DISSERTATION

A CASE FOR CONTEXT IN QUANTITATIVE ECOLOGY: STATISTICAL TECHNIQUES TO INCREASE EFFICIENCY, ACCURACY, AND EQUITY IN BIODIVERSITY RESEARCH

Submitted by

Hanna M. McCaslin

Graduate Degree Program in Ecology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2024

Doctoral Committee:

Advisor: Sara Bombaci

Mevin Hooten
David Koons
Jennifer Hoeting

ABSTRACT

A CASE FOR CONTEXT IN QUANTITATIVE ECOLOGY: STATISTICAL TECHNIQUES TO INCREASE EFFICIENCY, ACCURACY, AND EQUITY IN BIODIVERSITY RESEARCH

The current era of ecological research is characterized by rapid technological innovation, large datasets, and numerous computational and quantitative techniques. Together, big data and advanced computing are expanding our understanding of natural systems, allowing us to capture more complexity in our models, and helping us find solutions for salient challenges facing modern ecology and conservation, including climate change and biodiversity loss. However, large datasets are often characterized by noise, complex observational processes, and other challenges that can impede our ability to apply these data to address ecological research gaps. In each chapter of this dissertation, I seek to address a data problem inherent to the 'big data' that characterizes modern ecological research. Together, they extend the strategies available for addressing a problem facing many ecologists – how to make use of the large volumes of data we are collecting given (1) current computational limitations and (2) specific sampling biases that characterize various methods for data collection.

In the first chapter, I present a recursive Bayesian computing (RB) method that can be used to fit Bayesian hierarchical models in sequential MCMC stages to ease computation and streamline hierarchical inference. I also demonstrate the application of transformation-assisted RB (TARB) to a hierarchical animal movement model to create unsupervised MCMC algorithms and obtain inference about individual- and population-level migratory characteristics. This recursive procedure reduced computation time for fitting our hierarchical movement model by

half compared to fitting the model with a single MCMC algorithm. Transformation-assisted RB is a relatively accessible method for reducing the computational demands of fitting complex ecological statistical models, like those for animal movement, multi-species systems, or large spatial and temporal scales.

Biodiversity monitoring projects that rely on collaborative, crowdsourced data collection are characterized by huge volumes of data that represent a major facet of 'big data ecology,' and quantitative methods designed to use these data for ecological research and conservation represent a leading edge of contemporary quantitative ecology. However, because participants select where to observe biodiversity, crowdsourced data are often influenced by sampling bias, including being biased toward affluent, white neighborhoods in urban areas. Despite the growing evidence of social sampling bias, research has yet to explore how socially driven sampling bias impacts inference and prediction informed by crowdsourced data, or if existing data pre-processing or analytical methods can effectively mitigate this bias. Thus, in Chapters 2 and 3, I explored social sampling bias in data from the crowdsourced avian biodiversity platform eBird. In Chapter 2, I studied patterns of social sampling bias in the locations of eBird "hotspots" to determine whether hotspots in Fresno, California, U.S.A. are more biased by social factors than the locations of Fresno eBird observations overall. My findings support previous work showing that eBird locations are biased by demographics. Further, I found that demographic bias is most pronounced in the locations of hotspots specifically, with hotspots being more likely to occur in areas with higher proportions of non-Hispanic white residents than eBird locations overall. This relationship is reinforced because hotspots in these predominantly white areas also amass more eBird checklists overall than hotspots in areas with more demographic diversity. These findings raise concerns that the eBird hotspot system may be exacerbating spatial bias in sampling and

reinforcing patterns of inequity in data availability and eBird participation, by leading to datasets and user-facing maps of birding hotspots that mostly represent predominantly white neighborhoods. Then, in Chapter 3, I investigated the impacts of not accounting for socially biased sampling when using eBird data to study patterns of urban biodiversity. The luxury effect has emerged as a prominent hypothesis in urban ecology, describing a pattern of higher biodiversity associated with greater socioeconomic status observed in many cities. Using eBird data from 2015-2019, I tested whether an avian luxury effect is observed in Raleigh-Durham, North Carolina, U.S.A. before and after accounting for social sampling bias. By jointly modeling sampling intensity and species richness, I found that sampling intensity and species richness are positively correlated and sampling bias influences the estimated relationship between species richness and income. Thus, failing to account for sampling bias can hinder our ability to accurately observe social-ecological dynamics. Additionally, I found that randomly spatially subsampling eBird data prior to analysis, as recommended by existing guidelines to mitigate sampling bias in eBird data, does not reduce biased sampling related to demographics, because there are data gaps in communities of color and low-income communities that cannot be addressed via spatial subsampling. Therefore, it is paramount that crowdsourced and contributory science projects prioritize more equitable participation in their platforms, both for more ethical, equitable practice and because current sampling inequity negatively impacts data quality and project goals. Quantitative techniques can help us understand the complex observational processes influencing ecological data, and each chapter of this dissertation highlights how tailoring statistical or computing methods to these observational contexts can advance ecological knowledge – either by extending the complexity of models we can feasibly

fit, as in Chapter 1, or by acknowledging and accounting for sampling inequity, in Chapters 2 and 3.

We are all participants actively shaping the ecological processes we observe, and the actions, approaches, and assumptions used in our research reflect societal systems and biases. Data are never objective, and it is dangerous and false to assume that quantitative techniques can take data out of the contexts in which they were collected. Instead, quantitative frameworks that embrace, reflect, and seek to improve the ways in which social and observational contexts inform what is observed can elevate analytical techniques to tools towards more just, inclusive, and transparent ecological research and conservation.

ACKNOWLEDGMENTS

Thank you to my advisor, Dr. Sara Bombaci, for her mentorship and unwavering belief in my ability to succeed. I am beyond grateful for Sara's support and trust as my dissertation has evolved throughout my time at CSU and for her grace and understanding during moments when life took precedence over research. Sara has been an invaluable role model and mentor, and I am very grateful to have an advisor who has always acknowledged and honored my humanity and individuality first in our interactions and collaborations. I look to her for inspiration for how to create a successful academic career while centering justice, equity, and inclusion in research and teaching.

Thank you to my committee, Mevin Hooten, David Koons, and Jennifer Hoeting for their quantitative expertise, perspective, and flexibility. In particular, I am appreciative of Mevin's mentorship and guidance in developing a quantitative skillset and perspective that I am eager to apply to new ecological questions as I continue my career. I am also grateful for funding from a National Science Foundation Graduate Research Fellowship (grant 1840343) and the American Association of University Women, whose financial support has allowed me to explore diverse research interests.

I am incredibly grateful for my friends and family, and the invaluable community of friends, peers, and collaborators in Fort Collins who I have leaned on for support personally and professionally over the past several years. The community I have found here has become a home, and I would not be completing this degree without it. In particular, thank you to Abbey Feuka for collaborating on the first chapter of this dissertation – this experience set the bar for what I hope collaboration will always be (though rarely is). I have been fortunate to have been welcomed by

and learn from several lab groups within the Department of Fish, Wildlife, and Conservation Biology at CSU – thank you to the members of the Hooten lab, Kyle Horton and the AeroEco lab, and the Bombaci lab for their help crafting and refining these projects and their invaluable feedback and camaraderie along the way. I also thank the journal editor and anonymous reviewers whose insights improved chapter one during publication, and J. Tipton for early discussions on this work. Numerous others have been especially influential, helpful, and supportive as I have worked on this dissertation and grown as a scientist, educator, and person over the past several years, including Amy Collins, Kristin Davis, Gemara Gifford, Nathan Hahn, Emma Hanslowe, Audrey Harris, Mikko Jimenez, Jenna Parker, and Tawni Riepe.

Next, thank you to everyone who has participated in crowdsourced data collection and contributed data to eBird or other platforms, especially those who have had to overcome inequity, barriers to participation, and stereotypes to do so. I am grateful for everyone who is, in some way, working to making birding, contributory science, and conservation more inclusive. I am also very appreciative of the Fresno, Tucson, and Triangle Bird Counts for sharing their data, and to Deja Perkins, Jin Bai, and Madhusudan Katti from the Triangle Bird Count at North Carolina State University for their help and insights on inequitable sampling in eBird data.

Finally, I acknowledge, with the utmost respect and humility, the Nunt'zi (Ute), Inunaina (Arapaho), So'taa'e (Cheyenne) Nations and other Native peoples who violently lost their homeland, relatives, and lives for the founding of Colorado State University, the land-grant institution from where I have conducted this work.

DEDICATION

For Lily.

I love you, I miss you,

and I am ever inspired by your unwavering dedication to justice

and living the perfect outdoor leisure day.

TABLE OF CONTENTS

INTRODUCTION

The current era of ecology is characterized by big data and rapid technological advances (Farley et al., 2018; La Sorte et al., 2018; McCallen et al., 2019). Technologies including satellite imagery, passive audio recording devices, lightweight high-resolution telemetry units, smartphones, and countless others now enable us to collect ecological data in nearly every region of the world, across many scales (Nathan et al., 2022; van Klink et al., 2022). For example, in movement ecology, an array of tracking technologies including GPS, reverse GPS, and noninvasive techniques like radar and computer vision are facilitating the collection of data across broad spatiotemporal scales – from one-meter locations every second to global-scale tracking over multiple years (Nathan et al., 2022). These data are being leveraged to deepen our understanding of relationships between animal movement and physiology, behavior, and inter- and intraspecific interactions (Nathan et al., 2022), as well as linking individual variation in movement and population-level processes (Hooten et al., 2016). Large scale remote sensing products and initiatives like the National Ecological Observatory Network (NEON) in the United States then allow us to link ecological observations with other biotic and abiotic data to advance our knowledge of how climate change and landscape modification are influencing ecological processes (Nagy et al., 2021; Ustin and Middleton, 2021).

'Big data ecology' is also characterized by the increasing scope and applications of crowdsourced biodiversity platforms like eBird and iNaturalist (Sullivan et al., 2014; Nugent, 2018). The prevalence of smartphones has driven rapid growth in participation in these projects (also known as contributory science or citizen science projects), and the hundreds of millions of resulting observations inform ecological research and conservation around the world (Hampton

et al., 2013; Chandler et al., 2017). For example, the Global Biodiversity Information Facility (GBIF) is a platform that centralizes biodiversity data from projects of all types and scopes, and currently over 70% of the > 2.5 billion animal observations on GBIF come from crowdsourced projects (Bonney, 2021). Crowdsourced biodiversity data are being used to create dynamic, global maps of avian species distributions (Fink et al., 2020), inform progress toward the United Nation's Sustainable Development Goals (Fraisl et al., 2020), and conduct extensive biodiversity monitoring across scales (Hampton et al., 2013; Fraisl et al., 2022). Research into techniques for leveraging crowdsourced data for biodiversity forecasting, integration with other data sources, and creation of dynamic species distribution models represents the leading edge of statistical ecological research and computing (Callaghan et al., 2021)

Simultaneously, 'big data ecology' is complemented by advancements in computing and quantitative techniques, including cutting edge machine learning and artificial intelligence, to approach our biggest environmental challenges, like combatting climate change and biodiversity loss (Antonelli et al., 2022; Eastwood et al., 2022; Leal Filho et al., 2022). Quantitative methods can increase the utility of data that has been previously collected as well as prioritize, manage, and integrate new data that will be collected. Developments like hierarchical models and data integration techniques can extend the scope and relevance of existing data sets and facilitate the integration of multiple datasets to link ecological processes across multiple spatiotemporal scales (Zipkin and Saunders, 2018; Isaac et al., 2019; Frost et al., 2023). Computing advances and statistical innovations have enabled ecologists to progressively fit models to larger and more complex datasets with greater flexibility (McCallen et al., 2019), and ecologists are increasingly expected to learn skills and techniques to conduct data-intensive research (Visser et al., 2015; Hampton et al., 2017). Many novel statistical approaches have emerged as techniques for

managing complex observational processes and other challenges related to the application of crowdsourced data to ecological research (Johnston et al., 2023). For example, machine learning is preparing us for the massive volumes of data we are currently collecting and will continue to collect, and automation, black-box algorithms, and the online, real-time updating of data products are becoming more prevalent in ecology (Peters et al., 2014). Deep learning algorithms trained elsewhere can be used for image and sound classification and prediction in ecological contexts, drastically speeding up data processing and research progress, and machine learning and citizen science are being combined to automate species recognition (Willi et al., 2018; Koch et al., 2022).

Together, big data and computing are expanding our understanding of natural systems, allowing us to capture more complexity in our models than ever before, and helping us address research gaps and find solutions for salient challenges facing modern ecology and conservation. However, the collection and analysis of ecological data can have substantial impacts on people, animals, and the ecological systems being observed. Researchers and conservationists are increasingly questioning the ethics of conducting ecological research and practicing conservation, both with regards to how researchers impact the communities and systems they interact with and given the increasing prevalence of artificial intelligence in the field. In the United States and across western institutions, ecologists are increasingly acknowledging and grappling with conservation's roots in colonialism, extractive research approaches, and the exclusion of marginalized communities from conservation science and practice (Murdock, 2021). Conservation, data collection, and research often perpetuate and reinforce these colonial, exploitative, and violent dynamics (David-Chavez, 2019; Domínguez and Luoma, 2020; Liboiron, 2021). Many within our field are rejecting practices that violate Indigenous

sovereignty, interfere with cultural practices and livelihoods, and risk harming wildlife (Domínguez and Luoma, 2020; Cooke et al., 2022). Modifying or ceasing research practices that perpetuate these harms is the primary priority for more ethical ecology. Quantitative methodologies are no exception, and we must recognize the potential for quantitative ecological research to reinforce harmful systemic biases, exclusion, and injustice.

For example, crowdsourcing and citizen science projects often ask the public to voluntarily collect data without clear definitions of the goals, potential implications, and scope of use for the data. At the same time, participation in crowdsourced data collection is inequitable globally and locally, but it is often assumed in analysis that the identities of participants have negligible influence on the data collected (Haklay, 2016). As a result, the knowledge gained through crowdsourced projects often benefits those with greater power and privilege and can further marginalize areas without the ability to invest time and resources in data collection (Montanari et al., 2021; Mahmoudi et al., 2022). In the United States, crowdsourced data reflect the racialized, colonial, and class-based histories that have shaped current social and ecological landscapes (Mahmoudi et al., 2022). Analytical assumptions and emphases on maximizing data quantity that overlook the social context of data collection homogenize the social landscape and impede accuracy and understanding while promoting geographical discrimination in knowledge creation, environmental health, and conservation (Montanari et al., 2021; Mahmoudi et al., 2022).

Quantitative approaches can facilitate ethical data use by addressing pressing research gaps using extant data, interrupting the exploitative practice of indiscriminately collecting more data to keep pace with the advent of new observational technology (Gremillet et al., 2022). However, big data and quantitative advancements can also obstruct progress toward more just,

ethical ecological research and conservation. Quantitative ecology can reinforce unjust systems and hierarchies when used in attempts to delegitimize qualitative research or broader ways of knowing or promote a false sense of objectivity in ecological data and analysis (Reid and Rout, 2020; Pessach and Shmueli, 2020). Additionally, the proliferation of high-dimensional and nonlinear models, machine learning, and artificial intelligence in ecological research necessitates careful consideration and investigation of social biases hidden in data, because these techniques can easily reinforce systemic biases without detection (Garcia, 2017; Benthall and Haynes, 2019). Ecological data reflect the social factors that have modified landscapes, and model specification, assumptions, and tuning can perpetuate inequity or implicit biases of researchers (Gianfrancesco et al., 2018; Rudin et al., 2020). Thus, careful consideration of how data and modeling may perpetuate or reflect bias and acknowledgment of the subjectivity of data collection and quantification are paramount for minimizing the role of ecological research in the continuation of harmful systems.

With the proliferation of quantitative and statistical tools available for ecology, we must be cautious about generalizing techniques beyond their intended scopes and datasets. In many cases, we are still defining the appropriate applications and limitations of different approaches, and like many things in ecology, are working to find the balance between generalizability and specificity in our modeling techniques. As demonstrated in this dissertation, defining constraints and being cautious about the assumptions underlying our analytical techniques is critical not only for properly accounting for sources of variation, scale, and sample size, but also to understand the social context and potential implications of overlooking deeper issues like participation inequality or geographic discrimination hampering our data (Haklay, 2016; Montanari et al., 2022). Large ecological datasets share many similarities, including noise, effects of the

observational process, and scale and resolution challenges, but we must understand, account for, and communicate the unique ways they manifest in different datasets (Spake et al., 2022). Ultimately, this will improve the relevance, generalizability, and scalability of ecological research to advance largescale understanding, aid in local decision-making and action, and reduce widespread barriers to knowledge and environmental justice.

The chapters in this dissertation are unified by each seeking to address a data problem inherent to the 'big data' that characterizes modern ecological research. Together, they extend the strategies available for addressing a problem facing many ecologists – how to make use of the large volumes of data we are collecting given (1) current computational limitations and (2) specific sampling biases that characterize various methods for data collection. Though the approaches and findings of each chapter are generalizable beyond the specific study explored here, they also highlight the importance of considering the unique limitations and intentions distinguishing datasets and applications. Striving for more ethical data collection and use, as quantitative ecologists, it is our responsibility to remember context when working with data. With this dissertation, I hope to highlight how quantitative techniques can augment the use of existing datasets and emphasize the importance of always remembering context – of both the ecological system and observational method – when designing our analytical approaches.

REFERENCES

Antonelli, A., Dhanjal-Adams, K. L., and Silvestro, D. (2023). Integrating machine learning, remote sensing and citizen science to create an early warning system for biodiversity. *Plants, People, Planet*, *5*(3), 307-316.

Benthall, S., and Haynes, B. D. (2019, January). Racial categories in machine learning. In Proceedings of the conference on fairness, accountability, and transparency (pp. 289-298).

Bonney, R. (2021) Expanding the impact of citizen science. *BioScience,* 71(5), 448–451.

Callaghan, C. T., Poore, A. G., Mesaglio, T., Moles, A. T., Nakagawa, S., Roberts, C., Rowley, J.J., Vergés, A., Wilshire, J.H. and Cornwell, W.K. (2021). Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, *71*(1), 55-63.

Chandler, M., See, L., Copas, K., Bonde, A. M., López, B. C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A., and Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280-294.

Cooke, S. J., Michaels, S., Nyboer, E. A., Schiller, L., Littlechild, D. B. R., Hanna, D.E.L., Robichaud, C.D., Murdoch, A., Roche, D., Soroye, P., Vermaire, J.C., and Auld, G. (2022). Reconceptualizing conservation. *PLOS Sustainability and Transformation*, 1(5), e0000016.

