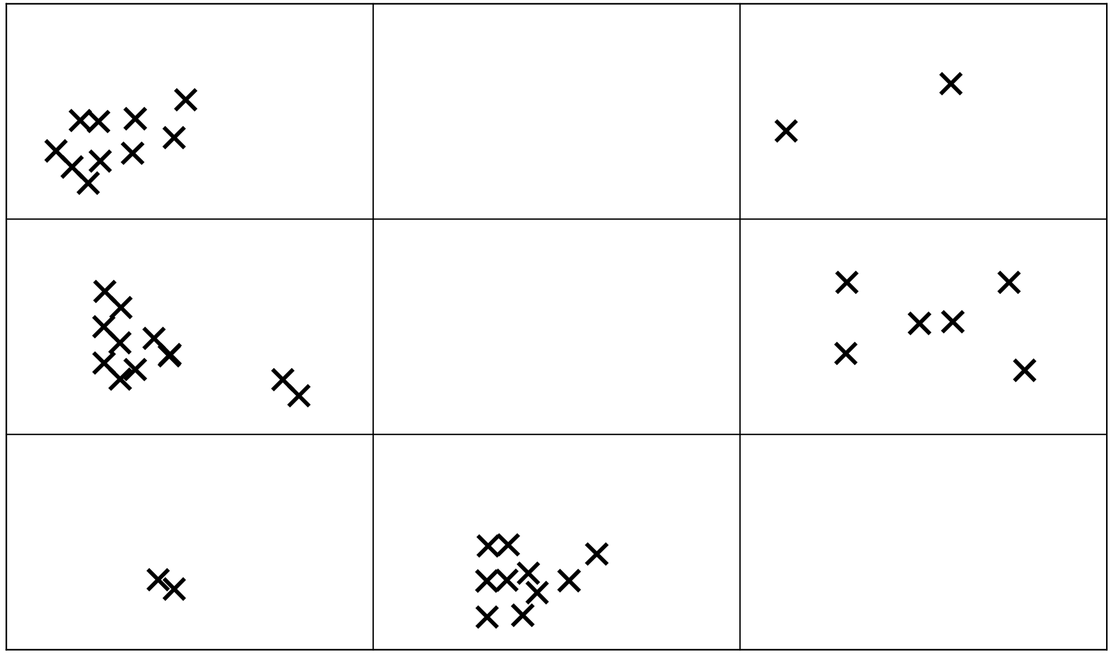


eBird data processing

Citizen Science

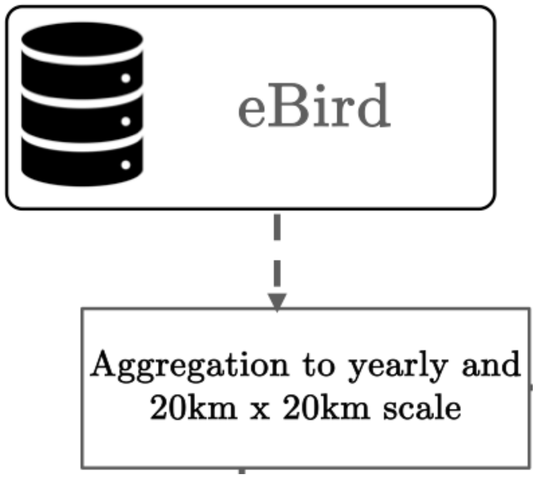


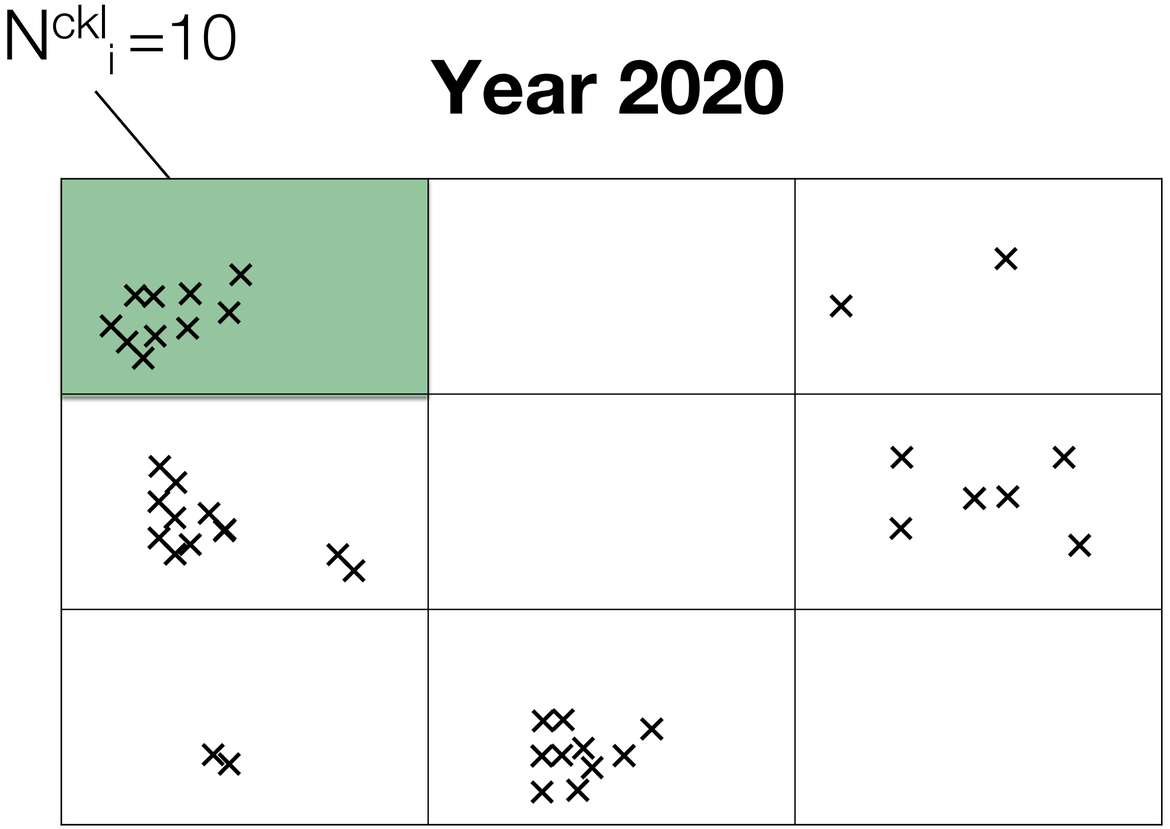
Year 2020



eBird data processing

Citizen Science

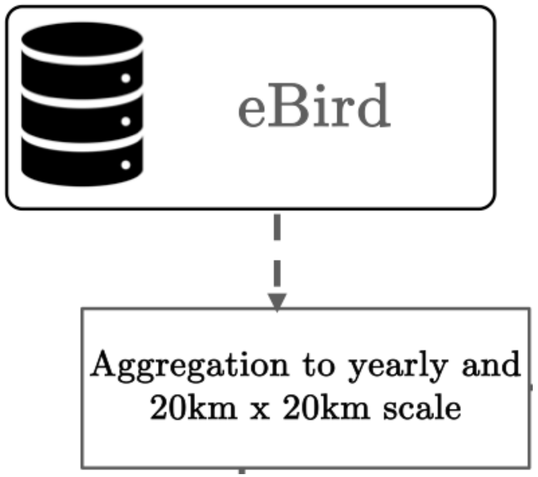




