To Merge or Not to Merge?

A circular optimal transport-based framework for comparing and fusing seasonality pattern data across participatory science platforms

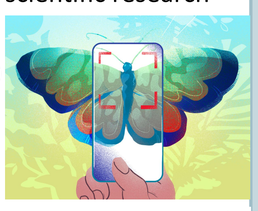
Cody Carroll[1], Robert E. Furrow[2], Laci M. Gerhart[2]

[1]University of San Francisco [2]University of California, Davis

Problem Overview

Crowd-sourced Data & Participatory Science

- Participatory / citizen science = publicly available programs/apps/platforms through which amateur and nonprofessional scientists engage and contribute to scientific research
- Examples:
  - iNaturalist (any organism)
  - eBird
  - Merlin (birds)
  - Monarch Watch
  - SCOOL (clouds)



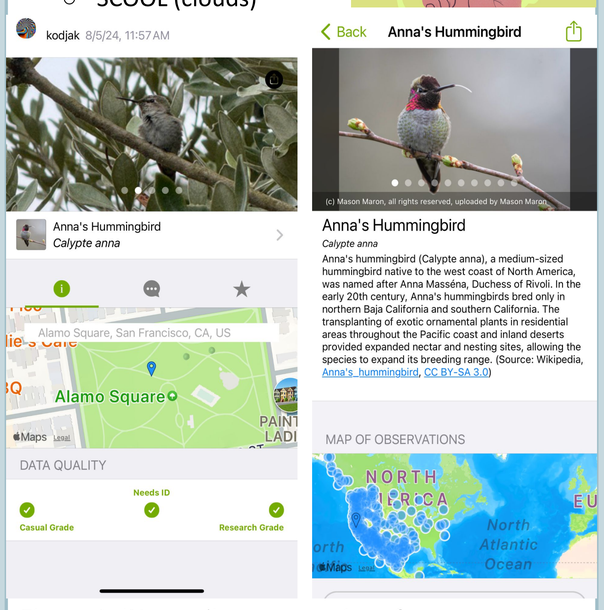


Figure 1. iNaturalist app user interface.

Is crowd-sourced data reliable for scientific use?

- Some show comparable precision between professional- and volunteer-collected data[6,7]; Others caution against overstating reliability of crowd-sourced data[1]
- Use in ecological studies or machine learning models requires care
  - Need to statistically address error and bias: spatial, temporal, taxonomic, observer, sociopolitical (+ more!)[4,9]

Our Main Question:

Can we develop a quantitative method to assess whether seasonality data for a given bird species is "consistent" across two platforms, e.g., eBird & iNaturalist?

- If we find consistency → **merge** data across platforms:
  - increases sample size
  - establishes a framework for leveraging multiple platforms' data at once
- If not → we learn something interesting about the data collection biases for that species across **eBird** & **iNaturalist**.

Data Sources

- Observation count data from *iNaturalist* and *eBird* APIs for 254 species of birds present in Northern CA and Nevada in 2022 (Fig. 2)

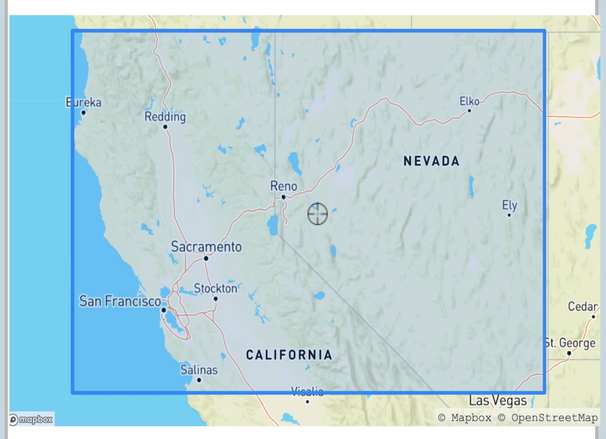
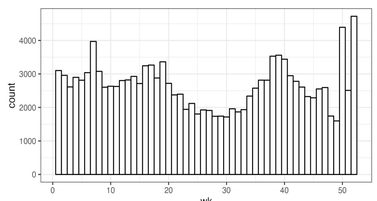


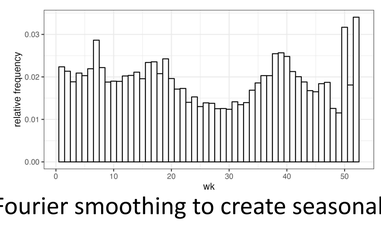
Figure 2. Bounding box for region of interest.