David-Chavez, D. M. (2019). A guiding model for decolonizing environmental science research and restoring relational accountability with Indigenous communities (Doctoral dissertation, Colorado State University).

Domínguez, L., and Luoma, C. (2020). Decolonising conservation policy: How colonial land and conservation ideologies persist and perpetuate indigenous injustices at the expense of the environment. *Land*, *9*(3), 65.

Eastwood, N., Stubbings, W. A., Abdallah, M. A. A. E., Durance, I., Paavola, J., Dallimer, M., Pantel, J.H., Johnson, S., Zhou, J., Hosking, J.S., Brown, J.B., Ullah, S., Krause, S., Hannah, D.M., Crawford, S.E., Widmann, M., and Orsini, L. (2022). The Time Machine framework: monitoring and prediction of biodiversity loss. *Trends in Ecology and Evolution*, 37(2), 138-146.

Farley, S. S., Dawson, A., Goring, S. J., and Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*(8), 563-576.

Fink, D., T. Auer, A. Johnston, V. Ruiz-Gutierrez, W. M. Hochachka, and S. Kelling. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* 30: e02056.

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J. L., Masó, J., Penker, M. and Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science* 15 (6), 1735–1751.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P. Y., Danielsen, F., Hitchcock, C.B., Hulbert, J.M., Piera, J., Spiers, H., Thiel, M., and Haklay, M. (2022). Citizen science in environmental and ecological sciences. *Nature Reviews Methods Primers*, *2*(1), 64.

Frost, F., McCrea, R., King, R., Gimenez, O., and Zipkin, E. (2023). Integrated population models: Achieving their potential. *Journal of Statistical Theory and Practice*, *17*(1), 6.

Garcia, M. (2016). Racist in the Machine. *World Policy Journal*, 33(4), 111-117.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine,* 178(11), 1544-1547.

Grémillet, D., Chevallier, D., and Guinet, C. (2022). Big data approaches to the spatial ecology and conservation of marine megafauna. *ICES Journal of Marine Science*, *79*(4), 975-986.

Haklay, M. E. (2016). Why is participation inequality important? Ubiquity Press.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C.S., and Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156-162.

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R.R., Boettinger, C., Collins, S.L., Gross, L., Fernández, D.S., Budden, A., White, E.P., Teal, T.K., Labou, S.G., and Aukema, J. E. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, *67*(6), 546-557.

Hooten, M. B., Buderman, F. E., Brost, B. M., Hanks, E. M., and Ivan, J. S. (2016). Hierarchical animal movement models for population-level inference. *Environmetrics*, *27*(6), 322-333.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., and O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology and Evolution*, *35*(1), 56-67.

Johnston, A., Matechou, E., and Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103-116.

Koch, W., Hogeweg, L., Nilsen, E. B., and Finstad, A. G. (2022). Maximizing citizen scientists' contribution to automated species recognition. *Scientific Reports*, *12*(1), 7648.

La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., and Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. *The Condor: Ornithological Applications*, *120*(2), 414-426.

Leal Filho, W., Wall, T., Mucova, S. A. R., Nagy, G. J., Balogun, A. L., Luetz, J. M., Ng, A.W., Kovaleva, M., Safiul Azam, F.M., Alves, F., Guevara, Z., Matandirotya, N.R., Skouloudis, A., Tzachor, A., Malakar, K., and Gandhi, O. (2022). Deploying artificial intelligence for climate change adaptation. *Technological Forecasting and Social Change*, *180*, 121662.

Liboiron, M. (2021). *Pollution is Colonialism*. Duke University Press.

Mahmoudi, D., Hawn, C. L., Henry, E. H., Perkins, D. J., Cooper, C. B., and Wilson, S. M. (2022). Mapping for whom? Communities of color and the citizen science gap. *UMBC Faculty Collection*.

McCallen, E., Knott, J., Nunez-Mir, G., Taylor, B., Jo, I., and Fei, S. (2019). Trends in ecology: shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment*, 17(2), 109-116.

Montanari, M., Jacobs, L., Haklay, M., Donkor, F. K., and Mondardini, M. R. (2021). Agenda 2030's, "Leave no one behind," in citizen science? *Journal of Science Communication*, 20(06), A07-A07.

Murdock, E. G. (2021). Conserving dispossession? A genealogical account of the colonial roots of western conservation. *Ethics, Policy and Environment*, *24*(3), 235-249.

Nagy, R. C., Balch, J. K., Bissell, E. K., Cattau, M. E., Glenn, N. F., Halpern, B. S., ... and Zhu, K. (2021). Harnessing the NEON data revolution to advance open environmental science with a diverse and data-capable community. *Ecosphere*, *12*(12), e03833.

Nathan, R., Monk, C. T., Arlinghaus, R., Adam, T., Alós, J., Assaf, M., ... and Jarić, I. (2022). Big-data approaches lead to an increased understanding of the ecology of animal movement. *Science*, *375*(6582), eabg1780.

Nugent, J. (2018). iNaturalist: citizen science for 21st-century naturalists. *Science Scope*, *41*(7), 12-15.

Pessach, D., and Shmueli, E. (2020). Algorithmic fairness. arXiv preprint arXiv:2001.09784. https://doi.org/10.48550/arXiv.2001.09784.

Peters, D. P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, *5*(6), 1-15.

Reid, J., and Rout, M. (2020). Developing sustainability indicators–The need for radical transparency. Ecological Indicators, 110, 105941.

Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 2(1), 1.

Spake, R., O'dea, R. E., Nakagawa, S., Doncaster, C. P., Ryo, M., Callaghan, C. T., and Bullock, J. M. (2022). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology and Evolution*, 6(12), 1818-1828.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... and Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31-40.

Ustin, S. L., and Middleton, E. M. (2021). Current and near-term advances in Earth observation for ecological applications. *Ecological Processes*, *10*, 1-57.

van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T.T., Jongejans, E., Menz, M.H.M., Miraldo, A., Roslin, T., Roy, H.E., Ruczynski, I., Schigel, D., Schäffler, L., Sheard, J.K., Svenningsen, C., Tschan, G.F, Wäldchen, J., Zizka, V.M.A., Åström, J., and Bowler, D. E. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology and Evolution*.

Visser, M. D., McMahon, S. M., Merow, C., Dixon, P. M., Record, S., and Jongejans, E. (2015). Speeding up ecological and evolutionary computations in R; essentials of high performance computing for biologists. *PLoS Computational Biology*, 11(3), e1004140.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., and Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80-91.

Zipkin, E. F., and Saunders, S. P. (2018). Synthesizing multiple data types for biological conservation using integrated population models. *Biological Conservation*, *217*, 240-250.

CHAPTER ONE

HIERARCHICAL COMPUTING FOR HIERARCHICAL MODELS IN ECOLOGY[1]

Introduction

Ecological systems are characterized by dynamics and uncertainty at many scales, but observing all relevant scales may be difficult or impossible (Wiens 1989). Instead, we must use models to scale and connect processes across multiple levels (Levin 1992), such as from the scale of observation to the hypothesized scale of biological process, or from a single individual or species to a population or community. For example, in movement ecology, we often collect telemetry data and observe movement at the individual-level, but wish to make inference on the population as a whole, like to better understand responses to environmental conditions that are similar among individuals (Hooten et al. 2016). Alternatively, modeling ecosystems or ecological communities often involves joint analysis of many taxonomic groups as well as the processes that connect them (Levin 1992, Warton et al. 2015). Finally, conducting ecological studies introduces additional uncertainty, including sampling and detection uncertainty as well as spatial and temporal variation between study sites and years, which must be considered when specifying ecological models (Royle and Dorazio 2008, Beissinger et al. 2016).

Bayesian hierarchical modeling has become a popular tool in ecology, facilitating scaling by relating process models at one level to parameters at another level (Royle and Dorazio 2008, Hobbs and Hooten 2015). Hierarchical models are flexible and facilitate the inclusion of multiple

[1] Reprinted from McCaslin, H. M., Feuka, A. B., & Hooten, M. B. (2021). Hierarchical computing for hierarchical models in ecology. *Methods in Ecology and Evolution*, *12*(2), 245-254. https://doi.org/10.1111/2041-210X.13513

sources of uncertainty in the data, process, and parameter components (Beliner 1996, Cressie et al. 2009). For example, many integrated population models (IPMs) use a Bayesian hierarchical framework to integrate multiple data sources to understand population dynamics and demographic processes (Schaub and Abadi 2011). However, IPMs and other hierarchical models can quickly become large and time-consuming to fit.

Ecological science has seen a rapid increase in the availability of big data, advanced statistical techniques, and collaborative research, and our ability to specify ecological models that capture more of the complexity of natural phenomena has improved substantially as a result (McCallen et al. 2019). However, many ecologists have also reached the point where computational demands limit what can be modeled. Further, as ecologists are increasingly interested in long-term monitoring and prediction (Dietze et al. 2008), statistical models must be fit each time data are added. Collaborations with computer and data scientists and new software packages for efficient computing have introduced sophisticated computational techniques (e.g., distributed computing) in ecological science, but barriers to wide implementation of these approaches are a bottleneck for advancing ecological modeling (Visser et al. 2015, Hampton et al. 2017). Therefore, more accessible approaches for reducing computational limitations are needed to support progress in ecological modeling and understanding.

Recursive computing techniques, also known as batch or modular computing or Bayesian filtering, are used to fit a statistical model in a series of steps (Särkä 2013). These techniques simplify computing at each step, without modifying the original model specification or resulting inference. One recursive Bayesian computing (RB) method, introduced by Lunn et al. (2013), leverages the properties of Markov Chain Monte Carlo (MCMC) sampling (Gelfand and Smith 1990) to lessen the computational burden of fitting hierarchical models. The authors used RB to

reconcile the results of several independent studies in a meta-analysis (Lunn et al. 2013), and the method has been applied in ecological contexts to facilitate online updating (Hooten et al. 2020), model individual and group variation in physiological measurements (Hooten and Hefley 2019), and scale movement and resource-selection models from individuals to populations (Hooten et al. 2016, Gerber et al. 2018). While not unique to ecology, RB is a natural computational technique for ecologists to consider because the RB framework mirrors many ecological study designs and hierarchical models.

Consider a study of invasive cheatgrass (*Bromus tectorum*) occurrence in grasslands in Montana, in the northwestern United States (Pearson et al. 2018). Cheatgrass occurrence was monitored at 20 grassland sites by sampling 20 randomly selected 1-m^2 plots within each site. Suppose we want to model the probability of cheatgrass occurrence y_{ij} in Montana grasslands using a Bernoulli generalized linear mixed model (GLMM) specified as

$$\begin{aligned} y_{ij} &\sim \text{Bern}(p_j), \qquad i = 1,\dots,N, j = 1,\dots,J, && (1)\\ \text{logit}(p_j) &\sim \text{N}(\mu, \sigma^2), && (2)\\ \mu &\sim \text{N}(\mu_0, \sigma_0^2), && (3)\\ \sigma^2 &\sim \text{IG}(q, r), && (4) \end{aligned}$$

where j indexes sites and i indexes plots within each site. In this model, p_j is the probability of cheatgrass at site j, and $\text{logit}(p_j)$ arises from a Gaussian distribution with study-wide parameters μ and σ^2, arising from Gaussian and inverse gamma distributions, respectively (Fig. 1.1). Thus, p_j are “random effects" because they will vary for each site but will arise from a single underlying distribution. We use Gaussian random effects, with the logit link function to constrain p_j to the proper support, and seek inference on μ. The full-conditional distributions for the $\text{logit}(p_j)$ are not analytically tractable, so the $\text{logit}(p_j)$ cannot be sampled using Gibbs updates

and will need to be tuned individually to fit the model (Gelfand and Smith 1990). This minimal example could be fit in a single, conventional MCMC algorithm, but we describe the procedure to fit it recursively to demonstrate RB methods.

We could fit this model using RB by first partitioning the data by site, $\mathbf{Y} = (\mathbf{y'_1}, \ldots, \mathbf{y'_J})'$. These individual partitions would be analyzed independently in a first-stage MCMC algorithm with a temporary prior for logit(p_j) to obtain temporary posterior distributions for the parameters logit(p_j). Then, the resulting temporary posterior distributions would be used as proposals in the second-stage algorithm to update the study-wide parameters μ and σ^2, and the logit(p_j) given μ and σ^2 (Lunn et al. 2013). However, we would still need to tune the updates for each logit(p_j) by hand in the first stage, because the full-conditional distributions are not analytically tractable. This would slow model fitting and may be difficult.

Instead, we propose a modification of RB, which we call transformation-assisted RB (TARB), to eliminate tuning in the first stage and ease model fitting with unsupervised algorithms and efficient Gibbs updates. In what follows, we demonstrate how to implement RB and TARB to fit ecological models and apply TARB to a hierarchical movement model for avian migration to make individual- and population-level inference. Additionally, we discuss the implementation of TARB to other ecological models to illustrate its wide applicability.

Methods

Our Bernoulli GLMM is a hierarchical model comprised of data, process, and parameter components (Berliner 1996), with a set of latent random effects $\boldsymbol{\theta}_j = \text{logit}(p_j)$ for $j = 1, \ldots, J$ (Fig. 1.1). The group-level parameters $\boldsymbol{\psi} = (\mu, \sigma^2)'$, which correspond to the full study area in

our example, describe the distribution underlying the partition-level (e.g., site-level) parameters $\boldsymbol{\theta}_j$. For data partitioned $\mathbf{Y} = \left(\mathbf{y_1}', \ldots, \mathbf{y_J}'\right)'$, this can be written

$$\mathbf{y_j} \sim \left[\mathbf{y_j}|\boldsymbol{\theta}_j\right], \qquad j = 1, \ldots, J, \tag{5}$$

$$\boldsymbol{\theta}_j \sim \left[\boldsymbol{\theta_j}|\boldsymbol{\psi}\right], \tag{6}$$

$$\boldsymbol{\psi} \sim [\boldsymbol{\psi}]. \tag{7}$$

Note that square brackets $[\cdot]$ denote probability distributions (Gelfand and Smith 1990).

In general, $\boldsymbol{\theta}_j$ could be an $m \times 1$ vector that describes the partition-level process with m covariates. The data partitions $\mathbf{y_j}$ do not need to be equal-sized, and can represent any natural data subset such as different field sites as in our example, telemetry fixes for distinct individuals, results from several studies in a meta-analysis, or data on different species in a community, as long as dependence within the data partitions is accounted for in the data or process models.

The RB approach presented by Lunn et al. (2013) is carried out by specifying prior distributions $\left[\boldsymbol{\theta}_j\right]$ in the first-stage to obtain a sample from the posterior distributions $\left[\boldsymbol{\theta}_j|\mathrm{y}_j\right] \propto \left[\mathbf{y_j}|\boldsymbol{\theta}_j\right]\left[\boldsymbol{\theta}_j\right]$ for each partition $j = 1, \ldots, J$ independently. Next, the hierarchical model in (5)-(7) is fit using a second-stage MCMC algorithm with Metropolis-Hastings (MH) updates for $\boldsymbol{\theta}_j$, in which random samples from the temporary, first-stage posterior distributions for $\boldsymbol{\theta}_j$ are used as the proposals $\boldsymbol{\theta}_j^{(*)}$. This eliminates the need for tuning in the second-stage MH updates. Also in the second-stage algorithm, the group-level parameters $\boldsymbol{\psi}$ are updated based on their full-conditional distributions $[\boldsymbol{\psi}|\cdot] \propto \left(\prod_{j=1}^{J}\left[\boldsymbol{\theta}_j|\boldsymbol{\psi}\right]\right)[\boldsymbol{\psi}]$. The MH acceptance probability for each $\boldsymbol{\theta}_j^{(*)}$ is $\min\left(r_j^{(*)}, 1\right)$ where

$$r_j^{(k)} = \frac{\left[\mathbf{y_j}|\boldsymbol{\theta}_j^{(*)}\right]\left[\boldsymbol{\theta}_j^{(*)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}|\mathbf{y_j}\right]}{\left[\mathbf{y}_j|\boldsymbol{\theta}_j^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(*)}|\mathbf{y_j}\right]}, \tag{8}$$

$$= \frac{\left[\mathbf{y_j}|\boldsymbol{\theta}_j^{(*)}\right]\left[\boldsymbol{\theta}_j^{(*)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\mathbf{y}_j|\boldsymbol{\theta}_j^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}\right]}{\left[\mathbf{y_j}|\boldsymbol{\theta}_j^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\mathbf{y}_j|\boldsymbol{\theta}_j^{(*)}\right]\left[\boldsymbol{\theta}_j^{(*)}\right]}, \tag{9}$$

$$= \frac{\left[\boldsymbol{\theta}_j^{(*)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}\right]}{\left[\boldsymbol{\theta}_j^{(k-1)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(*)}\right]}, \tag{10}$$

for MCMC iteration $k = 1, \ldots, K$. Notably, neither the MH ratio (10) nor the full-conditional distributions for $\boldsymbol{\psi}$ involve the data $\mathbf{y}$. For the data model to cancel in the numerator and denominator of the MH ratio (10), the proposals $\boldsymbol{\theta}_j^{(*)}$ should be independent draws from the first-stage posterior distributions for $\boldsymbol{\theta}_j$. Thus, in practice, we sample $\boldsymbol{\theta}_j^{(*)}$ randomly with replacement from the first-stage Markov chains so that the samples are uncorrelated (Lunn et al. 2013, Hooten et al. 2020).

If the hierarchical model is specified such that the conditional distributions for $\boldsymbol{\theta}_j$ are not analytically tractable, like in our GLMM, then the first stage of the model must be fit using MH or importance sampling (Geweke 1989) which must be tuned by the user for each partition (Hooten et al. 2016). Thus, rather than specifying a first-stage prior directly on $\boldsymbol{\theta}_j$, we use TARB and specify a prior $\left[\mathbf{g}\left(\boldsymbol{\theta}_j\right)\right]$ on a transformation $\mathbf{g}(\boldsymbol{\theta}_j)$ of the parameters $\boldsymbol{\theta}_j$. It is most advantageous to specify $\mathbf{g}$ so that the first-stage priors on $\mathbf{g}\left(\boldsymbol{\theta}_j\right)$ are conjugate with the data model to allow us to use an automated Gibbs sampler in the first stage. In GLMMs and other hierarchical models, we often specify models so that parameters and random effects arise from Gaussian distributions, and use a link function to constrain these parameters to the appropriate support. Thus, in these cases, $\mathbf{g}$ will likely be a back-transformation (i.e., the inverse of the link

function) that allows us to specify conjugate first-stage priors. However, unlike if we were to specify a different model to facilitate conjugacy, using TARB allows us to incorporate prior knowledge and obtain inference in terms of the original model specification. For example, if we let $\mathbf{g}(\boldsymbol{\theta}_j) = \text{logit}^{-1}(\boldsymbol{\theta}_j)$ in our cheatgrass example, then we can specify a temporary beta prior on p_j in the first-stage. In this example, the benefit of doing so extends beyond conjugacy to a first-stage posterior distribution that can be written analytically, and therefore does not require MCMC to sample. We provide the complete procedure to fit the cheatgrass GLMM using TARB, with code, in the Supporting Information (Appendix A).