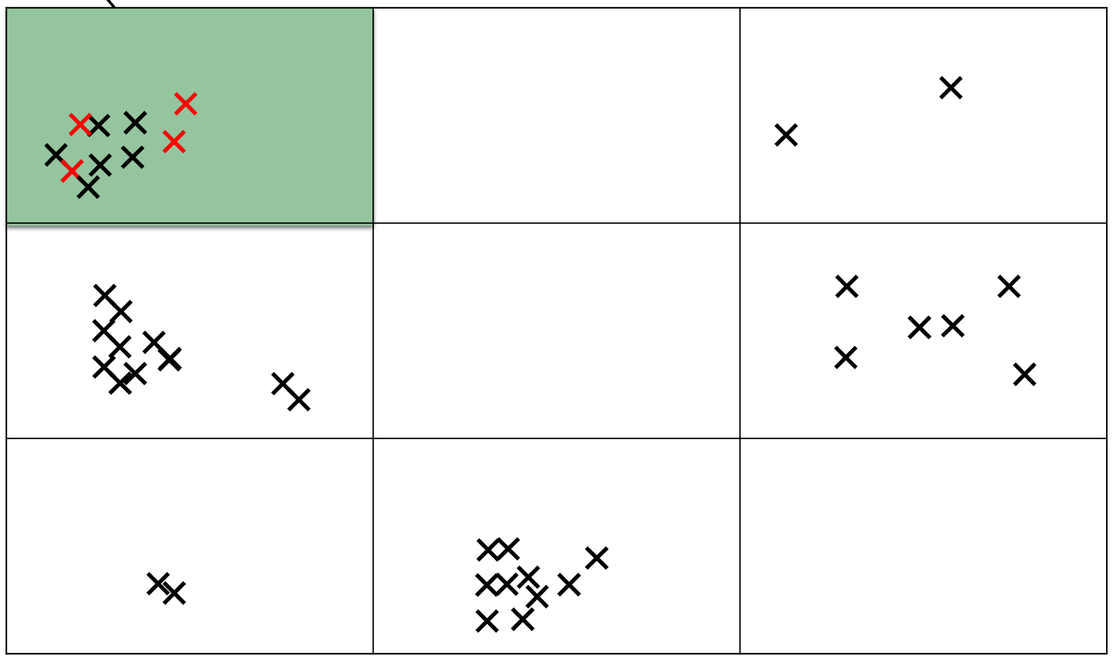
eBird data processing

Citizen Science

Purple Martin



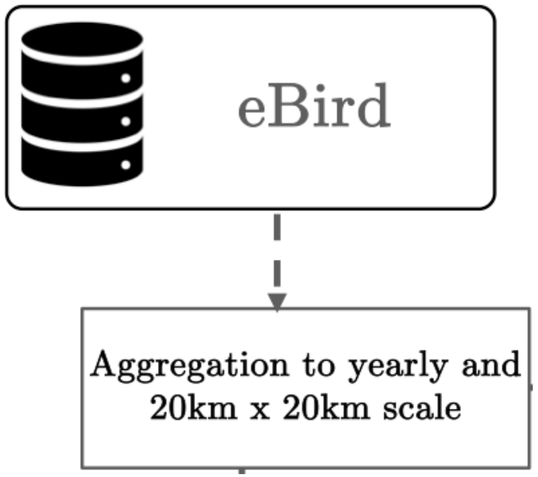
$N^{spc}_i = 4$

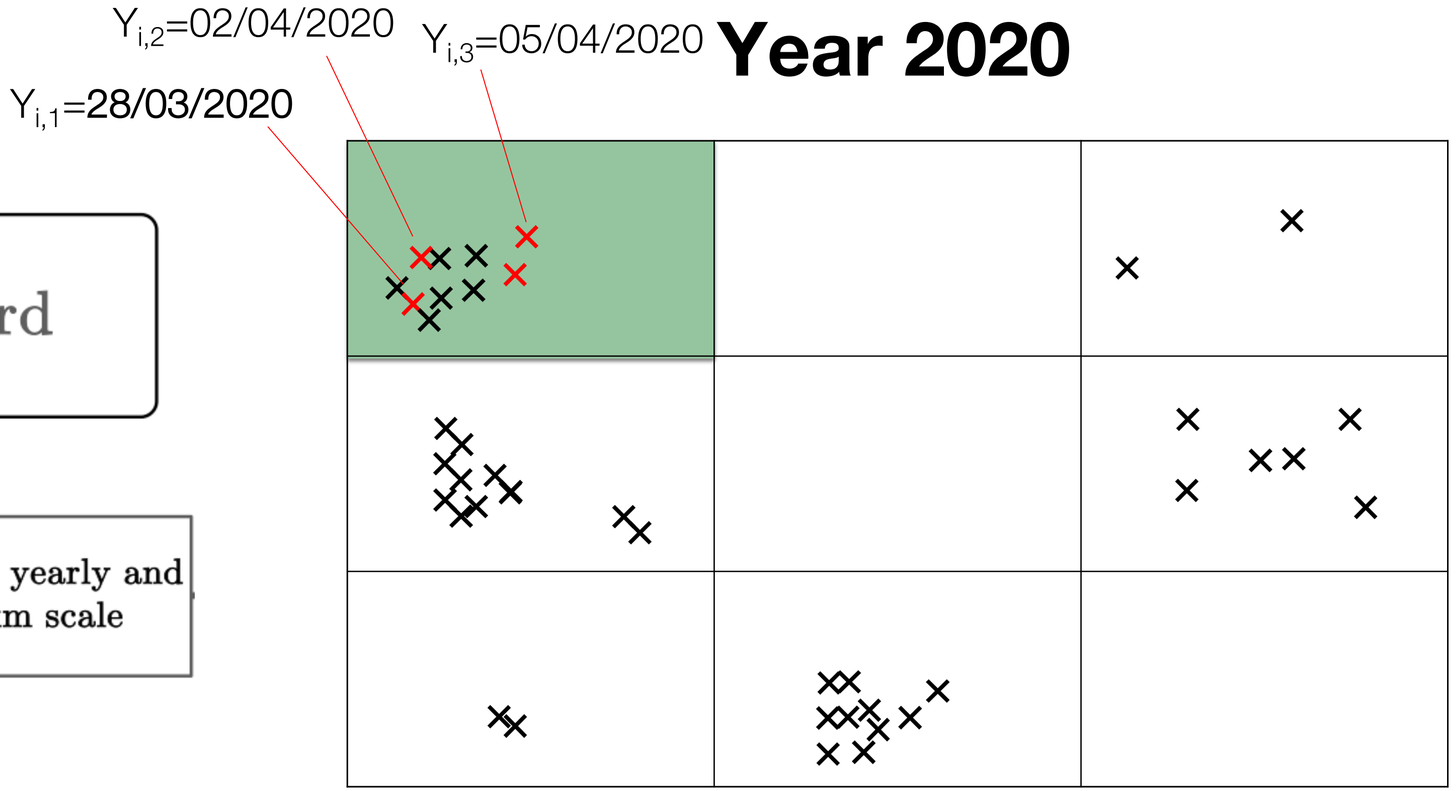


eBird data processing

Citizen Science

Purple Martin