Data Processing

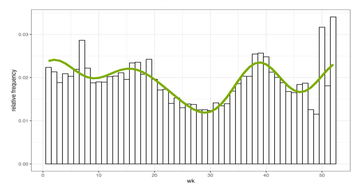
1. Start with weekly raw observation counts



2. Scale by total year's count for relative frequency



3. Apply Fourier smoothing to create seasonality curve



4. Visualize temporal distribution in polar coordinates

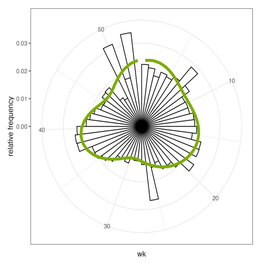
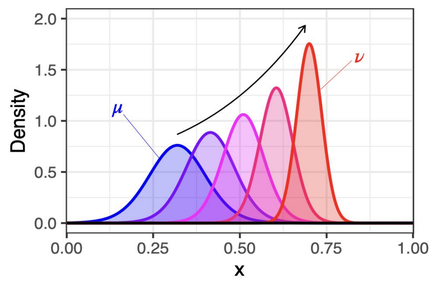


Figure 3. Example data processing pipeline for the California scrub-jay's eBird data. The resulting processed data is a temporal distribution over the span of the year per species.

Circular Optimal Transport Distance

How can we measure the discrepancy between two temporal distributions, μ and ν?



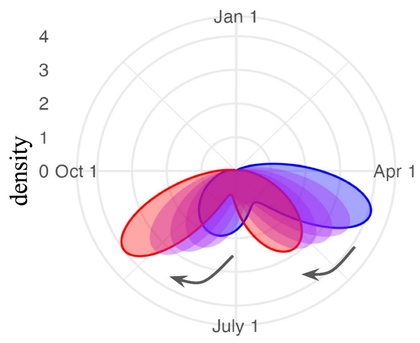


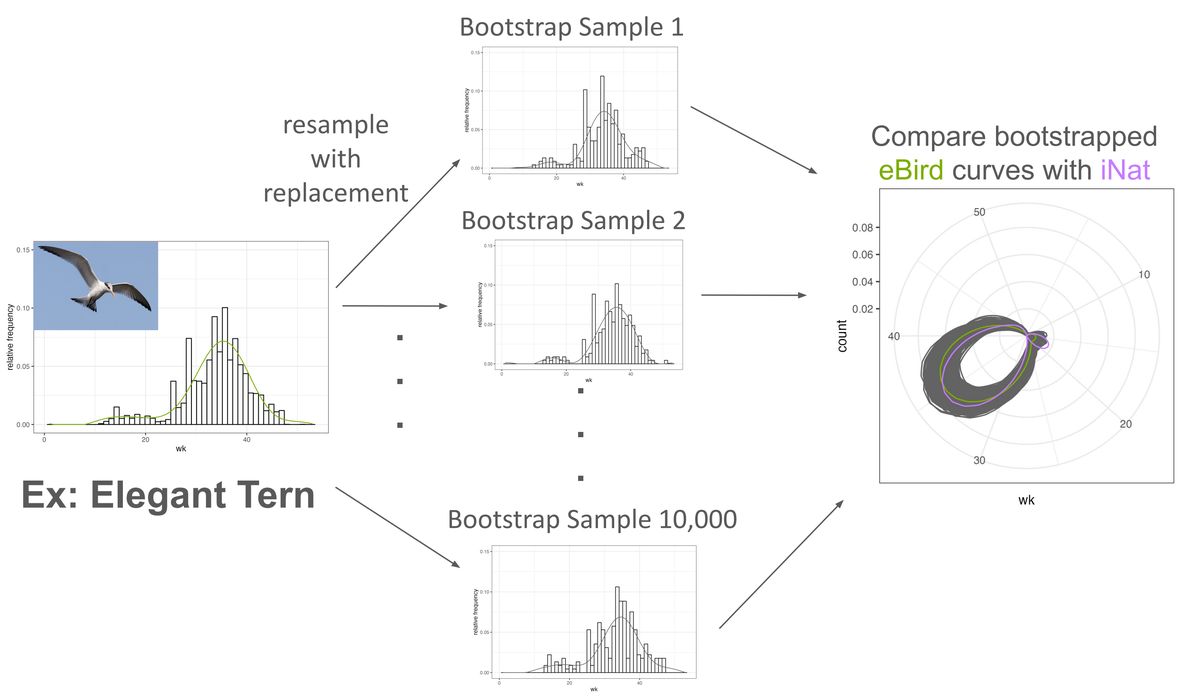
Figure 4. Optimal transport distance[10] (top) quantifies discrepancies between distributions as the energy required to move the mass of one distribution to the other. Circular optimal transport[5] (bottom) modifies this to respect a circular domain, reflecting the seasonality of the year. Mathematically,