We need to use the resulting first-stage posterior distribution as a proposal distribution in the second-stage MCMC algorithm, but the first stage posterior distribution $\left[\mathbf{g}(\boldsymbol{\theta}_j)|\mathbf{y}_j\right]$ is on the transformed parameters $\mathbf{g}(\boldsymbol{\theta}_j)$. Thus, to account for the first-stage prior on transformed parameters, we must modify the MH ratio (10) and use a change of variables technique to ensure the proposal is on the same transformation that appears in the process component (6) of the original hierarchical model. While we could easily use the first-stage posterior distribution to obtain a *sample* from the desired posterior distribution $\left[\boldsymbol{\theta}_j|\mathbf{y}_j\right]$, the MH ratio requires us to evaluate the probability density function $\left[\boldsymbol{\theta}_j|\mathbf{y}_j\right]$ rather than sample from it. There are many possible methods for obtaining this distribution, including analytical change of variable techniques and numerical approaches. For continuous random variables, we use a change of variables technique where

$$\left[\boldsymbol{\theta}_j|\mathbf{y}_j\right] = [\mathbf{g}(\boldsymbol{\theta}_j)|\mathbf{y}_j]\left|\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)\right)\right|, \tag{11}$$

in which $\mathbf{J}(\mathbf{g}(\boldsymbol{\theta}_j))$ is the Jacobian matrix defined as

$$\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)\right) \equiv \begin{bmatrix} \frac{\delta g_1(\boldsymbol{\theta}_j)}{\delta \theta_{j,1}} & \cdots & \frac{\delta g_1(\boldsymbol{\theta}_j)}{\delta \theta_{j,p}} \\ \vdots & \ddots & \vdots \\ \frac{\delta g_{p_g}(\boldsymbol{\theta}_j)}{\delta \theta_{j,1}} & \cdots & \frac{\delta g_{p_g}(\boldsymbol{\theta}_j)}{\delta \theta_{j,p}} \end{bmatrix}. \quad (12)$$

The Jacobian matrix consists of partial derivatives of each element of $\mathbf{g}(\boldsymbol{\theta}_j)$ with respect to each element of $\boldsymbol{\theta}_j$. Its determinant $\left|\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)\right)\right|$ maps the change in the transformed variables to the change in the non-transformed variables ($d\mathbf{g}(\boldsymbol{\theta}_j)$ onto $d\boldsymbol{\theta}_j$), yielding the correct probability distribution of the non-transformed variable when multiplied to the probability distribution of the transformed variable. Thus, substituting (11) for the proposal in the second-stage MH ratio (10) results in

$$r_j^{(k)} = \frac{\left[\mathbf{y}_j|\boldsymbol{\theta}_j^{(*)}\right]\left[\boldsymbol{\theta}_j^{(*)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}|\mathbf{y}_j\right]}{\left[\mathbf{y}_j|\boldsymbol{\theta}_j^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(k-1)}|\boldsymbol{\psi}^{(k-1)}\right]\left[\boldsymbol{\theta}_j^{(*)}|\mathbf{y}_j\right]}, \quad (13)$$

$$= \frac{[\boldsymbol{\theta}_j^{(*)}\left|\boldsymbol{\psi}^{(k-1)}\right]\left[\mathbf{g}(\boldsymbol{\theta}_j)^{(k-1)}\right]\left|\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)^{(k-1)}\right)\right|}{[\boldsymbol{\theta}_j^{(k-1)}\left|\boldsymbol{\psi}^{(k-1)}\right]\left[\mathbf{g}(\boldsymbol{\theta}_j)^{(*)}\right]\left|\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)^{(*)}\right)\right|}. \quad (14)$$

The data component of the hierarchical model cancels in the MH ratio (14) associated with the second-stage MCMC algorithm regardless of the transformation used in the first-stage temporary prior, and we account for the transformation via the determinant of the Jacobian in the modified TARB ratio (14). In our cheatgrass GLMM, because $\boldsymbol{\theta}_j = p_j$ is a scalar, the Jacobian simplifies to the derivative of $g = \text{logit}^{-1}(p_j)$ with respect to $\text{logit}(p_j)$ (Appendix A). Thus, we can use TARB to create unsupervised first-stage algorithms that can be easily parallelized and a second-stage MCMC algorithm that does not rely on the data model. This results in substantial

computational savings when the data model is complex or there are many data models to fit and allows the second stage to be updated easily if new data partitions become available.

Application: White Stork Migration

To demonstrate TARB, we developed a hierarchical animal movement model for the migratory behavior of white storks (*Ciconia ciconia*) in western Europe to obtain individual- and population-level inference for migration characteristics. We analyzed data from $J = 15$ individuals tracked with GPS units from 30 July 2018 – 29 Sept 2018 (Fig. 1.2, Cheng et al. 2019, Fiedler et al. 2019). These data are available in the R package 'moveVis' (Schwalb-Willmann et al. 2020).

Model statement

We specified a continuous-time hierarchical model for stork movement with the data component

$$\mathbf{s_j}(t_i) \quad \sim \mathrm{N}\big(\mathbf{s_j}(t_{i-1}) - \nabla p\big(\mathbf{s_j}(t_i), \boldsymbol{\beta}_j\big)dt_i, \sigma_j^2 dt_i \mathbf{I}\big), \tag{15}$$

where $\mathbf{s_j}(t_i)$ is the measured position of individual j at time i (for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$). We defined the potential function in (15) as $p\big(\mathbf{s}, \boldsymbol{\beta}_j\big) \equiv \mathbf{x}'(\mathbf{s})\boldsymbol{\beta}_j$, which describes a surface upon which an individual is more likely to move "downhill" (Brillinger 2010, Hooten et al. 2017). In our specification, this surface is a linear function of covariates $\mathbf{x}(\mathbf{s})$ and will influence the speed and directional persistence of movement. The term dt_i represents the change in time between successive positions $\mathbf{s_j}(t_{i-1})$ and $\mathbf{s_j}(t_i)$, and $\mathbf{I}$ is the 2×2 identity matrix. The statistical model in (15) converges to the stochastic differential equation (SDE)

$$\mathbf{s_j}(t) \quad = -\nabla p\big(\mathbf{s_j}(t), \boldsymbol{\beta}_j\big)dt + \sigma_j d\mathbf{b_j}(t), \tag{16}$$

as $dt \to 0$, where $d\mathbf{b_j}(t)$ is bivariate Gaussian white noise.

In the data model (15), the parameters σ_j^2 relate to the speed of the migrating individuals and will vary around a group-level speed. However, due to the positive support of the variance components σ_j^2, we chose to model the individual-level process relating to migration speed in the transformation $\log(\sigma_j)$, so that the support is unbounded and can be suitably modeled with a Gaussian distribution. Otherwise, to create Gibbs updates for σ_j^2 directly in a single-stage algorithm, we would need to specify a conjugate inverse gamma process model on σ_j^2, and specifying hyperpriors on the associated shape and scale parameters would be neither trivial nor biologically intuitive. Thus, we specified a process model for $\log(\sigma_j)$ instead of σ_j^2, implying the transformation function $\sigma_j^2 = \mathbf{g}\left(\log(\sigma_j)\right) = e^{2\log(\sigma_j)}$.

In our example, we expected migration to occur primarily in a single direction and specified $\mathbf{x(s)} = s_2$ where the second component of position **s** corresponds to latitude and the coefficient vector is comprised of a single parameter β. Thus, the negative gradient of the potential function in (15) simplifies to $-\nabla p(\mathbf{s_j}(t), \boldsymbol{\beta}_j) = -(0, \beta_j)'$. However, this simplification is based on the assumption that all individuals will migrate in a north/south orientation. To allow for individual variation in the bearing, we multiplied the potential function in (15) by the rotation matrix

$$\mathrm{M} \equiv \begin{pmatrix} \cos(\phi_j) & -\sin(\phi_j) \\ \sin(\phi_j) & \cos(\phi_j) \end{pmatrix}, \tag{17}$$

where ϕ_j is the angle from south of a migratory path, resulting in the data model,

$$\mathbf{s_j}(t_i) \sim \mathrm{N}\left(\mathbf{s_j}(t_{i-1}) - \beta_j \begin{pmatrix}\sin(\phi_j)\\ \cos(\phi_j)\end{pmatrix} dt_i,\ \sigma_j^2 dt_i \mathbf{I}\right), \tag{18}$$

Assuming that the variability in β_j and $\log(\sigma_j)$ across individuals can be accounted for as Gaussian random effects and that individual variability in ϕ_j does not arise from an underlying group-level distribution, we have $\beta_j \sim \mathrm{N}(\mu_\beta, \sigma_\beta^2)$, $\log(\sigma_j) \sim \mathrm{N}(\mu_\sigma, \sigma_\sigma^2)$, and $\phi_j \sim \mathrm{Unif}(0, \pi)$, where population-level means μ_β and μ_σ are modeled with Gaussian priors and σ_β^2 and σ_σ^2 arise from inverse gamma priors (full model in Supporting Information, Appendix B).

Two-stage implementation

We fit our model to a subset of the stork migration data (approximately two observations per day per individual) using TARB. In the first stage, we specified individual-level models using the temporary prior $[\beta_j, \sigma_j^2] = [\beta_j][\sigma_j^2]$ where $\beta_j \sim [\beta_j] \equiv \mathrm{N}(\mu_0, \sigma_0^2)$ and $\sigma_j^2 \sim [\sigma_j^2] \equiv \mathrm{IG}(q_0, r_0)$ for $j = 1, \ldots, J$. Thus, in the first stage, we sample from the posterior distribution,

$$[\beta_j, \sigma_j^2, \phi_j | \mathbf{S_j}] \propto \prod_{i=2}^{n_j} [\mathbf{s_j}(t_i) | \beta_j, \sigma_j^2, \phi_j]\, [\beta_j][\sigma_j^2][\phi_j], \tag{19}$$

for each individual $j = 1, \ldots, J$. We sampled sequentially from the conjugate full-conditional distributions $[\beta_j | \cdot]$ and $[\sigma_j^2 | \cdot]$ using Gibbs updates and from $[\phi_j | \cdot]$ using a MH update in an MCMC algorithm in R (version 3.6.1) that we parallelized over individuals with the 'parallel' package (R Core Team 2019).

To use samples from the first-stage models as proposals in the second-stage algorithm, we calculated the Jacobian determinant in (14). Letting $\boldsymbol{\theta}_j \equiv \left(\beta_j, \log(\sigma_j)\right)'$, and the 2×1 vector

transformation $\mathbf{g}(\boldsymbol{\theta}_j)$ be comprised of components $g_1(\boldsymbol{\theta}_j) = \beta_j$ and $g_2(\boldsymbol{\theta}_j) = e^{2\log(\sigma_j)}$, we calculated the Jacobian

$$\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)\right) \equiv \begin{bmatrix} \frac{\delta g_1(\boldsymbol{\theta}_j)}{\delta \beta_j} & \frac{\delta g_1(\boldsymbol{\theta}_j)}{\delta \log(\sigma_j)} \\ \frac{\delta g_2(\boldsymbol{\theta}_j)}{\delta \beta_j} & \frac{\delta g_2(\boldsymbol{\theta}_j)}{\delta \log(\sigma_j)} \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & 2\sigma_j^2 \end{bmatrix}, \quad (20)$$

which has the determinant $\left|\mathbf{J}\left(\mathbf{g}(\boldsymbol{\theta}_j)\right)\right| = 2\sigma_j^2$. Thus, the second-stage MH ratio from (14) to update β_j, $\log(\sigma_j)$, and ψ_j for individual *j* is

$$r_j^{(k)} = \frac{\left[\beta_j^{(*)} | \mu_\beta^{(k-1)}, \sigma_\beta^{2(k-1)}\right]\left[\log\left(\sigma_j^{(*)}\right) | \mu_\sigma^{(k-1)}, \sigma_\sigma^{2(k-1)}\right]\left[\beta_j^{(k-1)}\right]\left[\sigma_j^{2(k-1)}\right]\left[\phi_j^{(*)}\right] \times \sigma_j^{2(k-1)}}{\left[\beta_j^{(k-1)} | \mu_\beta^{(k-1)}, \sigma_\beta^{2(k-1)}\right]\left[\log\left(\sigma_j^{(k-1)}\right) | \mu_\sigma^{(k-1)}, \sigma_\sigma^{2(k-1)}\right]\left[\beta_j^{(*)}\right]\left[\sigma_j^{2(*)}\right]\left[\phi_j^{(k-1)}\right] \times \sigma_j^{2(*)}}. \quad (21)$$

The scalar multiple of 2 from the Jacobian determinant cancels in the numerator and denominator of (21). In the second-stage algorithm, we used the MH ratio in (21) to accept our proposals for $\beta_j^{(*)}$, $\log\left(\sigma_j^{(*)}\right)$, and $\phi_j^{(*)}$ which we sampled jointly at random (with replacement) from our first-stage MCMC sample. Then, we sampled the group-level model parameters ($\mu_\beta, \sigma_\beta^2, \mu_\sigma$, and σ_σ^2) sequentially from their full-conditional distributions using Gibbs updates (Appendix B).

Alternatively, it is possible to fit the full hierarchical model using a standard MCMC algorithm with Gibbs updates for $\beta_j, \mu_\beta, \sigma_\beta^2, \mu_\sigma$, and σ_σ^2. However, we would need to use MH updates for $\log(\sigma_j)$ and ϕ_j, and in cases where the number of individuals *J* is large, we may have to tune a prohibitively large number of proposal distributions to yield optimal acceptance rates in the MCMC algorithm. Nonetheless, to demonstrate that we obtain the same inference with

TARB as compared to a single MCMC algorithm, we also fit the full model with a single algorithm, updating β_j and $\log(\sigma_j)$ sequentially for each individual with Gibbs and MH updates, respectively, and the remaining model parameters as above.

Results

We fit our movement model to a subset of $n = 1675$ stork telemetry observations across $J = 15$ individuals using TARB with $K = 100{,}000$ MCMC iterations for each stage, computing the first stage in parallel over 8 cores, and using a single hierarchical MCMC algorithm with $K = 100{,}000$ MCMC iterations. The recursive approach required 2.95 minutes and the single algorithm required 9.87 minutes; thus computation was over three times faster using TARB. With a larger data set of $n = 155{,}161$ locations for 15 individuals and $K = 60{,}000$ MCMC iterations, computation time to fit the model recursively, in parallel over 15 cores, was 49 minutes, compared to 88 minutes to fit the model as a single algorithm.

Both computational approaches resulted in the same 95% credible intervals and posterior means for β_j and $\log(\sigma_j)$ and the same population-level means μ_β and μ_σ (Fig. 1.2). The stage-two posterior credible intervals for the β_j and $\log(\sigma_j)$ for each individual j indicate individual variation in speed and directional persistence of migration, but the population is centered around μ_β and μ_σ. First-stage credible intervals are included only to visualize the relationship between stage one and stage two in Figure 1.2, and are not used for inference. The shrinkage in interval width between the first- and second-stage posteriors of β_j and $\log(\sigma_j)$ indicates individual-level inference was informed by group-level parameters in the second stage, although this effect was relatively minor in this example. Further, fitting the model to simulated data shows that both computational approaches do equally well recovering 'true' simulated parameters (Appendix C).

Discussion

In our application, we illustrated how TARB can be used to efficiently fit a hierarchical animal movement model to telemetry data, but TARB could be implemented in many ecological models to improve computational efficiency. In Table 1.1, we highlight several studies from the ecological literature in which the authors used a Bayesian hierarchical model (or desired to, barring computational limitations, as in Breed et al. 2009) that could be fit with TARB. To demonstrate the application of TARB to existing ecological models, we discuss two examples in detail, outlining how the models can be specified in the two-stage framework for faster computation.

Harbor Seal Counts

Cressie et al. (2009) specified a Bayesian hierarchical model to explicitly account for uncertainty at the data and process levels while estimating abundance of harbor seals (*Phoca vitulina*) from census data (Ver Hoef and Frost 2003) in Prince William Sound

$$
\begin{aligned}
y_{ij} &\sim \text{Pois}(\lambda_{ij}), && (22)\\
\log(\lambda_{ij}) &\sim \text{N}(\mu_{ij}, \sigma_{ij}^2), && (23)\\
\mu_{ij} &= \theta_{0,j} + \mathbf{x_{ij}}'\boldsymbol{\theta}_j, && (24)\\
\boldsymbol{\theta}_j &\sim \text{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}), && (25)
\end{aligned}
$$

where y_{ij} is the number of hauled-out seals counted from photographs during each aerial survey i conducted at site j. In the observation model (22), counts arise from a Poisson distribution with intensity parameter λ_{ij} that represents the expected number of haul-outs in a given survey and location. The expected number of haul-outs (λ_{ij}) arises from a normal distribution with mean μ_{ij} that is a function of covariates $\mathbf{x_{ij}}$ with variance parameters σ_{ij}^2 for each survey and location.

Site-level coefficients $\boldsymbol{\theta}_j$ arise from a population-level multivariate Gaussian distribution, where $\boldsymbol{\Sigma}$ is a diagonal matrix with population-level variance parameters along the diagonal. Thus, the hierarchical model in (22)-(25) is a special case of a generalized linear mixed model.