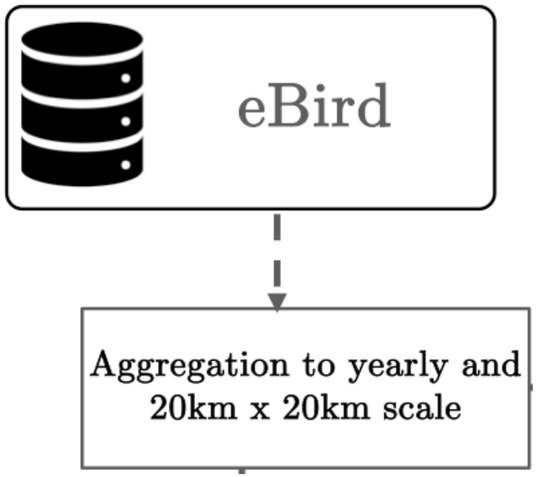




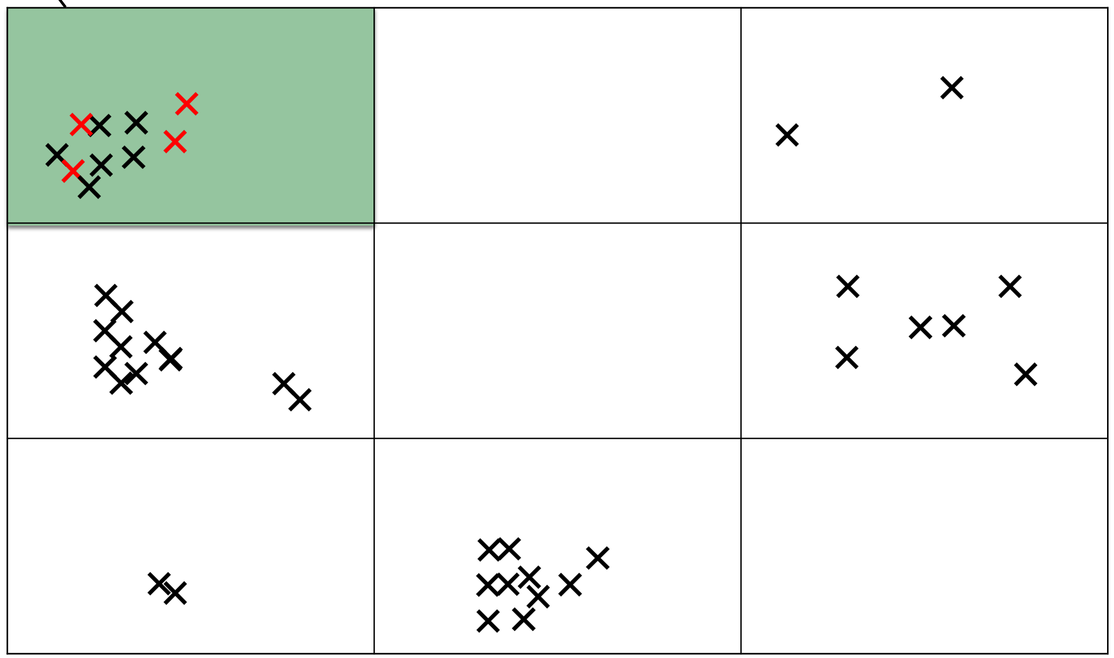
eBird data processing

Extreme-Value Theory

Purple Martin



Z'_i = First arrival date: 28/03/20

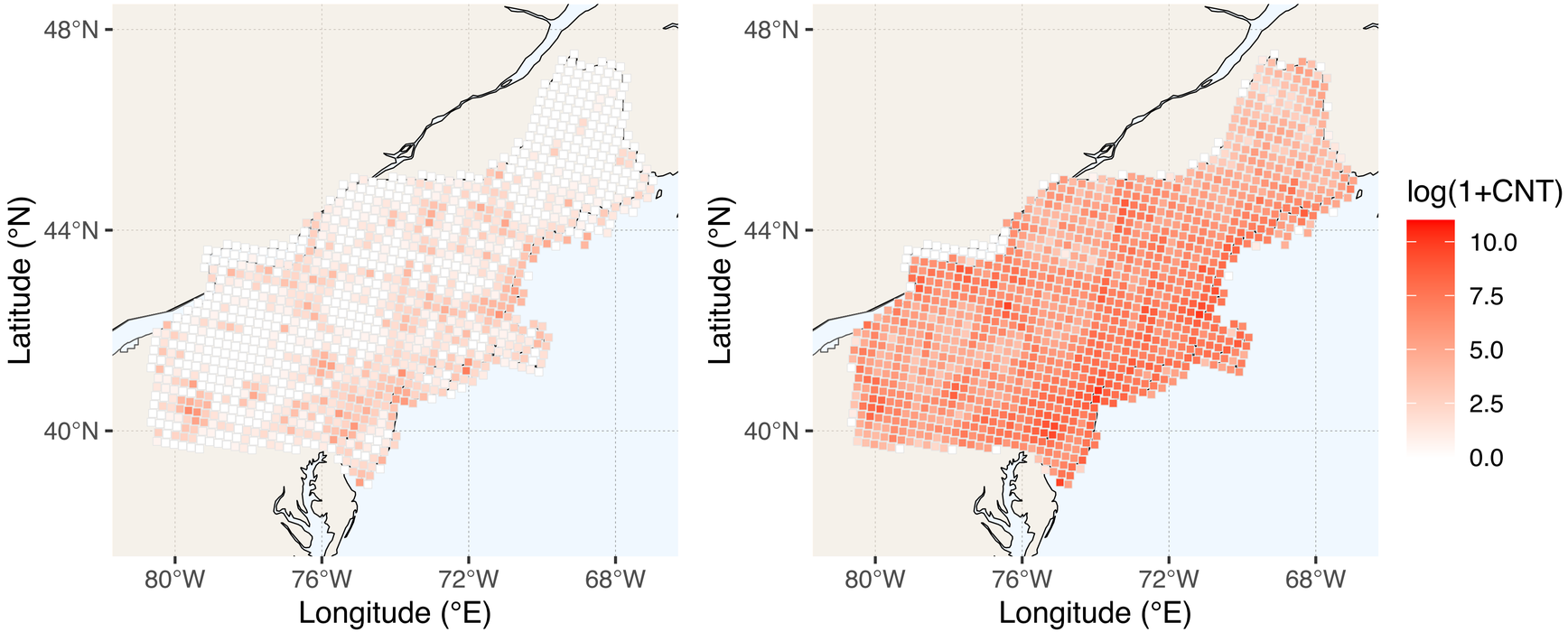


*Wijeyakulasuriya et al. (2024)