$$COT(\mu,\nu) = \inf_{\alpha\in\mathbb{R}} \int_0^1 |F_\mu(t) - F_\nu(t) - \alpha| dt$$

where F_μ and F_ν represent the cumulative distribution functions of μ and ν.

Bootstrap Hypothesis Testing Approach

Question: How much distance is too much to be considered "mergeable"?
Idea: Resample from a reference distribution (take, e.g., eBird) and see if the merge candidate distribution (iNaturalist) falls within the typical range of variation.



Use pairwise COT distances to quantify discrepancy and create a null distribution for comparison.

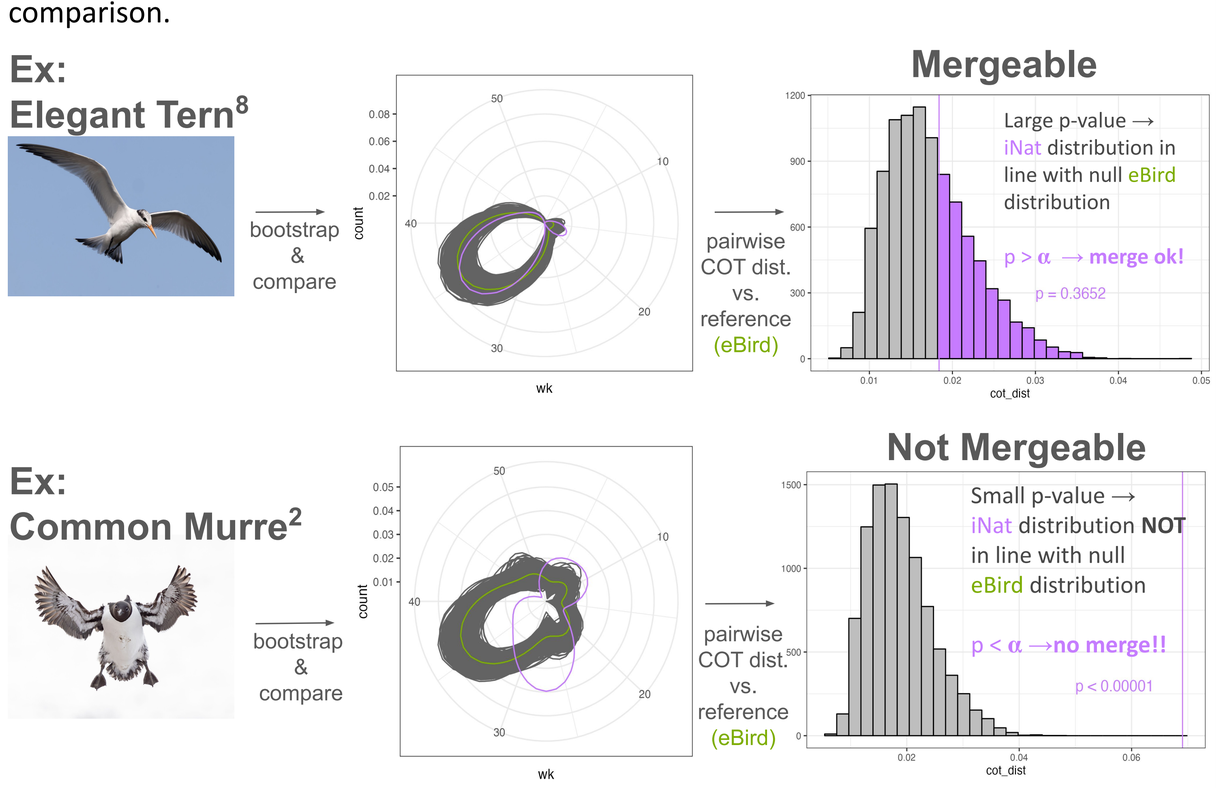


Figure 5. Schematic examples of bootstrap hypothesis testing procedure to test for mergeability, taking the significance level α. Process is replicated for all 254 species & adjusted for multiple testing using Benjamini-Yekutieli[3] procedure.

Results

What fraction of species were mergeable across eBird (2022) and iNaturalist (2022) data?

How about if we replicate the process comparing similar count data from 2019 with eBird 2022 data?

Database/ Year	iNat '22	iNat '19	eBird '19
% of mergeable species w/ eBird '22	97.6% (N = 248)	88.6% (N = 225)	97.2% (N = 247)

Table 1. Fraction of species (out of 254) mergeable with the eBird 2022 dataset according to the two-sample COT test after Benjamini-Yekutieli correction.

Main Result: For the vast majority of species, seasonality patterns are consistent & mergeable between eBird/iNaturalist across platforms and years despite differences in platform structures & user profiles/behaviors!

Discussion

- We established a merge criteria & statistical framework for participatory science projects (small or large!) to contribute to broadscale analyses by organizing and pooling data across projects.
- Post-hoc analysis showed that 7 archetypal seasonality patterns emerged from clustering temporal distributions: *spring, summer, fall, winter, year-round, bimodal, and anomalous*.