Surveys were conducted several times per year at each site. Thus, in the first stage of the TARB framework, counts could be modeled independently for each site with the model

$$y_{ij} \sim \text{Pois}(\lambda_{ij}), \tag{26}$$
$$\lambda_{ij} \sim \text{Gamma}(\alpha, \beta), \tag{27}$$

where a temporary gamma prior on λ_{ij} is conjugate with the data model (22) in the first stage so that the MCMC algorithm is unsupervised and could be parallelized over the sites. To complete model fitting in stage two, log-transformed first-stage samples for λ_{ij} would be used as proposals in the MH update for $\log(\lambda_{ij})$ in a second-stage algorithm,

$$\left[\log(\lambda_{ij})|\cdot\right] = \frac{\left[\log\left(\lambda_{ij}^{(*)}\right)|\mu_{ij}, \sigma_{ij}^2\right]\left[\lambda_{ij}^{(k-1)}|\alpha, \beta\right]\lambda_{ij}^{(k-1)}}{\left[\log\left(\lambda_{ij}^{(k-1)}\right)|\mu_{ij}, \sigma_{ij}^2\right]\left[\lambda_{ij}^{(*)}|\alpha, \beta\right]\lambda_{ij}^{(*)}}, \tag{28}$$

where $\frac{d}{d\log(\lambda_{ij})} e^{\log(\lambda_{ij})} = \lambda_{ij}$. All other parameters in the second stage would be updated in the same manner as in a conventional algorithm.

Host Plant Genetics

Evans et al. (2012) conducted a common garden experiment to determine the effects of cottonwood host (*Populus* spp.) genotype on the abundance of herbivorous mite (*Aceria parapopuli*) galls on trees. In our notation, their model was

$$
\begin{aligned}
y_{imt} &\sim \text{Pois}(\theta_{imt}), && (29)\\
\log(\theta_{imt}) &\sim \text{N}(\mu_{imt}, \sigma^2), && (30)\\
\mu_{imt} &= \beta_i + \mathbf{x_{tm}}'\boldsymbol{\alpha}, && (31)\\
\boldsymbol{\alpha} &\sim \text{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha), && (32)\\
\beta_i &\sim \text{N}(0, \tau^2), && (33)\\
\tau^2 &\sim \text{IG}(\text{a}_\tau, \text{b}_\tau), && (34)\\
\sigma^2 &\sim \text{IG}(\text{a}_\sigma, \text{b}_\sigma), && (35)
\end{aligned}
$$

where y_{imt} is the number of galls on tree i with genotype m in year t. The intensity parameter θ_{imt} is a log-linear function of fixed effects $\boldsymbol{\alpha}$ for year and genotype and random effect of tree, β_i. Modifying the process model to

$$\theta_{imt} \sim \text{Gamma}(\gamma_1, \gamma_2), \quad (36)$$

and using temporary priors on γ_1 and γ_2 results in an unsupervised first-stage algorithm. We make a similar adjustment to the second-stage MH ratio as in (28) for recursive computation.

Conclusion

Transformation-assisted RB is one of the most accessible approaches for fitting ecological models recursively with improved computational efficiency and ease. Transformation allows us to extend the benefits of RB to more model specifications, and the demonstrated approach with change of variables can be implemented for most continuous random variables. The ability to incorporate prior information into analyses is a well-known feature of Bayesian analysis, but it can be difficult to determine how to do so in a robust way, and TARB is a natural approach for using posterior estimates from a previous study as prior information in subsequent studies. Finally, TARB leverages the parallel computing capacity of modern multi-core computers (Visser et al. 2015) to reduce the computational bottleneck created by large data sets and conventional sampling techniques.

Decreased computation time is a major advantage of fitting hierarchical models using TARB, but reducing tuning and partitioning the data in the first stage are equally, if not more, advantageous. This is especially true for large hierarchical models where one might otherwise have to individually tune dozens or hundreds of individual-level parameters to achieve convergence, which would require repeatedly fitting the model. Further, because the first-stage algorithm is used to fit data partitions independently and the second-stage algorithm does not rely on the data directly, we expect additional computational gains. Finally, by design, TARB accommodates uneven sample sizes of partitions, because the first-stage posterior distributions will reflect the uncertainty associated with different sample sizes, thus implicitly weighting the partitions according to sample size in the second stage.

In many cases, the first-stage algorithms of RB and TARB approaches could be implemented in an existing package like JAGS, Stan, or NIMBLE (Plummer 2003, Stan Development Team 2018, NIMBLE Development Team 2019), but the second-stage algorithm cannot be easily implemented in this software. However, using TARB, it may be possible to fit models that are not feasible using these software packages at all. While automated software is convenient and well-suited to a wide range of models, it cannot accommodate all model specifications and users do not always have control over tuning. Although software packages can often fit large models quickly, this may be achieved via computation in C++ rather than R (e.g., Stan, Stan Development Team 2018) or by making approximate inference (e.g., INLA, Rue et al. 2009). Recursive techniques like TARB can also be implemented in C++ via R and *rcpp* for greater computational efficiency, and the results can be used to obtain both marginal and joint inference.

While TARB can be implemented for a broad range of hierarchical models, there are some cases for which TARB, as presented here using the Jacobian to perform a change of variables, is not ideal for model fitting. For example, hierarchical models that have common parameters at the data level, in addition to partition-level parameters, such as GLMMs with both fixed and random effects, are not easily implemented using TARB. In this case, prior-proposal RB may be helpful (Hooten et al. 2020). Additionally, the Jacobian approach for computing transformed densities is well-suited for transforming continuous random variables, but alternate approaches must be used for discrete random variables. We demonstrated TARB using this technique because it serves as a good introduction into recursive techniques with transformation. For other random variables or applications, there are many useful generalizations of this approach that could be used to obtain valid transformations.

Hierarchical models are powerful tools for understanding complex ecological systems, but the computational demands of fitting ecologically realistic models can make them impractical or impossible to implement. Recursive Bayesian computing techniques address these computational demands, and partitioning model-fitting into stages is natural in many ecological applications. For example, in adaptive management, RB and TARB would allow managers to fit first-stage individual-, year-, or site-level models as data are collected, and add new partitions to existing results by subsequently updating the second stage. Additionally, because the second-stage algorithm only requires first-stage posterior samples, partitions could represent data collected by different researchers during ongoing projects, and researchers could fit population-wide models without needing to share data (Hooten et al. 2020). Thus, in the current era of big data and complex modeling in ecology, TARB is an approachable technique that reduces the computational limitations on the ecological models ecologists can specify and fit.

All supplementary information available at https://doi.org/10.1111/2041-210X.13513. The R code used in our analyses is available at https://doi.org/10.5281/zenodo.4075393, and the white stork data set is available on Movebank (Fiedler et al. 2019, https://doi.org/10.5441/001/1.v1cs4nn0).

Tables and Figures

Table 1.1. Examples of ecological studies with Bayesian hierarchical models that could be implemented in a transformation-assisted recursive Bayesian framework.

Discipline	Study
Fish and Wildlife Ecology	Burton et al. 2012 Cressie et al. 2009 Breininger et al. 2019 Kuhnert et al. 2005 Monroe et al. 2017 Moore and Barlow 2011
Integrated Population Models	Cleasby et al. 2017 Eacker et al. 2017 Raiho et al. 2015 Schaub et al. 2013
Animal Movement	Breed et al. 2009 Eckert et al. 2008 Jonsen et al. 2006 McClintock et al. 2013 Muff et al. 2019
Forestry and Plant Ecology	Dietze et al. 2008 Evans et al. 2012 Hanks et al. 2011 Iijima and Otsu 2018 Vieilledent et al. 2010
Ecosystem Ecology	Borsuk et al. 2001 Coll et al. 2019 Shelton et al. 2016

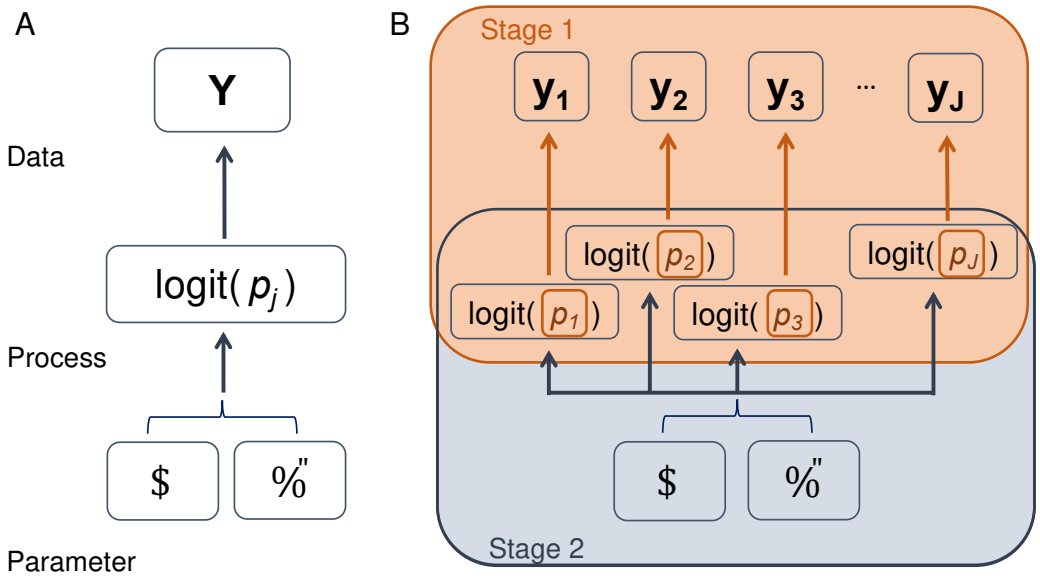


Figure 1.1. (A) Directed acyclic diagram (DAG) for Bernoulli GLMM of cheatgrass occurrence in Montana (1)-(4) and (B) schematic for partitioning DAG according to the TARB framework. In (A), **Y** is the matrix whose columns are the data vectors y_j for the sites $j = 1, \ldots, J$. In stage 1, the data **Y** are partitioned by site and fit to obtain the posterior distributions for the p_j. In stage 2, samples from these posterior distributions are used to sample logit(p_j), μ, and σ^2.

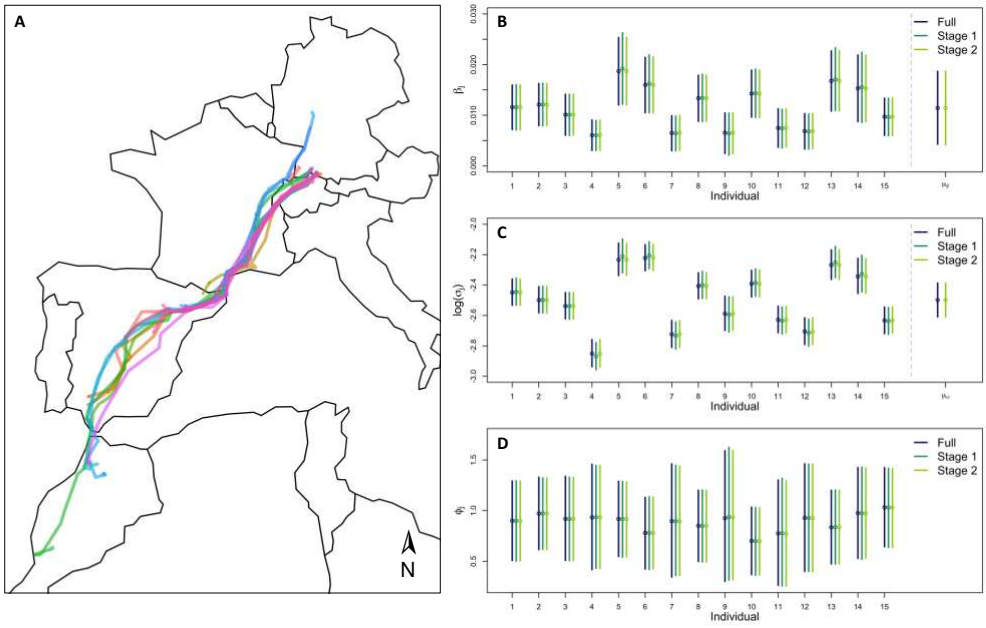


Figure 1.2. (A) Migratory trajectories for $J = 15$ white storks tracked via GPS loggers in fall 2018, with each individual represented by a different color, and (B)-(D) posterior means (points) and 95% credible intervals for model parameters resulting from fitting our hierarchical movement model to $n = 1675$ telemetry locations from $J = 15$ white storks as a single hierarchical algorithm and in two stages using TARB. It is important to note here that we show the posterior distributions for the first-stage estimates to illustrate how some individual-level parameters borrow strength from the group-level parameters in stage 2, but in practice, the first stage posterior estimates would not be used to make inference.

REFERENCES

Beissinger, S. R., K. J. Iknayan, G. Guillera-Arroita, E. F. Zipkin, R. M. Dorazio, J. A. Royle, and M. Kery. 2016. Incorporating imperfect detection into joint models of communities: A response to Warton et al. *Trends in Ecology and Evolution* 31: 736–737. doi: 10.1016/j.tree.2016.07.009.

Berliner, L. M. 1996. Hierarchical Bayesian time series models. Maximum Entropy and Bayesian Methods. Springer: 15–22.

Borsuk, M. E., D. Higdon, C. A. Stow, and K. H. Reckhow. 2001. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones. *Ecological Modelling* 143: 165–181. doi: 10.1016/S0304-3800(01) 00328-3.

Breed, G. A., I. D. Jonsen, R. A. Myers, W. D. Bowen, and M. L. Leonard. 2009. Sex-specific, seasonal foraging tactics of adult grey seals (*Halichoerus grypus*) revealed by state–space analysis. *Ecology* 90: 3209–3221. doi: 10.1890/07-1483.1.

Breininger, D. R., E. D. Stolen, D. J. Breininger, and R. D. Breininger. 2019. Sampling rare and elusive species: Florida east coast diamondback terrapin population abundance. *Ecosphere* 10 (8): e02824. doi: 10.1002/ecs2.2824.

Brillinger, D.R. 2010. Modeling spatial trajectories. Handbook of Spatial Statistics: 463–475.

Burton, A. C., M. K. Sam, C. Balangtaa, and J. S. Brashares. 2012. Hierarchical multi-species modeling of carnivore responses to hunting, habitat and prey in a West African protected area. *PloS One* 7. doi: 10.1371/journal.pone.0038007.

Cheng, Y., W. Fiedler, M. Wikelski, and A. Flack. 2019. “Closer-to-home” strategy benefits juvenile survival in a long-distance migratory bird. *Ecology and Evolution* 9: 8945–8952. doi: 10.1002/ece3.5395.

Cleasby, I. R., T. W. Bodey, F. Vigfusdottir, J. L. McDonald, G. McElwaine, K. Mackie, K. Colhoun, and S. Bearhop. 2017. Climatic conditions produce contrasting influences on demographic traits in a long-distance Arctic migrant. *Journal of Animal Ecology* 86: 285–295. doi: 10.1111/1365-2656.12623.

Coll, M., M. Pennino, J. Steenbeek, J. Sole, and J. Bellido. 2019. Predicting marine species distributions: Complementarity of food-web and Bayesian hierarchical modelling approaches. *Ecological Modelling* 405: 86–101. doi: 10.1016/j.ecolmodel.2019.05.005.

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical

statistical modeling. *Ecological Applications* 19: 553–570. url: https://doi.org/10.1890/07-0744.1.

Dietze, M. C., M. S. Wolosin, and J. S. Clark. 2008. Capturing diversity and interspecific variability in allometries: A hierarchical approach. *Forest Ecology and Management* 256: 1939–1948. doi: 10.1016/j.foreco.2008.07.034.

Eacker, D. R., P. M. Lukacs, K. M. Proffitt, and M. Hebblewhite. 2017. Assessing the importance of demographic parameters for population dynamics using Bayesian integrated population modeling. *Ecological Applications* 27: 1280–1293. doi: 10.1002/eap.1521.

Eckert, S. A., J. E. Moore, D. C. Dunn, R. S. van Buiten, K. L. Eckert, and P. N. Halpin. 2008. Modeling loggerhead turtle movement in the Mediterranean: Importance of body size and oceanography. *Ecological Applications* 18: 290–308. doi: 10.1890/06-2107.1.

Evans, L. M., J. S. Clark, A. V. Whipple, and T. G. Whitham. 2012. The relative influences of host plant genotype and yearly abiotic variability in determining herbivore abundance. *Oecologia* 168: 483–489. doi: 10.1007/s00442-011-2108-8.

Fiedler, W., A. Flack, W. Sch¨afle, B. Keeves, M. Quetting, B. Eid, H. Schmid, and M. Wikelski. 2019. Data from: Study “LifeTrack White Stork SW Germany”(2013-2019). doi: 10.5441/001/1.v1cs4nn0.

Gelfand, A. E. and A. F. Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85: 398–409. doi: 10.2307/2289776.

Gerber, B. D., M. B. Hooten, C. P. Peck, M. B. Rice, J. H. Gammonley, A. D. Apa, and A. J. Davis. 2018. Accounting for location uncertainty in azimuthal telemetry data improves ecological inference. *Movement Ecology* 6: 14. doi: 10.1186/s40462-018- 0129-1.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica: *Journal of the Econometric Society:* 1317–1339. doi: 10.2307/1913710.

Hampton, S. E., M. B. Jones, L. A. Wasser, M. P. Schildhauer, S. R. Supp, J. Brun, R. R. Hernandez, C. Boettiger, S. L. Collins, L. J. Gross, et al. 2017. Skills and knowledge for data-intensive environmental research. *BioScience* 67: 546–557. doi: 10.1093/biosci/bix025.

Hanks, E. M., M. B. Hooten, and F. A. Baker. 2011. Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. *Ecological Applications* 21: 1173–1188. doi: 10.1890/09-1549.1.

Hobbs, N. T. and M. B. Hooten. 2015. Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press.

Hooten, M. B., F. E. Buderman, B. M. Brost, E. M. Hanks, and J. S. Ivan. 2016. Hierarchical animal movement models for population-level inference. *Environmetrics* 27: 322–333. doi: 10.1002/env.2402.

Hooten, M. B. and T. J. Hefley. 2019. Bringing Bayesian Models to Life. CRC Press.

Hooten, M. B., D. S. Johnson, and B. M. Brost. 2020. Making recursive Bayesian inference accessible. *The American Statistician*: 1–10. doi: 10.1080/00031305.2019.1665584.

Hooten, M. B., D. S. Johnson, B. T. McClintock, and J. M. Morales. 2017. Animal Movement: Statistical Models for Telemetry Data. CRC press.

Iijima, H. and C. Otsu. 2018. The method of conserving herbaceous grassland specialists through silvicultural activities under deer browsing pressure. *Biodiversity and Conservation* 27: 2919–2930. doi: 10.1007/s10531-018-1577-z.

Jonsen, I. D., R. A. Myers, and M. C. James. 2006. Robust hierarchical state–space models reveal diel variation in travel rates of migrating leatherback turtles. *Journal of Animal Ecology* 75: 1046–1057. doi: 10.1111/j.1365-2656.2006.01129.x.