eBird

Citizen Science

Strong temporal trends in reported occurrences



2001

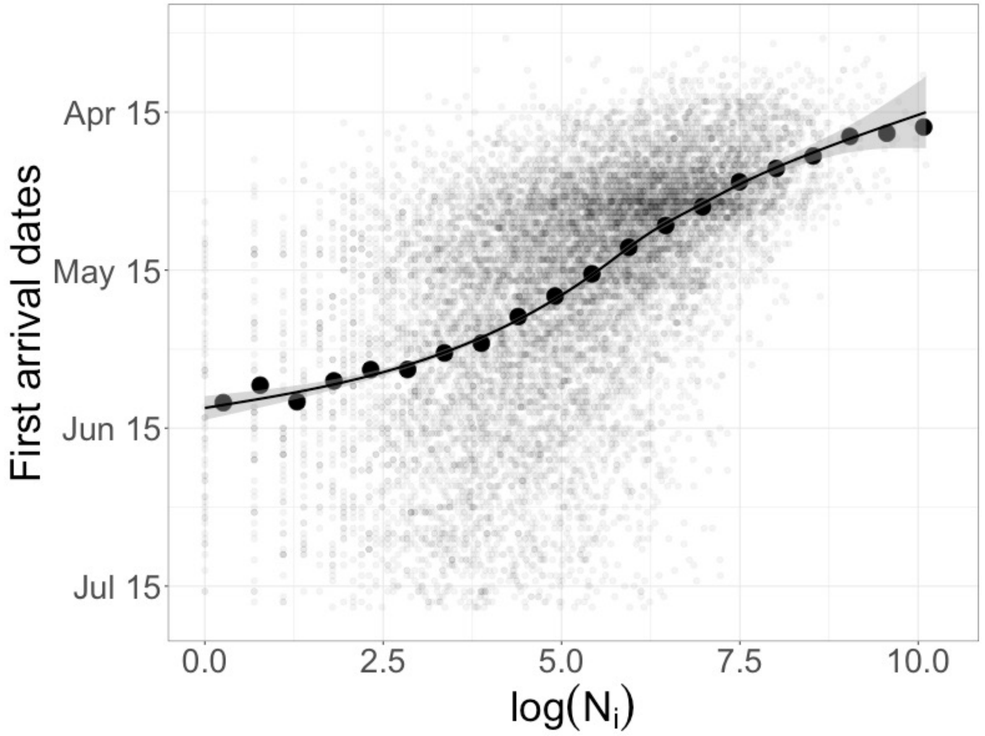
2021

First arrival dates vs. checklist counts

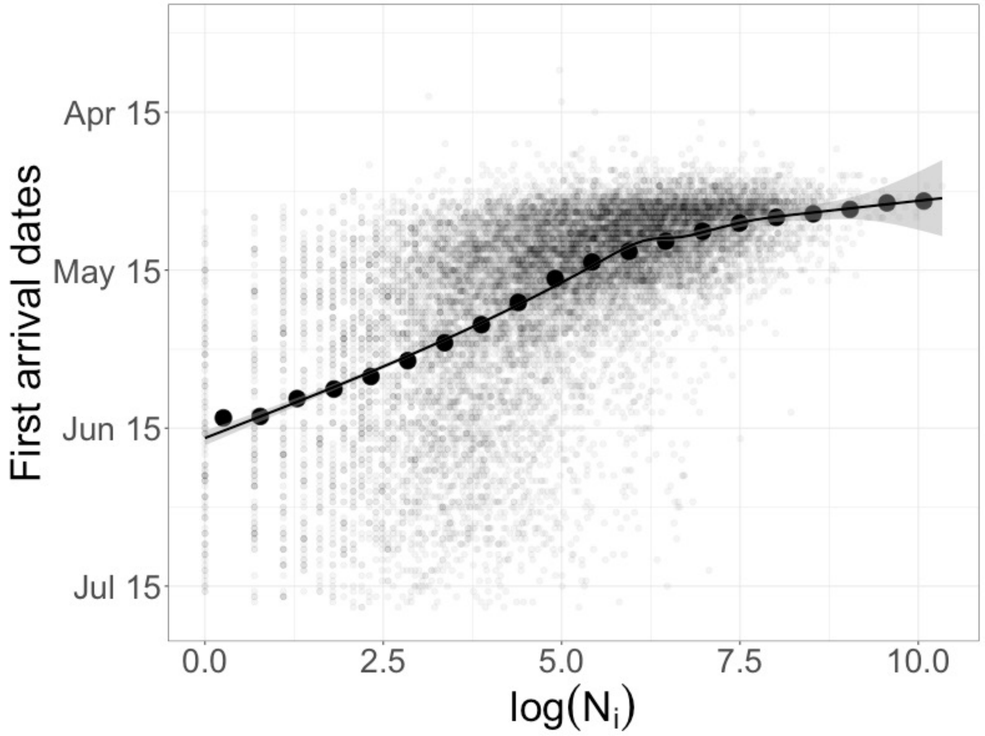
Citizen Science

Bayesian Hierarchical Models

Chimney-Swift



Chestnut-sided Warbler

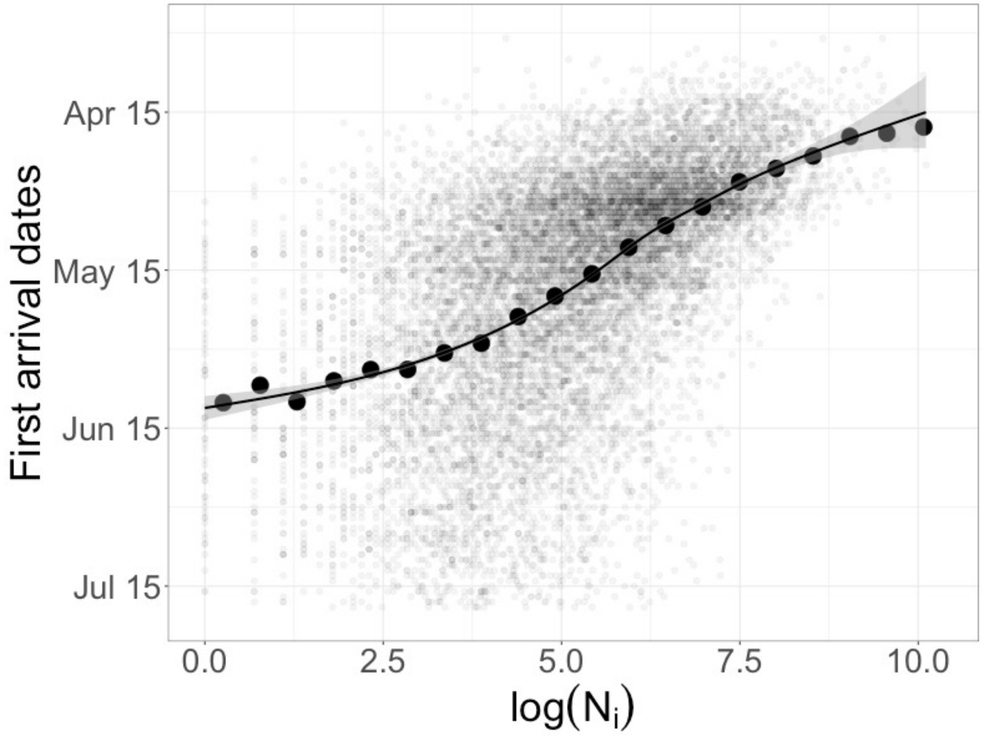


“Observational effort”/ Preference

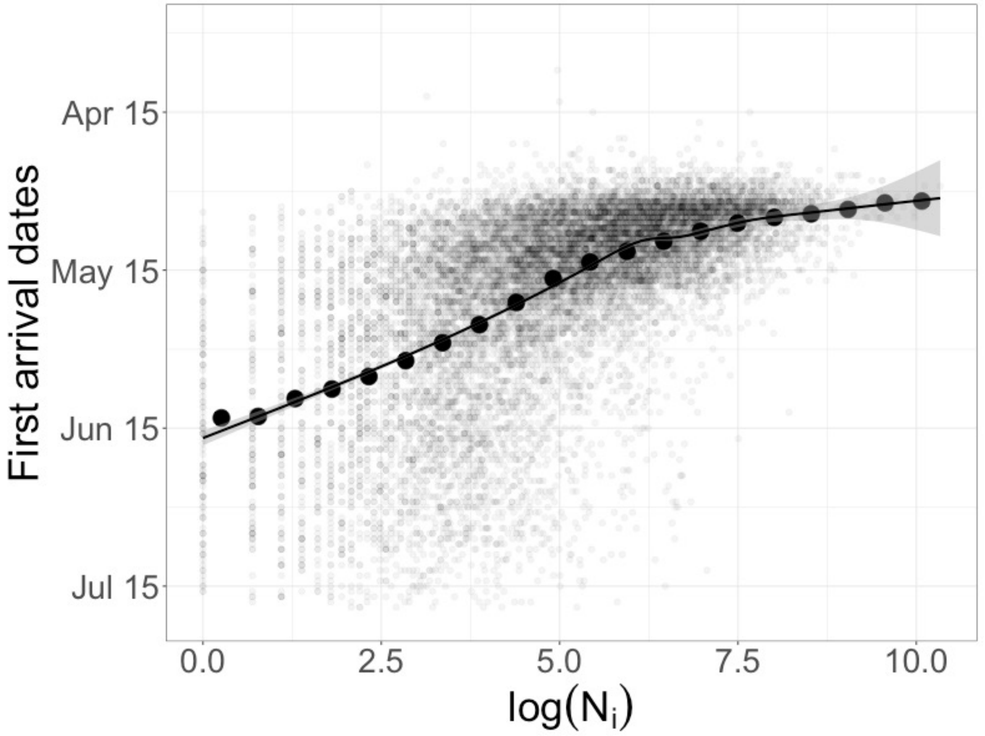
Citizen Science

Bayesian Hierarchical Models

Chimney-Swift



Chestnut-sided Warbler



Citizen Science

Observational effort = Preference + Activity

space-time varying

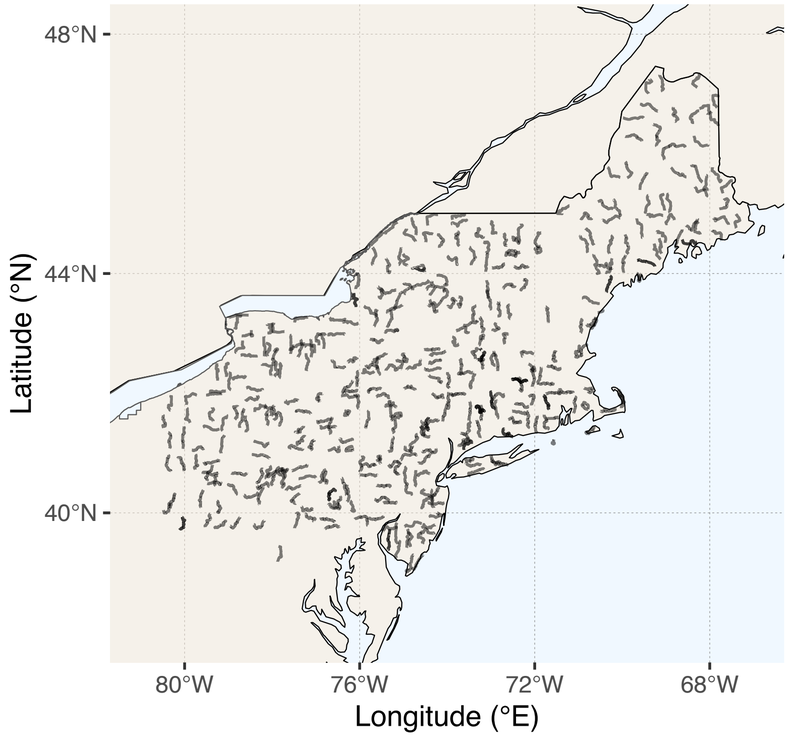
captured by the sampling intensity for the checklists

captured by the (median) time spent on the checklist

Breeding Bird Survey (BBS) sampling routes

Citizen Science

- For each route (~40km),
  bird occurrences are reported at 50 **equidistant** stops
- Only every **summer**
- Complex data preprocessing (missing observations, **missing stop** coordinates, etc.)



MODEL

Modelling goals

Bayesian Hierarchical Models

- Fit a realistic **model** to first arrival data, conditional on covariates
- Correct for the **observational bias** from these datasets

Modelling goals

Bayesian Hierarchical Models

- Fit a realistic **model** to first arrival data, conditional on covariates
- Correct for the **observational bias** from these datasets
- Use the model to make **posterior predictions**
- **Interpolate** spatially to locations not visited, in a **reasonable** way

A multi-response spatial regression system

Bayesian Hierarchical Models

Multi-response spatial regression

$$N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} \sim \mathrm{Pois}\left\{ \sum_{k \in \mathrm{route}_j} \omega_k \lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}}) \right\},$$

$$N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} \sim \mathrm{Pois}\left\{ \lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}}) \right\},$$

$$N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} \sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})\},$$

$$Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\},$$

where

$$\boldsymbol{\theta}_{\mathrm{bbs}}, \boldsymbol{\theta}_{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \mathrm{Hyperpriors}$$

A multi-response spatial regression system

Bayesian Hierarchical Models

Multi-response spatial regression

$$N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} \sim \mathrm{Pois}\left\{ \sum_{k \in \mathrm{route}_j} \omega_k \lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}}) \right\},$$

$$N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} \sim \mathrm{Pois}\left\{ \lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}}) \right\},$$