- Our team consisted of a participatory science expert, an ornithological expert, and a statistician. Deliberate, project-informed interdisciplinary team construction is essential in participatory science-driven research.

References

1. Aceves-Bueno, E., Adeleye, A.S., Feraud, M., Huang, Y., Tao, M., Yang, Y. and Anderson, S.E. 2017. The Accuracy of Citizen Science Data: A Quantitative Review. *The Bulletin of the Ecological Society of America*, 98(4): pp. 278–290.
2. Artemeva, L. 2023. Common Murre (*Uria aalge*) Photo 301430202, posted to iNaturalist 7/18/2023 (no rights reserved), accessed 11/18/2024.
3. Benjamini, Y. and Yekutieli, D. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency, *The Annals of Statistics*, 29(4): pp. 1165–1188.
4. Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., et al. 2014. Statistical solutions for error and bias in global citizen science datasets, *Biological Conservation*, 173: pp. 144–154.
5. Hundreiser, S., Klatt, M. and Munk, A. 2021. *The Statistics of Circular Optimal Transport. Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale*. Singapore: Springer Nature Singapore, 2022. pp. 57–82.
6. Kosmala, M., Wiggins, A., Swanson, A. and Simmons, B. 2016. Assessing data quality in citizen science, *Frontiers in Ecology and the Environment*, 14(10): pp. 551–560.
7. Lewandowski, E. and Specht, H. 2015. Influence of volunteer and project characteristics on data quality of biological surveys, *Conservation Biology*, 29(3): pp. 713–723.
8. Mac 2023. Elegant Tern (*Thalasseus elegans*) Photo 316600192, posted to iNaturalist 8/22/2023 (CC BY-NC), accessed 11/18/2024.
9. Sierra, E., Gillespie, L. E., Soltani, S., Exposito-Alonso, M., & Kattenborn, T. 2025. DivShift: Exploring Domain-Specific Distribution Shifts in Large-Scale, Volunteer-Collected Biodiversity Datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(27), 28386-28396.
10. Villani, C. 2021. *Topics in optimal transportation* (Vol. 58). American Mathematical Soc.