Kuhnert, P. M., T. G. Martin, K. Mengersen, and H. P. Possingham. 2005. Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environmetrics* 16: 717–747. doi: 10.1002/env.732.

Levin, S. A. 1992. The problem of pattern and scale in ecology: The Robert H. MacArthur award lecture. *Ecology* 73: 1943–1967. doi: 10.2307/1941447.

Lunn, D., J. Barrett, M. Sweeting, and S. Thompson. 2013. Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C* 62: 551–572. doi: 10.1111/rssc.12007.

McCallen, E., J. Knott, G. Nunez-Mir, B. Taylor, I. Jo, and S. Fei. 2019. Trends in ecology: Shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment* 17: 109–116. doi: 10.1002/fee.1993.

McCaslin, H. M., A. B. Feuka, and M. B. Hooten. 2020. Hierarchical computing for hierarchical models in ecology. *Methods in Ecology and Evolution* 12(2): 245-154.

McCaslin, H. M., A. B. Feuka, and M. B. Hooten. 2020. Release for Hierarchical computing for hierarchical models in ecology. Zenodo. doi: 10.5281/zenodo.4075393.

McClintock, B. T., D. J. Russell, J. Matthiopoulos, and R. King. 2013. Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. *Ecology* 94: 838–849. doi: 10.1890/12-0954.1.

Monroe, A. P., C. L. Aldridge, T. J. Assal, K. E. Veblen, D. A. Pyke, and M. L. Casazza. 2017. Patterns in greater sage-grouse population dynamics correspond with public grazing records at broad scales. *Ecological Applications* 27: 1096–1107. doi: 10.1002/ eap.1512.

Moore, J. E. and J. Barlow. 2011. Bayesian state-space model of fin whale abundance trends from a 1991–2008 time series of line-transect surveys in the California Current. *Journal of Applied Ecology* 48: 1195–1205. doi: 10.1111/j.1365-2664.2011.02018.x.

Muff, S., J. Signer, and J. Fieberg. 2019. Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using Bayesian or frequentist computation. *Journal of Animal Ecology* 89: 80–92. doi: 10.1111/1365-2656.13087.

NIMBLE Development Team. 2019. NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. Version 0.9.0. R package version 0.9.0. doi: 10.5281/zenodo. 1211190. url: https://cran.r-project.org/package=nimble.

Pearson, D.E.,O. Eren,Y.K. Ortega, D.Villarreal, M.Sentürk, M.F.Miguel, C.M.Weinzettel, A. Prina, and J. L. Hierro. 2018. Are exotic plants more abundant in the introduced versus native range? *Journal of Ecology* 106: 727–736. doi: 10.1111/1365-2745. 12881.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. url: https://www.R-project.org/.

Raiho, A. M., M. B. Hooten, S. Bates, and N. T. Hobbs. 2015. Forecasting the effects of fertility control on overabundant ungulates: White-tailed deer in the National Capital Region. *PLoS One* 10. doi: 10.1371/journal.pone.0143122.

Royle, J. A. and R. M. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities. Elsevier.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* 71: 319–392.

Särkkä, S. 2013. Bayesian filtering and smoothing. Vol. 3. Cambridge University Press.

Schaub, M. and F. Abadi. 2011. Integrated population models: A novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology* 152: 227–237. doi: 10.1007/s10336-010-0632-7.

Schaub, M., H. Jakober, and W. Stauber. 2013. Strong contribution of immigration to local population regulation: Evidence from a migratory passerine. *Ecology* 94: 1828–1838. doi: 10.1890/12-1395.1.

Schwalb-Willmann, J., R. Remelgado, K. Safi, and M. Wegmann. 2020. moveVis: Animating movement trajectories in synchronicity with static or temporally dynamic environmental data in R. *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.13374.

Shelton, A. O., J. L. O'Donnell, J. F. Samhouri, N. Lowell, G. D. Williams, and R. P. Kelly. 2016. A framework for inferring biological communities from environmental DNA. *Ecological Applications* 26: 1645–1659. doi: 10.1890/15-1733.1.

Stan Development Team. 2018. RStan: the R interface to Stan. R package version 2.17.3. url: http://mc-stan.org/.

Ver Hoef, J. and K. Frost. 2003. A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics* 10: 201–209. doi: 10.1023/A:1023626308538.

Vieilledent, G., B. Courbaud, G. Kunstler, J.-F. Dhôte, and J. S. Clark. 2010. Individual variability in tree allometry determines light resource allocation in forest ecosystems: a hierarchical Bayesian approach. *Oecologia* 163: 759–773. doi: 10.1007/s00442-010-1581-9.

Visser, M. D., S. M. McMahon, C. Merow, P. M. Dixon, S. Record, and E. Jongejans. 2015. Speeding up ecological and evolutionary computations in R: Essentials of high performance computing for biologists. *PLoS Computational Biology* 11. doi: 10.1371/journal. pcbi.1004140.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. Hui. 2015. So many variables: Joint modeling in community ecology. *Trends in Ecology and Evolution* 30: 766–779. doi: 10.1016/j.tree.2015.09.007.

Wiens, J. A. 1989. Spatial scaling in ecology. *Functional Ecology* 3: 385–397. doi: 10.2307/2389612.

CHAPTER TWO

HOTSPOT DESIGNATION AMPLIFIES SOCIAL SAMPLING BIAS IN CROWDSOURCED DATA COLLECTION PLATFORM EBIRD

Introduction

Crowdsourced voluntary science projects (also known as citizen science, contributory science, or volunteered geographic information) can engage millions of people in biodiversity data collection and advance research and conservation worldwide (Sullivan et al. 2014, Young et al. 2019). Platforms that rely on collaborative, crowdsourced data collection have become powerful research and conservation tools, facilitating huge volumes of data and filling critical gaps in ecological knowledge. Smartphone-based platforms like iNaturalist (Nugent, 2018) and eBird (Sullivan et al., 2014) allow users to submit observations from any location, creating a wealth of incidental, presence-only biodiversity data. These observations tend to be located near urban areas, making them particularly valuable for monitoring areas that have historically received little attention from structured biodiversity research studies (Callaghan et al., 2020; de Camargo Barbosa et al., 2021). Further, crowdsourced contributory science is often heralded as a tool for public participation and engagement in science and an opportunity for connecting people with the world around them (Adler et al., 2020; Jimenez et al., 2021; Diprose et al., 2022). Thus, accessible and approachable crowdsourced data collection platforms have the potential to offer societal and personal wellbeing benefits in addition to advancing scientific research and conservation (Young et al., 2019; Peter et al., 2021).

However, evidence is accumulating that crowdsourced biodiversity data in cities, such as observations submitted to iNaturalist and eBird, are biased toward affluent, white neighborhoods

(Baker et al., 2018; Perkins, 2020; Grade et al, 2022; Mahmoudi et al., 2022; Ellis-Soto et al, 2023). eBird observations have been shown to overrepresent neighborhoods with higher median family incomes and greater proportions of white residents across several United States cities (Grade et al., 2022; Perkins, 2020). Similar bias has been found in Community Collaborative Rain, Snow, and Hail Network data (Mahmoudi et al., 2022) and Global Biodiversity Information Facility (GBIF) records (Ellis Soto et al., 2023). In GBIF records, patterns of sampling bias are not only associated with current patterns of socioeconomics, but also with historical, codified residential segregation, illustrating how legacy effects impact not only urban social and ecological landscapes, but also how we interact with them (Ellis-Soto et al., 2023). This pattern likely reflects inequity in who participates in birding and where they prefer to recreate, as well as potentially reflecting unjust distributions of other environmental attributes like noise pollution, or other social-ecological phenomena like the luxury effect, wherein wealthier neighborhoods are associated with greater plant and animal biodiversity (Hope et al, 2003; Leong et al., 2018). Consequentially, as Grade et al. (2022) point out, the biased distribution of crowdsourced contributory science observations casts doubt upon any urban ecological research that employs citizen science data without explicitly considering the social landscape and its relevance for sampling patterns. Thus, gaining a more complete understanding of the social biases present in data collection is paramount for conducting sound ecological research with crowd-sourced data and mitigating the impacts of sampling bias on research outcomes, conservation actions, and social justice.

Though most crowdsourced data collection is opportunistic (“unstructured”), some applications, including eBird (Sullivan et al., 2009), are considered “semi-structured” because they collect data about the observation process so that data filtering and analysis can mitigate

variation due to sampling (Robinson et al., 2020; Johnston et al., 2021). By asking users to designate checklists as "complete" if they have recorded all species they observed in a location, eBird data can be viewed as presence-absence data to facilitate more robust inference (Altwegg and Nichols, 2019; Johnston et al., 2021). Another distinct feature of eBird is that birders can submit observations at any location of their choice, or they may use the eBird website or phone app to search for "hotspots" in a region. These hotspots are public locations that offer convenience, good birding, or both. eBird users recommend locations for hotspot consideration, and a location receives hotspot designation after a local, volunteer "expert" reviews the location for qualification (eBird FAQ, 2023). Designated hotspots are highlighted and searchable on the app and therefore can amass frequent use and many observations.

Hotspot observations comprise a substantial portion of all available eBird data applied in research and conservation, yet sampling differences between hotspots and personal locations are a potential source of sampling bias that is not regularly accounted for in analyses. Birders may seek out hotspots as recreational destinations more frequently than other possible birding locations, such as private backyards or locations in densely populated areas they might encounter incidentally, likely resulting in different spatial patterns of hotspots versus eBird locations generally. This might make hotspots more prone to preferential sampling, or sampling biased towards where observers expect greater richness or bird counts, which can lead to overestimation in predictions unless analytically accounted for (Diggle, 2010; Sicacha-Parada et al., 2021). Finally, hotspots are typically located in public spaces considered desirable and accessible by birders, like parks and public open spaces, which, along with canopy cover, development intensity, and impervious surface cover, are distributed inequitably throughout cities (Grove et

al., 2014; Schwarz et al., 2015). Thus, current patterns of eBird hotspot availability and use may exacerbate sampling bias, including bias that is related to social inequity in urban areas.

Several studies (Kramer-Schadt et al., 2013; Robinson et al., 2020; Johnston et al., 2021) have recommended spatial subsampling (thinning, balancing) procedures to mitigate the influence of spatial bias in eBird and other crowdsourced data. However, others have noted that the choices researchers make regarding their spatial subsampling or balancing procedures can have substantial influence on inferential outcomes (Steen et al., 2021), and in the context of eBird specifically, these procedures have not been applied in a way that reflects the distinct mechanisms and behaviors underlying hotspot versus personal location selection.

While several studies have documented relationships between social factors (e.g., median household income, race) and eBird sampling locations in general, the degree to which hotspots, specifically, are biased has not been explored. Past studies of eBird sampling inequity did not distinguish between hotspots and other observation locations, and did not consider other sources of spatial dependence in the observations, such as spatial autocorrelation, which can lead to overestimation of relationships between observations and spatial covariates (Cliff and Ord, 1970). We expected hotspots to be at least as biased by the social landscape as eBird observations in general, but hypothesized exaggerated bias in hotspots relative to personal locations. In this case, bias placing hotspots in closer proximity to affluent neighborhoods may compound with other barriers preventing Black, Brown, and historically marginalized people from safely and easily accessing public outdoor space, further exacerbating exclusion in outdoor recreation and participation in crowdsourced science programs. Such a result would drastically weaken claims that eBird and similar programs are effective tools for inclusion and equitable participation in science, outdoor recreation, and conservation.

Thus, we aimed to understand the distribution of urban eBird hotspots and how it relates to landscape and social factors. We applied modeling approaches from spatial statistics to formalize the spatial relationships among all eBird observation locations, hotspots, and landscape factors within and around the city of Fresno, California, USA. Specifically, (1) we sought to describe the spatial distribution of hotspots within a city relative to ecological and social factors, including tree cover, neighborhood wealth, and demographics. Further, (2) we aimed to evaluate how spatial patterns, surrounding landscapes, and use of eBird hotspots differ from those for all eBird locations, and assess whether certain landscape or social factors are more associated with hotspots than with eBird checklist locations in general. Finally, (3) we aimed to assess whether given a set of hotspot locations, the number of checklists collected at these locations is associated with demographic and landscape characteristics of the locations. We highlight the associations between landscape and social factors and urban hotspot designation to encourage careful evaluation of how hotspots act as either barriers or facilitators for increasing access and inclusivity in birding and crowd-sourced data collection.

Methods

Study area

Fresno is the county seat and most populous city in Fresno County, California, with a population of about 540,000 as of the 2020 US Census (US Census Bureau, 2020). According to US Census data, the city of Fresno is 50.0% ethnically Hispanic or Latino and 50.1% non-white (single other race, or more than one race); 25.9% of the population is white and not of Hispanic or Latino ethnicity. As of 2021, median household income in Fresno is reported to be $57,211, and 22.9% of the population is considered to live in poverty according to the 2020 Census poverty threshold (US Census Bureau, 2020). The city is bounded to the north by the San

Joaquin River, with Madera County sitting north of the river. Madera County has a total population of around 156,000, of which 59.6% is ethnically Hispanic or Latino and 9.4% non-white (single other race, or more than one race); 31.0% of the population is white and not of Hispanic or Latino ethnicity (US Census Bureau, 2020). Together, Fresno and Madera counties comprise the US Census Fresno-Madera Combined Statistical Area (CSA).

Fresno is a largely agricultural city in the San Joaquin Valley of central California, characterized by a hot semi-arid climate. Fresno is in close proximity to Yosemite National Park and several major highways, making it a major gateway for tourists to the park (World Atlas, 2023). However, within the city, Fresno's park system is ranked 98th amongst the 100 largest US cities by ParkScore, based on park size, equity and access, investment, and amenities (The Trust for Public Land, 2023).

We conducted our study in Fresno because bias in crowdsourced avian biodiversity sampling has previously been explored in the city (Perkins 2020), enabling us to build upon and extend previous work to understand the importance of applying explicitly spatial methods to the challenge of sampling bias. Further, data from the Fresno Bird Count, a long-term systematic bird survey program in the city (Schleder 2010; Hensley et al., 2019), was helpful as a comparative dataset for assessing the efficacy of our modeling approach. We defined our spatial region of analysis by creating a 20 km buffer from the centroid of the jurisdictional Fresno city limits (City of Fresno, 2022), because this boundary captured the transition from urban to the surrounding suburban and rural regions. The selection of this boundary was thus independent from the distribution of observation locations, which is appropriate for the point process model we fit. Our boundary included a portion of Madera County, north of the San Joaquin River, and

allowed us to consider eBird checklists collected on both sides of the river to understand how the river may function as a barrier between for eBird users on either side.

Data

We downloaded the eBird Basic Dataset (EBD) and sampling event data for Fresno and Madera Counties, CA (Fresno County: April 2022 release, accessed 08 June 2022; Madera County: April 2023 release, accessed 16 May 2023, eBird 2022). We downloaded data for all species in these regions from April 2011-June 2019, and used the 'auk' package in R to filter data prior to analysis (Strimas-Mackey et al., 2018). We filtered data to include only records from April 1 to June 30 across all years, and only included complete checklists with 'stationary' or 'traveling' protocols, durations less than 5 hours, distances less than 5 km, and up to 10 observers following the eBird best practices outlined by Johnston et al. (2021). Using these data, we created a dataset of all unique checklists collected within our study period, and a second dataset containing each unique location ("hotspot" or "personal") used for at least one checklist during this period. Additionally, we downloaded the set of all eBird hotspots from eBird (https://confluence.cornell.edu/display/CLOISAPI/eBird-1.1-HotSpotsByRegion), and extracted all hotspot locations within our study area, which may or may not have checklists within our study timeframe.

Additionally, we removed eBird observations associated with the Fresno Bird Count (FBC) from our dataset. The FBC was an annual systematic survey in which volunteer observers conducted spring avian point counts at a designated set of locations throughout the city, from 2008 to 2015 (Schleder 2010; Hensley et al., 2019), based on the protocols developed for the Tucson Bird Count (Turner 2003). Many FBC volunteers uploaded their observations collected during the survey, at the official survey sites, to eBird as personal checklists. We removed these

checklists from our dataset because as a systematic, gridded survey, FBC checklists arose via markedly different site selection and sampling protocols than other eBird checklists. We obtained FBC data from program founder Professor Madhusudan Katti in November 2022 and removed all checklists collected as part of the Fresno Bird Count by cross-referencing recorded locations with Bird Count locations and filtering by eBird comments and location IDs.

We obtained percent impervious surface cover and landcover class data from the 2019 National Landcover Dataset (NLCD) release and percent canopy cover from the 2016 NLCD release (Dewitz, 2021) using the 'FedData' R package (Bocinsky, 2023). We combined landcover class with land use data from the National Land Use Dataset (Theobald, 2014) to derive additional spatial predictors, including residential and commercial development intensity and combined land cover/use classes. We downloaded a digital elevation model from The National Map (US Geological Survey, 2021), and calculated distance to nearest river or stream using a hydrology shapefile from the California Department of Fish and Wildlife (https://data-cdfw.opendata.arcgis.com/maps/CDFW::california-streams-1). Finally, we used the ParkServe shapefile of all designated parks (The Trust for Public Lands, 2023*)* to identify park locations and compute distance to nearest park throughout the study area. We conducted all spatial analyses in an Albers Equal Area projection in R using the 'sf,' 'raster,' and 'terra' packages (Pebesma, 2018; Pebesma and Bivand, 2023; Hijmans, 2023(a); Hijmans, 2023(b)). We calculated mean percent canopy cover, percent impervious surface, and development intensity at several scales around observation locations (radii 100m, 500m, 1 km, 2km) as possible predictor metrics.

For all demographic and socioeconomic variables, we used 5-year estimates from the U.S. Census American Community Survey (ACS; https://www.census.gov/data/developers/data-

sets/acs-5year.html) for the five-year period covering 2015-2019. These estimates are obtained from analyzing all surveys collected over the 5-year time period, and thus represent "period estimates" rather than "point-in-time" estimates. However, including samples over five years allows for a larger sample size and thus greater precision estimates. We used ACS estimates provided at the block group level, and because the eBird observations were collected over several years, it was more important to prioritize greater spatial resolution and precision and detailed community information than current, annual estimates (e.g., ACS 1-year estimates) or data covering fewer relevant variables (e.g., U.S. Decennial Census). Census data were downloaded, visualized, and processed via the R package 'tidycensus' (Walker and Herman, 2023).