$$N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} \sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})\},$$

$$Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\},$$

where

$$\boldsymbol{\theta}_{\mathrm{bbs}}, \boldsymbol{\theta}_{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \mathrm{Hyperpriors}$$

Sharing random effects

Bayesian Hierarchical Models

Multi-response spatial regression

$$N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} \sim \mathrm{Pois}\left\{ \sum_{k \in \mathrm{route}_j} \omega_k \boxed{\lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}})} \right\},$$

$$N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} \sim \mathrm{Pois}\left\{\lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}})\right\},$$

$$N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} \sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, \boxed{p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})}\},$$

$$Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \mathrm{GEV}\{\boxed{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu)}, \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\},$$

$$X^{\mathrm{niche}}(\cdot) \sim \mathcal{GP}(\boldsymbol{\omega}_2),$$

where

$$\boldsymbol{\theta}_{\mathrm{bbs}}, \boldsymbol{\theta}_{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \mathrm{Hyperpriors}$$

Sharing random effects

Bayesian Hierarchical Models

Multi-response spatial regression

$$N_j^{\mathrm{BBS}} \mid \lambda^{\mathrm{BBS}}, \boldsymbol{\theta}_{\mathrm{bbs}} \sim \mathrm{Pois}\left\{ \sum_{k \in \mathrm{route}_j} \omega_k \lambda^{\mathrm{BBS}}(\boldsymbol{s}_k; \boldsymbol{\theta}_{\mathrm{bbs}}) \right\},$$

$$N_i^{\mathrm{ckl}} \mid \lambda^{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{ckl}} \sim \mathrm{Pois}\{\lambda^{\mathrm{ckl}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{ckl}})\},$$

$$N_i^{\mathrm{spc}} \mid N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}, \boldsymbol{\theta}_{\mathrm{spc}} \sim \mathrm{Bin}\{N_i^{\mathrm{ckl}}, p^{\mathrm{spc}}(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mathrm{spc}})\},$$

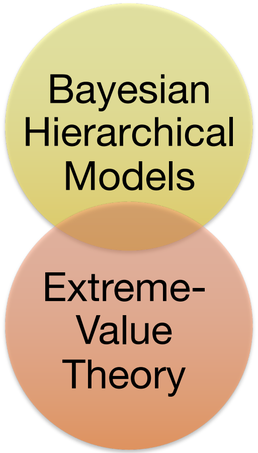
$$Z_i \mid \mu, \boldsymbol{\theta}_\mu, \sigma, \boldsymbol{\theta}_\sigma \sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_\mu), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_\sigma), \xi\},$$

$$X^{\mathrm{pref}}(\cdot) \sim \mathcal{GP}(\boldsymbol{\omega}_1)$$

where

$$\boldsymbol{\theta}_{\mathrm{bbs}}, \boldsymbol{\theta}_{\mathrm{ckl}}, \boldsymbol{\theta}_{\mathrm{spc}}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\sigma \sim \mathrm{Hyperpriors}$$