We included race, ethnicity, median household income, proportion of households living below poverty line, population density, median house age, and proportion of owner-occupied housing units, and Gini index (census tract level) in our set of possible demographic variables. We obtained each variable as an estimate of the number of people or households within the block group fitting each category, and calculated the proportion per block group by dividing this estimate by the estimate of total population or households within the block group for proportion of population in each race, proportion below poverty line, and proportion of owner-occupied housing units. The US Census reports the following races: Asian, Black, Hawaiian/Pacific Islander (HIPI), American Indian/Alaskan Native (Native), White, Single Other, and Two or More Races and Hispanic/Latino ethnicity (US Census Bureau, 2020). From these groupings, we disaggregated white into white/Hispanic or Latino, and white/not Hispanic or Latino. We also combined Native, HIPI, single other, and two or more races into one group, because the sample sizes of each of these were small and did not enable us to include them individually. We obtained

population density by dividing the estimated total population within a block group by the area of the block group. Additionally, we computed local residential segregation indices for race and income using 'OasisR' (Tivadar, 2019) and included these as possible predictor variables. Finally, because many of Census variables were correlated, we used Principal Components Analysis (PCA, Wold et al., 1987) to compute principal components that capture the primary variation across a set of variables, and included the first and second principal components as possible alternative predictors to individual census variables. We centered and scaled all predictor variables to have a mean of 0 and a standard deviation of 1 prior to analysis to assist with model fitting and to enable comparison of estimated effect sizes across predictors.

Model and Analysis

The locations of eBird observations are related to each other in space via multiple types of spatial structure, including relationships with spatial covariates – like canopy cover and human population density – and spatial autocorrelation, or correlations between observations that arise due to their proximity to each other. To account for these sources of spatial dependence, we modeled eBird observations within a city as a spatial point pattern by fitting a Log-Gaussian Cox Process (LGCP) model to the observations. In this model, the observation locations are modeled as a nonhomogeneous Poisson process where the underlying intensity surface varies through space via a Gaussian random field and may be governed by a set of spatial covariates (Banerjee et al., 2015).

We specified an LGCP model with multiple likelihoods to estimate the intensities of the eBird hotspot and personal observation location point patterns separately,

$$\log(\lambda_h) = \beta_{0h} + \omega_C(\boldsymbol{s}) + \frac{1}{2}\omega_D(\boldsymbol{s}), \quad (1)$$

$$\log(\lambda_p) = \beta_{0p} + \omega_C(\boldsymbol{s}) - \frac{1}{2}\omega_D(\boldsymbol{s}), \quad (2)$$

where λ_h is the intensity of the hotspot locations and λ_p is the intensity of the personal observation locations. The Gaussian spatial processes $\omega_C(\boldsymbol{s})$ and $\omega_D(\boldsymbol{s})$ capture the spatial structure that is shared between the intensities and that differs between the intensities, respectively. Thus, in this model, personal locations and hotspots have separate intensities, which have two Gaussian random fields in common – one that reflects the dynamics that are shared between the two point processes and the other capturing the difference between the two patterns. Additionally, the two intensities can include distinct or shared spatial covariates to assess how covariates influence one or both of the processes and how the addition of covariates mediates the residual shared versus different spatial structure. We used this model in a primarily exploratory capacity to quantify how hotspot intensity and spatial distribution differs from the intensity of eBird locations overall. We fit this model to the Fresno eBird data under an approximate Bayesian inferential framework using integrated nested LaPlace approximation via the R-INLA and 'inlabru' R packages (Rue et al., 2009; Lindgren et al., 2011; Bachl et al., 2019). The Gaussian random fields were represented in model fitting as spatial partial differential equation (SPDE) random effects, using a Matérn covariance function with penalized complexity priors on the hyperparameters (Fuglstad et al., 2016; Simpson et al., 2017).

We then specified a point process model with the following log-Gaussian Cox Process (LGCP) intensity,

$$\log(\boldsymbol{\lambda}(\boldsymbol{s})) = \beta_0 + \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \omega(\boldsymbol{s}), \quad (3)$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{x}(\mathbf{s})$ is a vector of spatial covariates at location $\mathbf{s}$, standardized to mean 0, standard deviation 1 for computational efficiency and ease of interpretation of importance across covariates. The Gaussian random field $\omega(\boldsymbol{s})$ captures spatial structure in the intensity of locations not described by the spatial

covariates. We again implemented $\omega(\boldsymbol{s})$ using an (SPDE) random effect with a Matérn covariance function with penalized complexity priors in INLA. We fit this model first to all eBird locations, including hotspots, used at least once within the years of our study, and second to the set of eBird hotspots exclusively. We evaluated the estimated SPDE ($\omega(\boldsymbol{s})$) for evidence of overfitting and conducted model comparison using marginal log likelihood (Fong and Holmes, 2020) and analysis of the spatial residuals for each model across several spatial scales (Peddinenikalva 2023). We considered the same set of possible spatial covariates when fitting to both all locations and hotspots only, to evaluate which covariates were most associated with each of these point patterns. We assessed the sensitivity of the model to mesh and prior specification by fitting the model using a sequence of mesh sizes and hyperparameter values.

Finally, we evaluated if, given the set of eBird hotspot locations, landscape or social factors were associated with the intensity of use (i.e., number of checklists) of individual hotspot locations. We considered a binary-Poisson hurdle model (R package 'pscl', Zeileis et al., 2008) and fit this model to all hotspots. We also fit a Poisson regression to the set of hotspot locations that were used at least once during the study period. For each of these models, we again considered the same possible spatial covariates as in other models, constructing models using additive and interactive effects of these covariates. We conducted model selection between these model specifications using AICc (Burnham and Anderson, 2004) to determine which covariates were associated with the count of uses per hotspot.

Results

There were 313 unique eBird checklist locations within our study boundary in Fresno, CA for the period from 2011 to 2019; of these locations, 48 were designated eBird hotspots (15.3%) and the remaining 265 were 'personal' locations (Fig. 2.1). Across these locations, there were 1263 complete checklists submitted to eBird between 2011 and 2019 that were conducted

independently of the Fresno Bird Count – 720 (57.0%) of these checklists were collected at hotspots and the remaining 543 (43.0%) at personal locations. Hotspots had an average of 49 (range [1,94]) complete checklists per hotspot, and personal locations, on average, had 6 (range [1, 27]) checklists. Additionally, there were 25 locations designated as hotspots that did not have any checklists collected between 2011 and 2019. Not accounting for differences in sampling duration, distance, or other eBird effort variables, mean species richness for hotspot checklists was 21.3 (range [1, 62]), and mean richness at personal locations was 10.4 (range [1, 80]).

The full set of spatial covariates we considered as predictor variables included: percent impervious surface cover, percent canopy cover, elevation, distance from San Juan River, distance to nearest park, built development intensity (residential and commercial), population density, median household income, median house age, proportion of non-Hispanic white residents, local racial diversity index, local Shannon diversity index for race, and Gini index. Although we initially considered additional covariates, we removed these because they were highly correlated with the retained variables. Correlations between the reduced covariate set were low to moderate (range [0.02, 0.69]), and many demographic variables were moderately correlated (range [0.1, 0.92]).

Multiple likelihood model

The locations of all eBird checklists and the locations of eBird hotspots were not distributed uniformly throughout the city, and were described by separate underlying intensities Fig. 2.2). The SPDE describing the spatial effect shared by the two intensities had an estimated spatial range of 34.04 (sd 16.23) and an estimated sd of 1.56 (sd = 0.55), and the SPDE describing the difference between the intensities had an estimated spatial range of 44.66 (sd 41.83) and an estimated sd of 0.71 (sd = 0.34). The largest difference between the two intensities

occurs in the NE to SW gradient where there is greater predicted intensity for hotspots in the NE and fewer predicted hotspots in the SW relative to the predicted intensity for all locations (Fig 2.1b).

The shared intensity surface was partially described by the percent of impervious surface cover (1 km buffer), which was positively associated with location intensity ($\hat{\beta}$=1.130, 95% CI (0.820, 1.467)). Additionally, there was evidence that median household income explained some of the differing spatial structure between the two intensities, with a positive association between income and hotspot location intensity ($\hat{\beta}$=0.293, 95% CI (0.076, 0.515)) but not all eBird locations generally.

All eBird sampling locations

The distribution of unique eBird locations was best described by a point process model containing the proportion of impervious surface cover, distance from the nearest formal park, and to a mild degree, the surrounding demographics. Models that included only impervious cover, impervious cover and distance from park, and impervious cover and demographics were similarly competitive based on marginal log likelihood comparisons (Table 2.1). Therefore, we considered all of these models when making inference, but emphasize the model containing all three covariates, because the more reduced models tended to show evidence of underestimation of point intensities throughout the study region. The coefficients were consistent in direction and 95% credible intervals (CIs) for each parameter overlapped substantially from model to model, and did not result in notable differences in log-intensity predictions. Further, all models show evidence of reflecting exaggerated influence of a small, heavily sampled region within the study area, and this is lessened when including all variables. Lastly, estimates of the parameters (range and standard deviation) of the Gaussian random field SPDE term were sensitive to the prior and

showed evidence of spatial confounding between the SPDE and the fixed effects due to the covariation between the spatial autocorrelation in the points and fixed effects.

The proportion of impervious surface cover was positively associated with log-intensity ($\hat{\beta} = 1.061$ (95% CI (0.79, 1.332)), where impervious surface cover was calculated in a 1 km buffer around each location (Fig. 2.3). Distance to the nearest formal park was negatively associated with log-intensity ($\hat{\beta}$ = -0.818, 95% CI (-1.467, -0.168)), meaning that the intensity of points was more associated with being near parks and decreased moving farther from a park. The most supported demographic metric was the first principal component of a principal components analysis including proportion of non-Hispanic white residents, proportion of Hispanic white residents, proportion of Black residents, proportion of Asian residents, proportion of all other races, proportion of residents living below the poverty line, and the median house age in the block group ($\hat{\beta}$ = 0.293, (0.102, 0.484)). This principal component reflects positive correlations between the proportions of Black, Asian, and Hispanic/Latino residents and the proportion of households living below the poverty line versus the proportion of non-Hispanic white residents. Thus, the positive estimated coefficient with this predictor is opposite our expectation that sampling intensity is lower in predominantly low-income communities and communities of color. However, the sign (direction) of this coefficient was reversed when the spatial random effect (SPDE) was omitted from the model. This suggests that the estimated spatial random effect was confounded with the demography fixed effect. The cause of this behavior and how to address it is an unresolved topic in spatial modeling with INLA (Sørbye et al., 2019; F. Lindgren, J. Illian, and S. Martino, pers. comm.), but indicates that the spatial random effect is "absorbing" or reversing some dimension of the fixed effect, and we must be thoughtful with interpretation. In this case, it is likely arising at least in part because there are relatively few observations given

the spatial area and range of spatial covariates. The estimated spatial effect had a range of 17.95 km (sd 6.02) and a standard deviation of 1.14 (sd 0.288) (Fig 2.3e). The estimation of the spatial random effect and the estimation of fixed effects were minimally sensitive to changes to mesh specification and hyperparameters in the SPDE prior.

Residual analysis showed that the model fits the data best in the central to northeast portion of Fresno County, which is the portion of the study area with the greatest number of observations and the greatest heterogeneity in covariate values. The largest-magnitude residuals occur north of the San Juan River in Madera County, where there is lower population density and fewer observations than Fresno County.

Hotspot locations

The most supported LGCP model fit to hotspot locations included fixed effects of distance from the nearest park and the proportion of non-Hispanic white residents in the block group containing the hotspot (Table 2.2, Fig. 2.4). Median household income was also supported as a predictor of hotspot intensity, though income and proportion of white residents were correlated and could not be included in the same model, so we chose the most predictive of the two covariates. Similarly to the case of all eBird locations, distance to nearest park was negatively associated with the intensity of hotspot locations $\boldsymbol{s}$ ($\hat{\beta}$ = -3.797, (-5.080, -2.508)), and the estimated coefficient indicates that hotspots are more strongly associated with proximity to a park than eBird locations overall. The proportion of non-Hispanic white residents was positively associated with hotspot locations ($\hat{\beta}$ = 0.355, (0.059, 0.637)), suggesting that hotspot intensity tends to be higher in areas with a greater proportion of non-Hispanic, white residents. Median household income was nearly as supported as the proportion of non-Hispanic, white residents,

and considering median household income instead of proportion of non-Hispanic white residents, median income was positively associated with hotspot intensity ($\hat{\beta}$ = 0.136, (-0.030, 0.302)).

The estimated range parameter of the spatial random effect was 173 km, although fixing this hyperparameter to 50 km did not influence inference or prediction from the model. This indicates evidence that the spatial random field was estimated to be constant within our study area after the fixed effects were included, and removing the spatial random effect entirely resulted in the most supported model by maximum likelihood.

Visual inspection of the spatial residuals for this model fit to hotspots showed consistent residuals of small magnitude throughout the spatial region, suggesting there were no portions of the study area that were notably better or worse fit under this model. This is consistent with the evidence that there was not strong spatial dependence remaining after covariates were included. Relative to the predictions from the multiple likelihood model, we see lower local predicted hotspot intensity, because the mass is spread out over a greater area overall, reflecting the importance of the fixed effects in predicting intensity.

Hotspot use

Social factors were also associated with the intensity of use of different hotspots. First, we did not find support that any of our covariates associated with the hurdle component of the use model. In other words, we did not find any covariates that predicted whether a designated hotspot was used or not used between 2011 and 2019. Given the process for hotspot designation (hotspots are designated after there are existing checklists created), these hotspots had to have been used either prior to 2011 or after 2019, though it cannot be determined from the data when they were designated as hotspots.

On the other hand, considering the hotspot locations used at least once between 2011 and 2019, race and impervious cover (0.5 km buffer), and the proximity of a hotspot to a park were associated with the number of checklists at a location (Table 2.3). A model that included these fixed effects plus the distance to the San Juan River ('distSJR') received nearly as much support, and we selected the most parsimonious model. The proportion of non-Hispanic white residents was positively associated with the number of checklists collected at a hotspot ($\hat{\beta}$ = 0.0738, 95% CI (0.003, 0.149)). The number of checklists at a hotspot was negatively associated with the proportion of impervious cover at a hotspot ($\hat{\beta}$ = -0.491, 95% CI (-0.552, -0.428)) and with increasing distance from a park ($\hat{\beta}$ = -0.208, 95% CI (-0.304, -0.123).

Discussion

We found that eBird observation locations in Fresno, California tend to occur in areas of the city with greater impervious surface cover that are in close proximity to parks. Further, although there is evidence that eBird locations overall are biased by demographics, this bias is much more pronounced in the locations of hotspots specifically, where hotspots are positively associated with higher proportions of non-Hispanic white residents. This relationship became stronger still when considering how heavily different hotspots are used, in addition to their spatial distribution alone. In Fresno, hotspot locations and use were more strongly associated with race than median household income (though these demographic metrics covary) and the proportion of non-Hispanic white residents is positively correlated with median household income.

We found that personal observation locations and hotspot locations do not share the same underlying spatial intensities (Fig. 2.1). This was evident both when fitting models to the two sets of points separately, and when fitting a single model to the two sets to estimate the portions

of the sampling intensity that were shared and distinct between the two. When estimating the two intensities in a single model, the common structure between the two sets of locations was associated with the proportion of impervious surface cover, whereas the difference between the two processes appears primarily driven by demographic factors. Understanding these differences between the sampling intensities of personal locations and hotspots is important when considering how to account for sampling bias in analyses of eBird data, because the differing observational processes may need to be treated differently when accounting for sampling bias.

Impervious surface cover was likely predictive of eBird locations overall because areas of high impervious cover reflect areas that are highly populated in Fresno. Thus, we expect these relationships capture the tendency of eBird locations to be collected in areas with more people, as has been found in other research exploring crowdsourced data collection in urban areas (Geldmann et al., 2016; Mair and Ruete 2016; Baker et al 2018). We did not find that the relationship between impervious cover and locations carried through to hotspot locations specifically, which implies that among the set of all eBird locations, hotspots tend to be less associated with impervious cover, instead reflecting birders' assumptions or preferences regarding where to observe birds.

The result that parks are associated with eBird locations in general, and particularly with hotspots, matches our expectation, because parks are frequently used for recreational birdwatching (Kuldna et al., 2020; Lopez et al., 2020; Kurnia et al., 2021). Given that Fresno has a relatively high proportion of impervious surface cover and comparatively little canopy or vegetated ground cover and that parks tend to be located within heavily developed areas, it makes sense that parks located within a context of predominant impervious surface cover would be frequent sampling locations, as these parks may act as a 'refuge' for wildlife and recreational

birders alike (Vasquez and Wood, 2022). Further, hotspots must occur in public locations, so it is unsurprising that they are more strongly associated with parks than other eBird locations, which may occur in private backyards or informal greenspaces.

We found that the general set of eBird locations was somewhat associated with a composite demographic metric reflecting multiple socioeconomic axes. This general set of eBird locations includes 'personal' lists that occur when birders submit a checklist from anywhere other than a hotspot, often observations occurring their own yards or incidentally. Therefore, these points likely reflect locations that are convenient for eBird contributors, like locations close to their homes or neighborhoods, and this relationship between checklist locations and demographics could reflect existing disparities in participation in recreational activities like birding (Carver, 2009; Blake et al., 2020; Rutter et al., 2021). Thus, this result provides evidence that there is a demographic gradient associated both with hotspot designation and patterns of convenience-based eBird sampling.

We found evidence of stronger relationships between the hotspot-only distribution and use and demographics, specifically race and income, implying that social sampling bias is more pronounced amongst eBird hotspots relative to all checklist locations. This suggests that eBird users prefer areas with greater proportions of white residents and higher incomes for hotspots, either because these areas do in fact reflect higher bird biodiversity or better bird habitat, indicating the presence of a luxury effect (i.e., a pattern of higher biodiversity associated with greater socioeconomic status, Hope et al., 2003), or the *perception* that these areas contain greater biodiversity due to overall habitat characteristics or disparities in the availability of parks or canopy cover.

Alternatively, this pattern could reflect the demographics of eBird's userbase and corresponding convenience, familiarity, or accessibility of these locations for typical eBird contributors, and/or users who feel entitled or qualified to suggest or review hotspots (Kuldna et al., 2020). The demographics of birders in the U.S. are not representative of the general population, overrepresenting white, affluent people who were able to access higher education, often a master's degree or higher (Pateman et al., 2021; Rutter et al., 2021). This pattern is further exaggerated amongst people who volunteer biodiversity data to eBird and other crowdsourcing platforms – eBird participants represent the most specialized subset of recreational birders in the U.S. (Cooper and Smith, 2010; Haklay, 2016; Rosenblatt et al., 2022). These trends in participation likely underly the social spatial bias we observe in crowdsourced data collection and in hotspot distributions in particular, as participants suggesting or reviewing hotspots likely reflect the most specialized of the most specialized.

Capturing the true relationships between bird distributions and social factors and disentangling these possibilities is challenging, because it is a significant hurdle to account for the effects of sampling bias on observed data when the sampling bias is strongly confounded with the biological process of interest. Thus, developing and applying statistical methods to address this challenge and distinguish between possible mechanisms is beyond the scope of this study. At the same time, this work accounting for spatial autocorrelation in observations and incorporating spatial covariates is an important step toward disentangling sampling and biological processes. Our spatially explicit point process approach is less prone to overestimating relationships between sampling intensities and spatial covariates, which can occur when autocorrelation is not considered. Although other studies have established evidence of social sampling bias in crowdsourced data collection (Perkins, 2020; Grade et al., 2022; Ellis-Soto et

al., 2023), we demonstrate that sampling bias associated with demographic factors persists in eBird sampling locations even after accounting for spatial autocorrelation.

However, these spatial models are complex and can be limited in practice because they can be extremely computationally intensive and often contain model components that are difficult to identify (Sørbye et al., 2019; Cole, 2020). The point process modeling approach via INLA and inlabru is appealing because it has more familiar syntax and out-of-the-box readiness than other methods for implementing spatial models, and is more computationally feasible than methods relying on numerical integration and Monte Carlo approximations of the point process likelihood (Green et al., 2015; Dinsdale and Salibian-Barrera 2019). However, mitigating spatial confounding between the fixed effects associated with spatial covariates and the estimated spatial Gaussian random field is difficult (Azevedo et al., 2023), and as a result, the fixed effects may be conservative estimates of the relationship between a spatial covariate and the response (Sørbye et al., 2019; F. Lindgren, J. Illian, and S. Martino, pers. comm.). These issues related to identifiability between model components are exacerbated in cases like ours with limited data. The small set of unique sampling locations and correlated spatial covariates necessitated strong priors on the spatial random effect to attempt to lessen confounding. Correlations between covariates, particularly the demographic variables, made it difficult to distinguish which factors underlie the associations observed. For example, the proportion of canopy cover was not a supported predictor in any of our final models, despite receiving some model support for some response variables prior to the inclusion of demographic variables. The covariation between canopy cover and demographics and the overall low proportion (and variance) of canopy cover throughout Fresno likely prevented this variable from explaining any additional structure in the distribution of eBird locations when included with demographic variables.

Our finding of mild support for social sampling bias in locations overall aligns with other work that found a weak positive correlation between median household income and eBird sampling density in Fresno (Perkins 2020). However, Fresno is likely not representative of eBird processes in many cities, because Fresno has far fewer eBird observations and much lower eBird participation than most similarly sized United States cities. Further, Fresno has greater impervious surface cover and less canopy cover and parks than most U.S. urban areas (The Trust for Public Land, 2023).

However, it may be the case that Fresno represents the early stages of establishing socially biased sampling patterns that become more severe as crowdsourced data collection scales up in a region. We found that among the set of all designated eBird hotspots in Fresno, use of hotspots was associated with race, with hotspots in whiter areas amassing more checklists than other hotspots. Thus, evidence that hotspot locations and use are more biased by social factors than personal locations suggest that relationships between demographic factors (specifically race and income) and eBird sampling may become more pronounced as eBird participation and data volumes increase, particularly at designated hotspots. We see a similar pattern in crowdsourced data at the national level, where social bias in sampling has increased by 35% as programs like eBird have grown over the past 20 years (Ellis-Soto et al., 2023). Stronger relationships between sampling density, race, and median household incomes have been documented in other U.S. cities with greater participation and data volumes than Fresno, including Tucson, Arizona; Raleigh, North Carolina; Boston, Massachusetts; and Phoenix, Arizona (Perkins 2020; Grade et al., 2022). Future work should explore the extent to which hotspots and checklists collected at these hotspots drive social sampling bias in cities like these, which represent greater diversity in climate, land cover, demographics, and eBird participation. We expect that in these cities, which

have more unique sampling location and more observations collected at personal locations, there may be a smaller difference in the distributions and drivers of hotspots versus personal locations. However, we also expect that a consistent pattern of socially biased sampling will be more evident with larger datasets, in corroboration with previous studies, and that cities with a broader range of demographics will display wider sampling disparities. Finally, we expect that factors like climate and the development history of cities may impact social sampling bias in cities. For example, evidence of a luxury effect has been most pronounced in arid regions where supplemental watering supports greater plant biodiversity (Chamberlain et al., 2020), which may in turn create a wider gradient of avian habitat that either seems to or actually supports greater avian biodiversity, thus influencing eBird participant behavior.

Evidence that social sampling bias is more pronounced among eBird hotspots than personal locations, and that both hotspot locations and use overrepresent predominantly white areas in Fresno, raises concerns that the eBird hotspot system may be exacerbating spatial bias in sampling. Existing urban eBird hotspots, paired with how users interact with the eBird platform, may reinforce patterns of inequity in data availability and perpetuate barriers to inclusivity in birding and crowdsourced data collection. Biased hotspot distributions perpetuate oversampling in predominantly white neighborhoods and undersampling in low income and predominantly Black and Brown neighborhood, because hotspots are advertised prominently on the eBird platform for other users to visit. As a result, hotspots are disproportionately represented in eBird checklists – though they make up only 15% of all locations, they represent 57% of all checklists collected in Fresno. Finally, because the number of checklists at a hotspot is positively associated with whiteness, hotspots in the already most over-sampled areas amass the most data, reinforcing these patterns and widening the inequity in data coverage.

As a result, data "cold spots" in historically marginalized communities prevent us from being able to monitor biodiversity in these areas or effectively understand patterns of biodiversity across representative demographics gradients (Ellis-Soto et al., 2023). For example, because inequitable data gaps and sampling patterns parallel relationships expected under the luxury effect, we lack data representative of the full range of socioeconomic gradients in Fresno, precluding a rigorous test of this hypothesis.

eBird and other crowdsourced data programs are often cited as potential tools for democratizing science and increasing inclusivity and equity in recreational activities like birding (Dickinson et al., 2012; Pocock et al., 2018; Paleco et al., 2021). However, the prevalence of social sampling bias in crowdsourced programs and evidence that the eBird hotspot system exacerbates this bias challenge the idea that simply engaging broader communities in birding and eBird participation will promote inclusion and accessibility in outdoor recreation and conservation. Recognized barriers to participation in birding and crowdsourced data collection include stereotypes about who belongs in or enjoys outdoor recreation and birding, exposure and access to technology and resources, and safety and accessibility of public outdoor spaces where these activities often take place (Hobbs and White, 2012; Pandya, 2012; Pateman et al., 2021). Underrepresentation of certain neighborhoods on the user-facing side of the eBird platform may reinforce these barriers, spatially reflecting stereotypes about who belongs in birding and uses eBird (Jones, 2020; Martin et al., 2023). It also can exacerbate hurdles like requiring a car or other transportation to access hotspots in other neighborhoods or necessitating that Black and Brown birders risk their safety and wellbeing to visit hotspots in public green spaces where they are at increased risk of experiencing violence, harassment, or policing (Bittel, 2020; Hoover and Lim 2021).

Our findings highlight the need to reimagine the process for designating eBird hotspots to lessen how the current design and use of crowdsourced data platforms contribute to disparities in sampling densities, data cold spots, and exclusion (Callaghan et al., 2019a; Blake et al., 2020; Montanari et al., 2021; Mahmoudi et al., 2022). The current procedure for designating eBird hotspots is unsystematic and lacks transparency, and combined with unequal demographic representation in eBird participation, likely reinforces inequitable sampling distributions and data collection. Under the current system, locations are suggested as hotspots by eBird users, and anonymous, volunteer reviewers approve or deny these recommendations based on their own familiarity with the area (eBird FAQ, 2023). Thus, hotspots reflect where people like to bird rather than what areas support the greatest biodiversity (eBird FAQ, 2023) and reflect recommenders' and reviewers' general perception of a location (i.e., familiarity, convenience, amenities). This reinforces confounding between the sampling process and the pattern being observed, making analytically accounting for sampling more difficult, and, if the demographics of hotspots reviewers reflect the demographics of eBird users in general, propagates the overrepresentation of whiteness and affluence in eBird participation into the spatial distribution of hotspots.

In general, the hotspot procedure should be updated to improve transparency and clarify the intent of hotspots. There should also be a dedicated effort to augment the current set of urban hotspots to reflect the diversity of neighborhoods and demographics within cities and fill in data cold spots that often occur alongside other environmental and social injustices. Beyond eBird, other biodiversity monitoring programs that rely on voluntary data submission could provide equitably distributed locations as entry points into observing biodiversity.

However, it is important to note that simply adding additional hotspot locations is only a starting point and investment in additional interventions and continued monitoring of the outcomes of these interventions are critical for meaningfully increasing equity in contributory science and access to its benefits. We echo concerns raised by Mahmoudi et al. (2022) that caution against tokenizing participation in data collection to extract data from low-income communities and communities of color and recognize that addressing data cold spots likely necessitates shifting the burden of remedying these data gaps onto these communities. Thus, partnering with existing community organizations, aligning with community values, and centering equity throughout the stages of a crowdsourced project are critical for improving equity in participation and spatial coverage in data collection (Pandya, 2012; Mahmoudi et al., 2022). It is important that the creation of more equitable distributions of eBird hotspots is combined with these strategies to create more opportunities for birders from all backgrounds to participate in eBird data collection near their homes and foster a sense of community, ownership, belonging, and relevance (Pandya 2012). Continued monitoring of outcomes should be used to check that increasing birding traffic in Black and Brown neighborhoods does not have unintended harmful consequences like green gentrification or limiting residents' safety and use of public green space near their homes through increased policing (Mullenbach and Baker, 2020; Rigolon et al., 2020).

Evening out data collection is paramount for advancing our understanding of urban ecological processes and confidently using crowdsourced data to produce just ecological knowledge and conservation. eBird and other crowdsourced data are increasingly being applied to habitat protection, regulatory, planning, and policy contexts (Sullivan et al., 2017; Callaghan et al., 2019b; Young et al., 2019; Stuber et al., 2022). Therefore, eliminating data cold spots is

necessary not only for capturing true patterns of urban biodiversity, but also to prevent crowdsourced biodiversity data from reinforcing social and environmental injustices. Affluent, predominantly white neighborhoods already benefit from greater research attention, power to influence conservation initiatives, and investment in creating and maintaining green spaces (Warren et al., 2011; Rigolon 2016; Schell et al., 2020). For example, when participation in a crowdsourced project to monitor river quality in Illinois overrepresented white residents, volunteers primarily monitored streams in that were not environmental justice concerns, while overlooking streams that may pose health risks (Blake et al., 2020). We need spatially balanced data that do not privilege these areas to accurately understand how resulting inequitable distributions of parks and pollutants affect biodiversity and access to the ecosystem services they may provide.

It is important to keep in mind that our study does not address the quality of the related eBird observations or how bird occupancy or biodiversity may correlate with social and demographic factors, such as has been suggested in the luxury effect hypothesis (Leong et al., 2018). However, deepening our understanding of related sampling bias is critical to begin to disentangle these processes, and future work should aim to understand how sampling bias impacts inference, species distribution models, and other understanding gained through crowdsourced data products. Systemic undersampling in historically marginalized communities perpetuates inequitable urban ecological knowledge and conservation, and works against efforts to make birding and contributory science more inclusive. We cannot claim eBird and similar crowdsourcing programs as tools for democratizing science or equitably engaging the public in data collection and the scientific process while simultaneously promoting systems that spatially exclude many communities. While the eBird hotspot program may be one such system,

reimagining eBird hotspots to promote sampling in data cold spots and create more equitable data coverage across urban landscapes represents one immediately actionable, relatively low investment approach for beginning to mitigate social sampling bias in eBird data collection.

Tables and Figures

Table 2.1. Marginal log-likelihoods (MLL) for LGCP point process models fit to all eBird sampling locations in Fresno and Madera counties, CA between 2011-2019 to identify environmental and demographic factors associated with sampling intensity across all eBird observations. The most supported model for describing the distribution of eBird sampling observations in Fresno included effects of the proportion of impervious surface cover (Imperv), the distance from the nearest park (DistPark), and first principal component from a principal components analysis to summarize variation across several demographics metrics (DemPCA).

Model	MLL
Imperv + DistPark + DemPCA	-2298.29
DistPark	-2299.21
DistSJR	-2299.27
Imperv + DistPark	-2299.78
SPDE only	-2299.91
DistPark + DistSJR	-2301.12
Canopy + DistPark + DistSJR	-2301.29
Income + Imperv + DistPark	-2305.69
Income	-2306.01

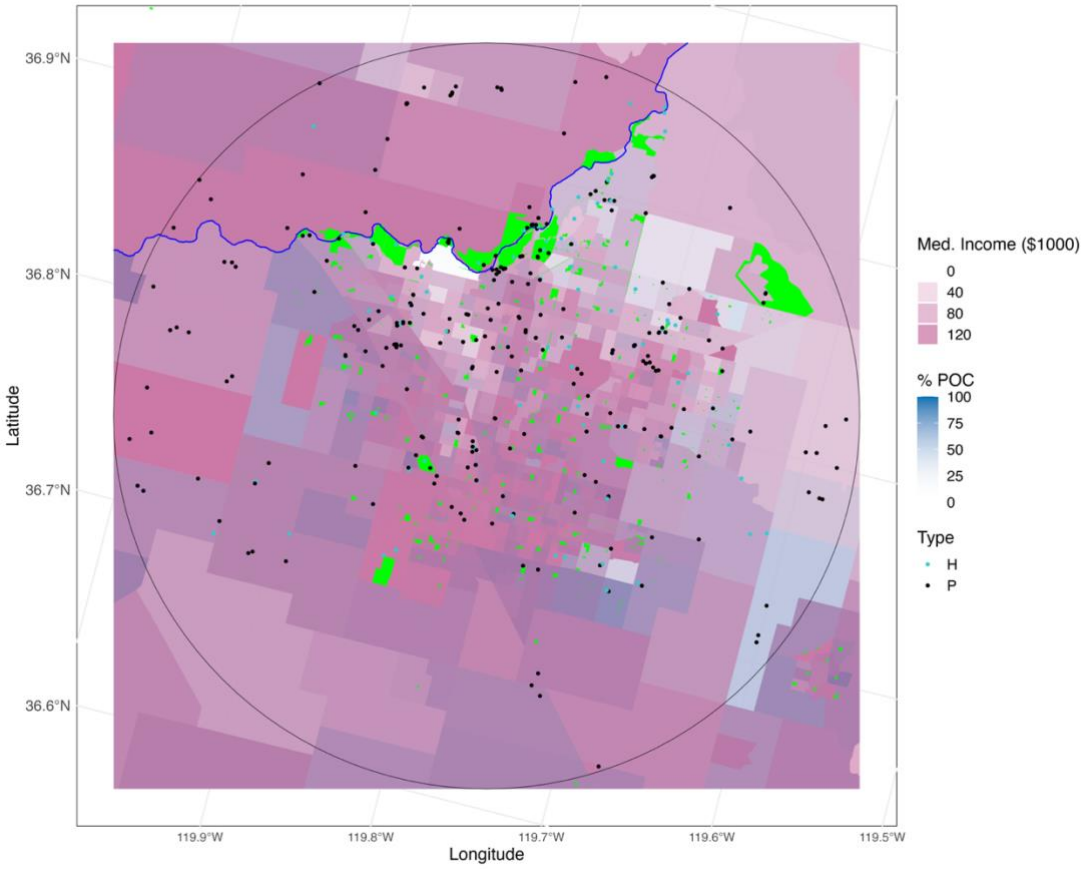
Table 2.2. Marginal log-likelihoods (MLL) for LGCP point process models fit to eBird hotspots in Fresno and Madera counties, CA (n=73) to identify environmental and demographic factors associated with hotspot sampling intensity. The most supported model for describing the distributions of hotspots in Fresno included effects of the distance from the nearest park (DistPark) and the proportion of non-Hispanic white residents in the block group containing the hotspot (PropNHW).

Model	MLL
PropNHW + DistPark	-485.19
DemPCA + DistPark	-485.65
DistPark + DistSJR	-486.72
Income + DistPark	-487.86
Canopy + DistPark + DistSJR	-489.23
Canopy	-501.41
Canopy + DistSJR	-502.34
DistSJR	-502.84
SPDE only	-503.57
DemPCA	-506.98
PropNHW	-507.09
Imperv	-507.18
Income + DistSJR + Canopy	-507.98
Income	-508.61
Loc. Div. index	-508.62
Income + Imperv + DistSJR	-511.91

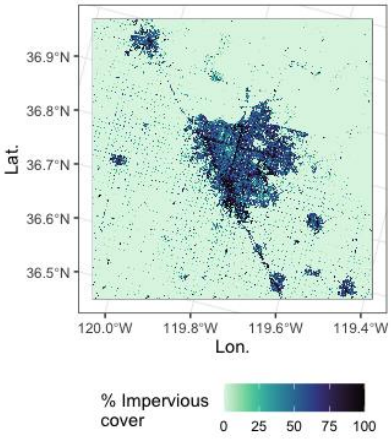
Table 2.3. Akaike information criteria (AIC) for Poisson generalized linear models (GLMs) exploring factors that are associated with the number of checklists collected at eBird hotspot locations in Fresno, CA, using data collected in spring (April-June) 2011-2019. The model that received the most support included fixed effects of the proportion of impervious surface cover (Imperv), the proportion of non-Hispanic white residents (PropNHW), and the distance to the nearest park (distPark).