Saturating effect of observational effort

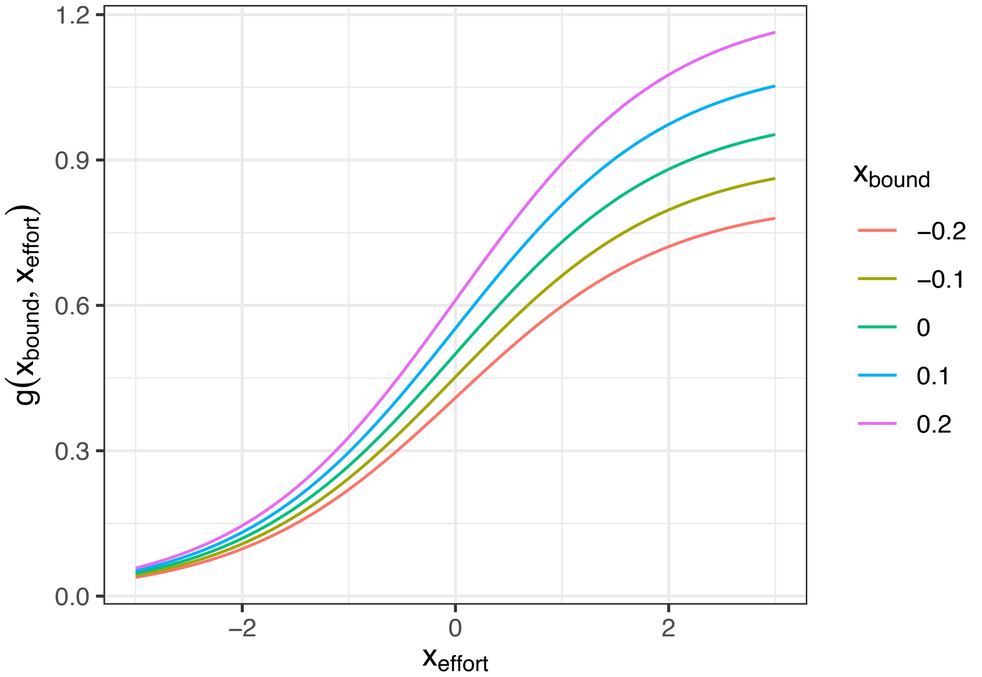


- Observed first arrival is **biased** towards later dates for low effort, but is the true one for very high effort

- Implementation: $Z_i \sim \mathrm{GEV}(\mu_i, \sigma_i) \quad \text{with} \quad \mu_i = g(\mathrm{Predictors}_i, \mathrm{Effort}_i)$

  → Nonlinear function g reaches (unknown) finite upper bound for very high effort
  → Infer g from data
  → Set very high effort for bias-corrected predictions

⚠ Source of high computational complexity



Implementation as a Bayesian hierarchical model

- **MCMC** implementation with Gibbs updates and Metropolis-Adjusted Langevin Algorithm (MALA) updates for the latent Gaussian components to improve mixing of chains. ~3secs per iteration, 80k burn-in, 20k for posterior evaluations with thinning

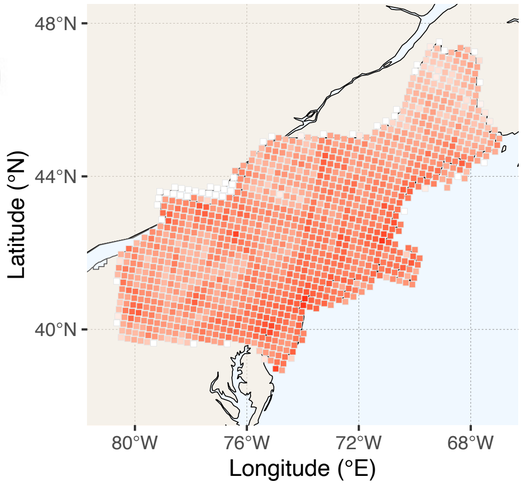
- **Vecchia** approximations for Gaussian random effects with exponential covariance (penalized complexity priors and sum-to-zero constraints)

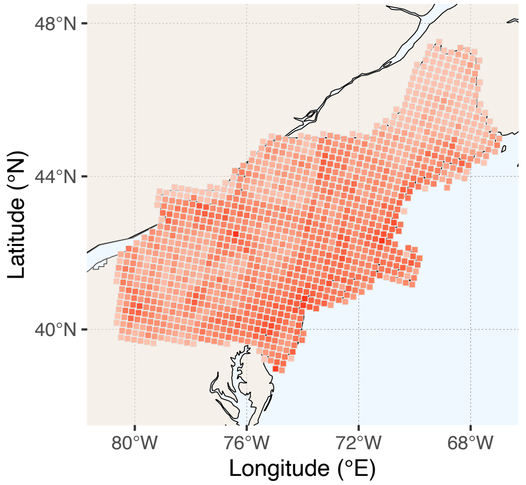
- **Simulation** study confirms relevant parameters/ecological processes are identifiable.

Goodness-of-fit of estimated models

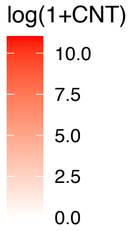
- Generally good match of eBird observations (left maps) with posterior means (right maps)
- Slight differences due to information shared from BBS

$N_i^{\mathrm{ckl}}(2021)$



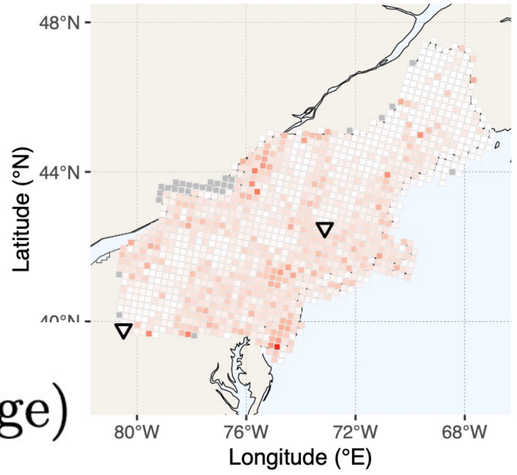


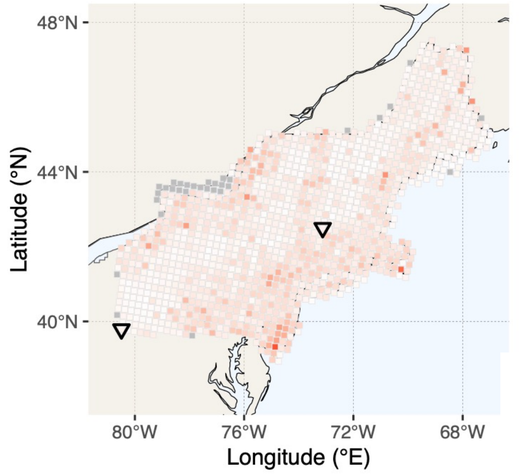
$\widehat{\lambda}_i^{\mathrm{ckl}}(2021)$

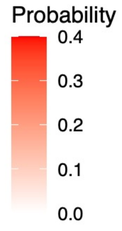


Example species:

Great Crested Flycatcher







$\dfrac{N_i^{\mathrm{spc}}}{N_i^{\mathrm{ckl}}}(\text{Time-average})$

$\widehat{p}_i^{\,\mathrm{spc}}(\text{Time-average})$

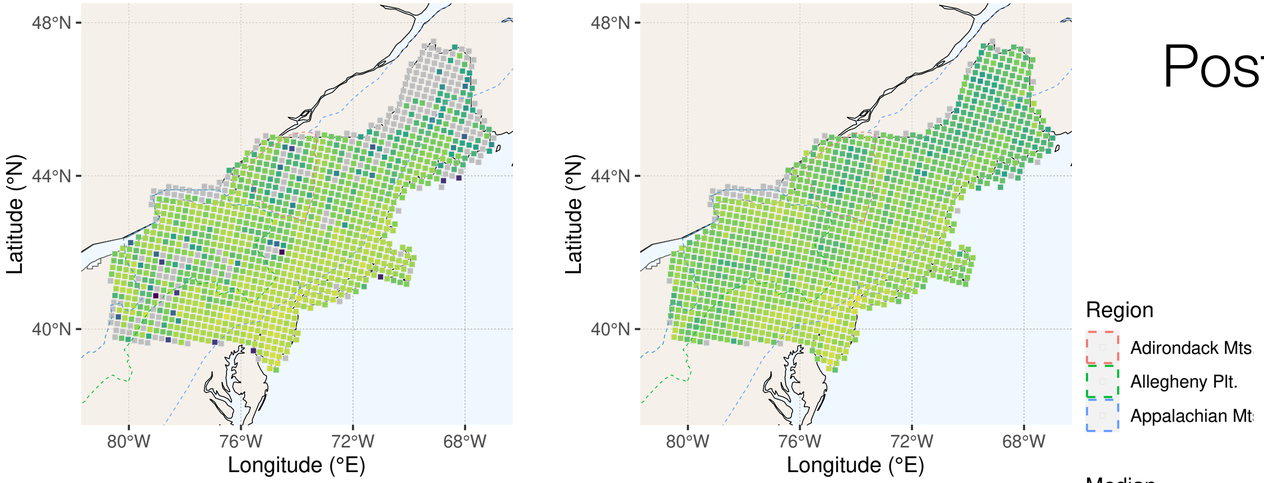
Illustration of bias-corrected prediction of first arrivals (2022)

- Based on Generalised Extreme-Value response
- Bias-corrected prediction by fixing saturated observational effort

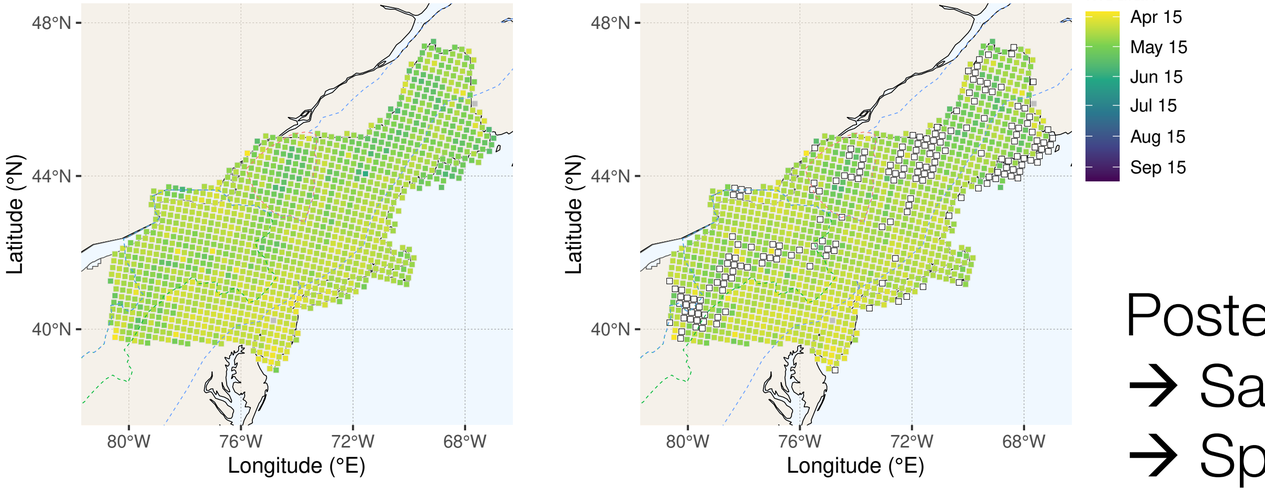
$$Z_i \mid \mu, \boldsymbol{\theta}^{\mu}, \sigma, \boldsymbol{\theta}_{\sigma} \sim \mathrm{GEV}\{\mu(\boldsymbol{s}_i, t_i; \boldsymbol{\theta}_{\mu}), \sigma(\boldsymbol{s}_i; \boldsymbol{\theta}_{\sigma}), \xi\}$$

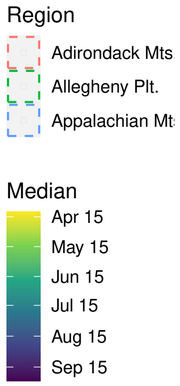
Great Crested Flycatcher

Observed



Posterior





Posterior predictive
→ Saturated effort

Posterior predictive
→ Saturated effort
→ Species niche

Illustration of bias-corrected prediction of first arrivals (2022), cont'd

- Table of estimated key parameters and first arrival dates for **two pixels**
- Estimated (not bias-corrected) first arrivals tend to occur relatively earlier for
  - higher Preference,
  - higher Activity and
  - in the core area of the niche




Species	Chimney Swift	Great Crested Flycatcher	Chestnut-sided Warbler	Purple Martin
$\hat{\theta}^{\text{pref}}$	0.191 (0.184,0.202)	0.204 (0.199,0.21)	0.187 (0.183,0.191)	0.2 (0.178,0.217)
$\hat{\theta}^{\text{act}}$	-0.15 (-0.217,-0.061)	-0.818 (-0.911,-0.696)	-0.548 (-0.619,-0.454)	-0.03 (-0.269,0.236)
$\hat{\theta}^{\text{niche-GEV}}$ $(\times 10^{-2})$	4.9 (4.664,5.134)	4 (3.894,4.133)	0.2 (0.17,0.278)	6 (5.541,6.443)
Observed	NA	NA	NA	NA
Predicted	09/05	03/05	21/05	~~07/06~~
Debiased	03/04	13/04	03/05	~~28/03~~
Observed	01/05	04/05	04/05	29/06
Predicted	09/05	15/05	12/05	~~12/05~~
Debiased	22/04	05/05	03/05	~~07/04~~

Discussion: Ecological data fusion using latent processes

Bayesian Hierarchical Models

- **Incomplete** and **biased** observation of true processes
- **Interpretable** latent processes for effort and relevant ecological properties
  → Identifiability thanks to shared random effects,
  but challenging validation
- Towards **spatiotemporal**, not purely spatial, modelling
  → Improve modelling of temporal dynamics
  ⚠ Requires disentangling complex observational/ecological dynamics
- Could we implement shared latent processes in other **learning** algorithms? (GAMs, ANNs, Random Forests…)
- **Multi-species** modelling?

Discussion: Bias and uncertainty reduction

Citizen Science

- Checklist data, such as eBird, allow generating **pseudo-absences**, but many opportunistic datasets are less structured
- Could we use these opportunitistic datasets in **machine learning** algorithms?
- Data fusion of opportunistic and structured data in ***Integrated Species Distribution Models*** is crucial (Fithian et al 2015; Isaac et al 2020)
- Collecting additional exhaustive field data may be necessary
  → Explore optimal sampling design through simulation studies?

Discussion: Opportunities for ecological extreme-value analysis

Extreme-Value Theory

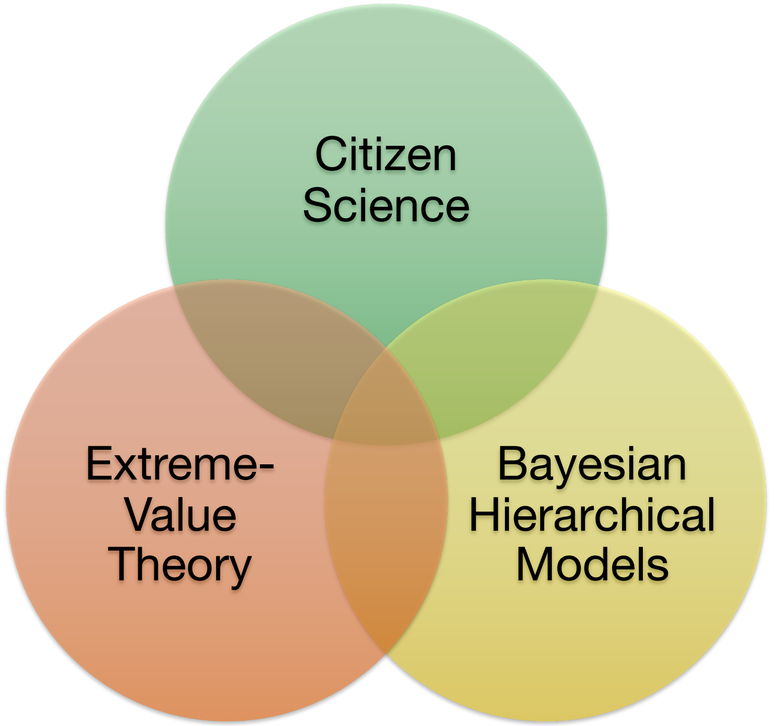
- EVT generally less relevant for discrete data but promising for modelling extreme phenological events, such as first arrivals
- GEV isn‘t the only model
- Lots of suggestions/ideas to improve mixing of MCMC
- EVT widely used for extreme climate and environmental events
  → Such events can drive strong species population shifts
  → Focus on specific events, not only long-term climate averages

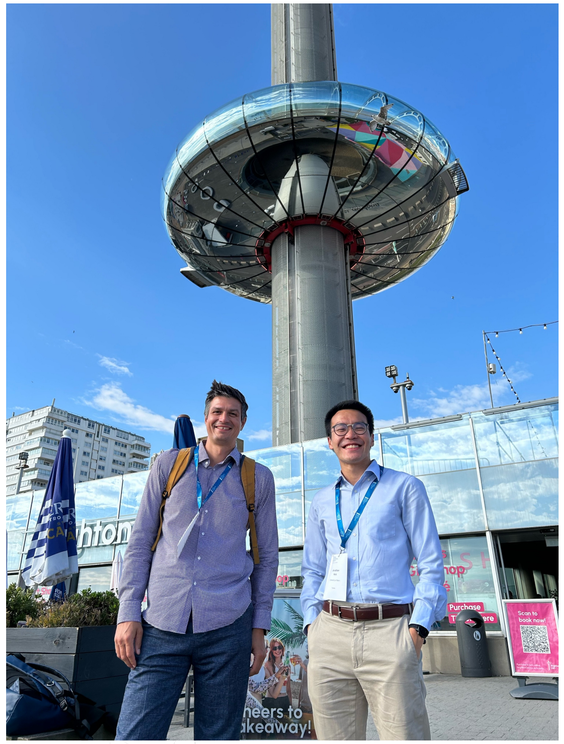
Outlook

- Citizen science datasets: *Small Data* and *Big Data*, but always complex…

  → Wide opportunities for modelling and decision support

  → An exciting potential playground for statisticians
  (and also those in extremes!)





ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

The RSS is a professional body for all statisticians and data analysts wherever they may live.

Membership | News and Publications | Training and events | Policy and campaigns | Jobs and careers | Resources

Sign in | Join

About RSS

Home > News > 2023 > Member callouts > **Call for discussion papers 2023: Analysis of citizen science data**

Call for discussion papers 2023: Analysis of citizen science data

03.05.23 | Member callouts

Citizen science involves volunteers who participate in scientific research by collecting data, monitoring sites, and even taking part in the whole process of scientific inquiry (Roy et al. 2012, Scyphers et al. 2015). In the past two decades, citizen science (also called participatory or community-based monitoring) has gained tremendous popularity (Bonney et al. 2009, Danielsen et al. 2014; Aceves-Bueno et al., 2017). These opportunistic datasets can be substantially large, numbering hundreds of thousands of units and present many statistical challenges in terms of sparse and missing data, issues with data collection

Special thanks to the discussion contributors!

- Kerrie Mengerson (Seconder)
- Ben Swallow (Proposer)
- Kuldeep Kumar
- Leo Belzile & Rishikesh Yadav
- Andrej Srakar
- Raphael Huser & Andrew Zammit-Mangion
- Abdelaati Daouia & Gilles Stupfler
- Allan Reese
- Stefano Rizzelli
- Maozai Tian

Food for thought

This work:

Koh, Opitz (2024). Extreme-value modelling of migratory bird arrival dates: Insights from citizen science data. Journal of the Royal Statistical Society, Series A (Statistics in Society).

Other literature:

- Adjei et al. (2023). A structural model for the process of collecting biodiversity data. Authorea Preprints.
- Adjei et al. (2023). The Point Process Framework for Integrated Modelling of Biodiversity Data. arXiv:2311.06755.
- Belmont et al. (2024). Spatio-temporal Occupancy Models with INLA. arXiv:2403.10680.
- Coles (2001). An introduction to statistical modeling of extreme values. Springer.
- Diggle et al. (2010). Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society Series C: Applied Statistics.
- Fithian et al. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution.
- Gelfand & Shirota (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs.
- Isaac et al. (2020). Data integration for large-scale models of species distributions. Trends in Ecology & Evolution.
- Lindgren et al. (2024). *inlabru*: software for fitting latent Gaussian models with non-linear predictors. arXiv:2407.00791.
- Tang et al. (2021). Modeling spatially biased citizen science effort through the eBird database. Environmental and Ecological Statistics.
- Wijeyakulasuriya et al. (2024). Modeling First Arrival of Migratory Birds Using a Hierarchical Max-Infinitely Divisible Process. Journal of Agricultural, Biological and Environmental Statistics.

Upcoming Talks

19 December 2025 - Jonathan Koh (ETH Zürich)

(16:00 UTC) = (11:00 New York) = (17:00 Paris) = (00:00 Shanghai) = (03:00 Sydney)

Tail calibration of probabilistic forecasts

Abstract: Probabilistic forecasts comprehensively describe the uncertainty in the unknown future outcome, making them essential for decision making and risk management. While several methods have been introduced to evaluate probabilistic forecasts, existing evaluation techniques are ill-suited to the evaluation of tail properties of such forecasts. However, these tail properties are often of particular interest to forecast users due to the severe impacts caused by extreme outcomes. In this work, we introduce a general notion of tail calibration for probabilistic forecasts, which allows forecasters to assess the reliability of their predictions for extreme outcomes. We study the relationships between tail calibration and standard notions of forecast calibration, and discuss connections to peaks-over-threshold models in extreme value theory. Diagnostic tools are introduced and applied in a case study on European precipitation forecasts.

22 January 2026 - Axel Bücher (Ruhr University Bochum)

(15:00 UTC) = (10:00 New York) = (16:00 Paris) = (23:00 Shanghai) = (02:00 Sydney)

Latent linear factor models for tail dependence in high dimensions

Abstract: A common object to describe the extremal dependence of a d-variate random vector X is the stable tail dependence function L. Various parametric models have emerged, with a popular sub-class consisting of those stable tail dependence functions that arise for linear and max-linear factor models with heavy tailed factors. We study such models under the assumption that the factors are possibly dependent, which results in a model for L that depends on a (d x K) loading matrix