Model	K	AICc	Delta_AICc	AICcWt	LL
Imperv + PropNHW + distPark	4	1084.81	0	0.47	-537.94
Imperv + PropNHW + distPark + distSJR	5	1084.88	0.07	0.45	-536.73
Imperv + Income + distPark	4	1088.31	3.5	0.08	-539.69
Imperv + DemPCA + distPark	3	1099.16	14.36	0	-546.31
Imperv + PropNHW	3	1110.43	25.63	0	-551.94
Imperv + Income	3	1122.36	37.55	0	-557.91
Imperv	2	1124.26	39.46	0	-560
Park (binary)	2	1277.35	192.54	0	-636.54
PropNHW + distPark	3	1306.87	222.06	0	-650.16
distPark	2	1310.65	225.84	0	-653.19
DemPCA	2	1310.91	226.1	0	-653.32
Income	2	1319.71	234.91	0	-657.72
PropNHW	2	1325.11	240.3	0	-660.42
Canopy	2	1328.65	243.84	0	-662.19
DistSJR	2	1332.2	247.4	0	-663.97
PropPov	2	1332.6	247.79	0	-664.17
Loc. Div. index	2	1339.35	254.55	0	-667.54
Intercept	1	1341.09	256.28	0	-669.5
Gini index	2	1341.75	256.94	0	-668.74

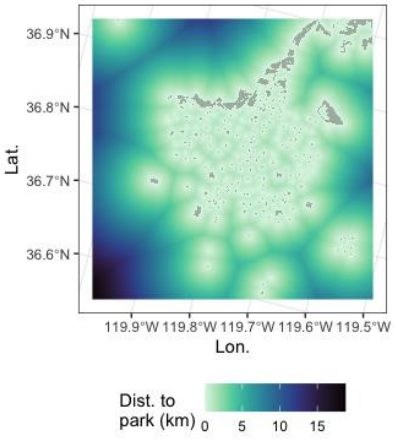
(a)


36.9°N
36.8°N
36.7°N
36.6°N
Latitude
119.9°W
119.8°W
119.7°W
119.6°W
119.5°W
Longitude
Med. Income ($1000)
0
40
80
120
% POC
100
75
50
25
0
Type
H
P

(b)


36.9°N
36.8°N
36.7°N
36.6°N
36.5°N
Lat.
120.0°W
119.8°W
119.6°W
119.4°W
Lon.
% Impervious cover
0
25
50
75
100

(c)


36.9°N
36.8°N
36.7°N
36.6°N
Lat.
119.9°W
119.8°W
119.7°W
119.6°W
119.5°W
Lon.
Dist. to park (km)
0
5
10
15

(d)

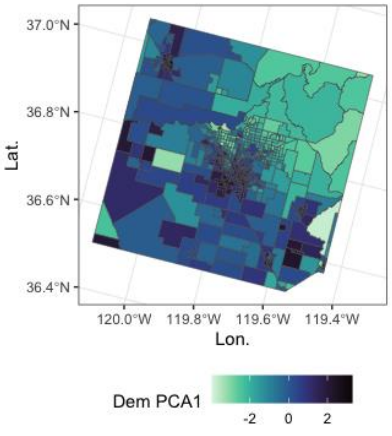

37.0°N
36.8°N
36.6°N
36.4°N
Lat.
120.0°W
119.8°W
119.6°W
119.4°W
Lon.
Dem PCA1
-2
0
2

Figure 2.1. (a) Map of study area in Fresno and Madera Counties, California, showing the study area boundary (black line), the San Juan River (blue line) separating the two counties, and parks (green). Map background shows the distribution of median household income (pink) and the proportion of people of color (blue), so that more saturated, overlapping regions reflect areas with lower median incomes and higher proportions of people of color. Points show eBird personal sampling locations (black) and hotspots (teal). (b) – (d) Selected environmental and census covariates considered in point process analysis of eBird sampling locations, on the original data scales: (b) the percentage of impervious surface cover; (c) distance to the nearest formal park or natural area; and (d) the first principal component from a principal components analysis to summarize variation across several demographic metrics.

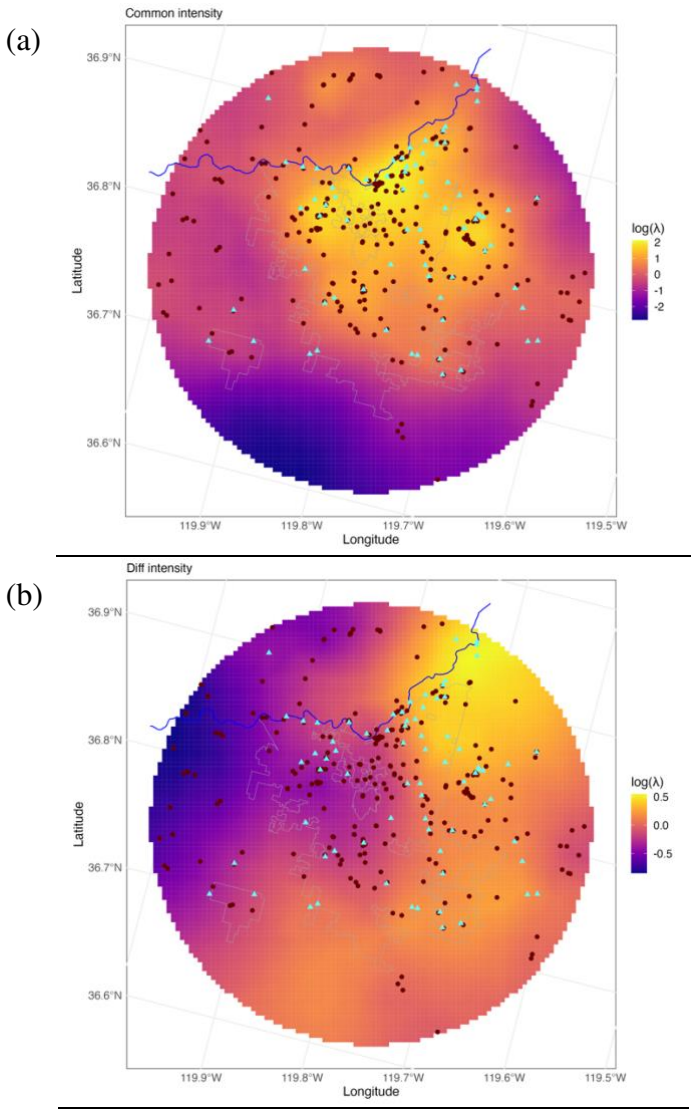


Figure 2.2. Spatial random effects predicted from the multiple likelihood point process model fit to eBird sampling locations in Fresno, CA that were visited in the spring between 2011 and 2019. Hotspots are shown as blue triangles, and all other eBird locations are represented as brown circles. (a) The estimated spatial effect that is shared between the hotspot and personal location intensities; (b) The estimated spatial effect underlying the intensity of hotspots that is not shared by personal locations not shared by the two intensities. Predictions were made using the model specification without covariates, so predicted spatial effects reflect all estimated spatial structure, including strucutre that is associated with spatial covariates.

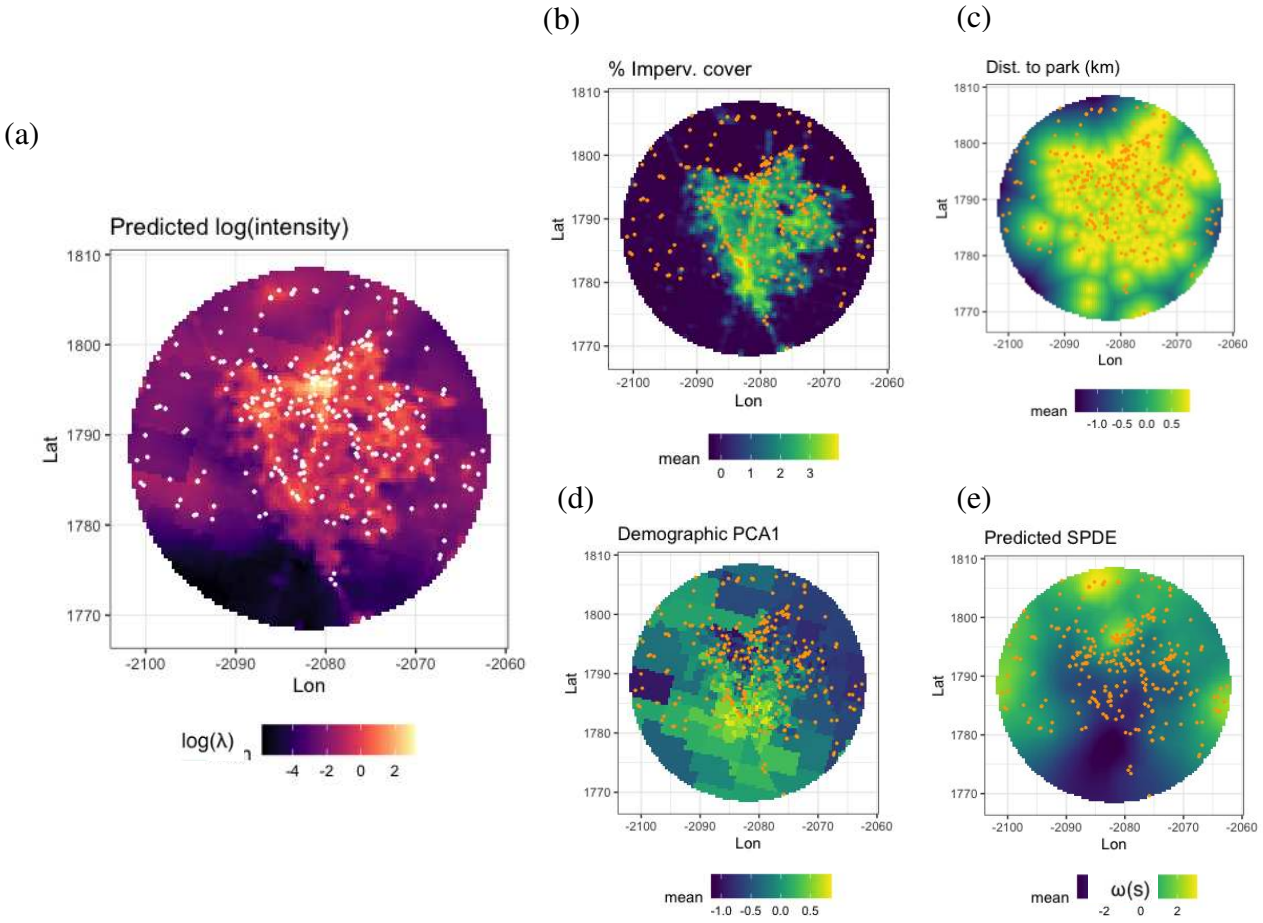


Figure 2.3. Model predictions from point process model (LGCP) fit to the set of all eBird locations (white or orange points) sampled in Fresno, CA in April-June 2011-2019 (n=338). (a) The intensity of eBird locations (shown on log-scale) was associated with (b) the percentage of impervious surface cover, (c) distance to the nearest park, and (d) the first principal component of a demographic PCA. Coefficients and predictions shown are relative to standardized covariate values. (e) Estimated spatial random effect reflects the remaining spatial structure (or autocorrelation) between locations that is not explained by fixed effects.

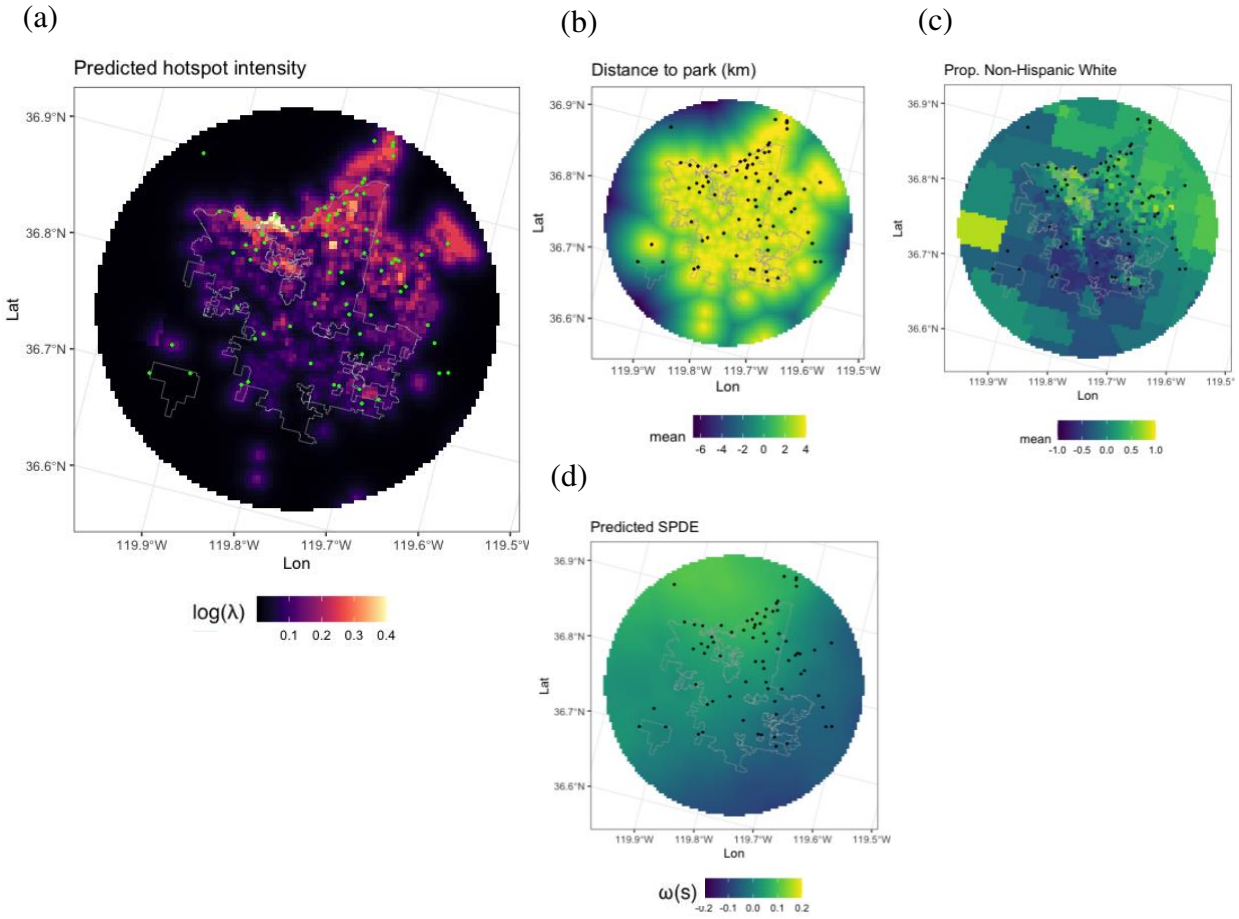


Figure 2.4. Model predictions from point process model of eBird hotspot intensity fit to the set of all hotspot locations in study area in Fresno, CA (n=73). (a) Predicted hotspot intensity (log-scale); (b) Estimated relationship between the distance to a park (km) and sampling intensity. Sampling intensity decreases as distance to a park increases. (c) Estimated relationship between the proportion of non-Hispanic white residents and sampling intensity. Areas with higher proportions of non-Hispanic white residents are associated with greater sampling intensity. (d) Estimated spatial random effect capturing the residual spatial structure in estimated sampling intensity that is not explained by the fixed effect. There is little evidence of remaining spatial autocorrelation between points after accounting for fixed effects.

REFERENCES

Adler, F. R., Green, A. M., and Şekercioğlu, Ç. H. (2020). Citizen science in ecology: a place for humans in nature. *Annals of the New York Academy of Sciences*, 1469(1), 52-64.

Altwegg, R., and Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, 10(1), 8-21.

Azevedo, D. R., Prates, M. O., and Bandyopadhyay, D. (2023). Alleviating spatial confounding in frailty models. *Biostatistics*, 24(4), 945-961.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6), 760-766.

Baker, F., Smith, C. L., and Cavan, G. (2018). A combined approach to classifying land surface cover of urban domestic gardens using citizen science data and high resolution image analysis. *Remote Sensing*, 10(4), 537.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). Hierarchical Modeling and Analysis for Spatial Data, second edition. Chapman and Hall/CRC.

Bittel, J. (2020) People called police on this black birdwatcher so many times that he posted custom signs to explain his hobby. Washington Post.

Blake, C., Rhanor, A. and Pajic, C. (2020). The Demographics of Citizen Science Participation and Its Implications for Data Quality and Environmental Justice. *Citizen Science Theory and Practice*. 5, 21.

Bocinsky, R. K., Beaudette, D., Chamberlain, S., and Bocinsky, M. R. K. (2015). 'FedData': Functions to Automate Downloading Geospatial Data Available from Several Federated Data Sources. R package version 3.0.4, <https://CRAN.R-project.org/package=FedData>.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2), 261-304.

Callaghan, C. T., Rowley, J. J., Cornwell, W. K., Poore, A. G., and Major, R. E. (2019a). Improving big citizen science data: Moving beyond haphazard sampling. *PLoS Biology*, 17(6), e3000357.

Callaghan, C. T., Major, R. E., Lyons, M. B., Martin, J. M., Wilshire, J. H., Kingsford, R. T., and Cornwell, W. K. (2019b). Using citizen science data to define and track restoration targets in urban areas. *Journal of Applied Ecology*, 56(8), 1998-2006.

Callaghan, C. T., Ozeroff, I., Hitchcock, C., and Chandler, M. (2020). Capitalizing on opportunistic citizen science data to monitor urban biodiversity: A multi-taxa framework. *Biological Conservation*, 251, 108753.

Carver, E. (2009). Birding in the United States: A demographic and economic analysis: addendum to the 2006 national survey of fishing, hunting, and wildlife-associated recreation. US Fish and Wildlife Service, Division of Economics.

Chamberlain, D., Reynolds, C., Amar, A., Henry, D., Caprio, E., and Batáry, P. (2020). Wealth, water and wildlife: Landscape aridity intensifies the urban luxury effect. *Global Ecology and Biogeography*, 29(9), 1595-1605.

City of Fresno (2022). Fresno City Council Districts. FresnoGIS Hub. https://gis-cityoffresno.hub.arcgis.com/. *Accessed 29 August 2022.*

Cliff, A. D., and Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1), 269-292.

Cole, D. (2020). Parameter Redundancy and Identifiability. CRC Press.

Cooper, C. B., and Smith, J. A. (2010). Gender patterns in bird-related recreation in the USA and UK. *Ecology and Society*, 15(4).

de Camargo Barbosa, K. V., Develey, P. F., Ribeiro, M. C., and Jahn, A. E. (2021). The contribution of citizen science to research on migratory and urban birds in Brazil. *Ornithology Research*, *29*, 1-11.

Dewitz, J., and U.S. Geological Survey, (2021). National Land Cover Database (NLCD) 2019 Products (ver. 2.0, June 2021): U.S. Geological Survey data release, https://doi.org/10.5066/P9KZCM54 *Accessed 16 December 2022*

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T., and Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment,* 10(6), 291-297.

Diggle, P. J., Menezes, R., and Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(2), 191-232.

Dinsdale, D., and Salibian-Barrera, M. (2019). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(1), 181-198.

Diprose, G., Greenaway, A., and Moorhouse, B. (2022). Making Visible More Diverse Nature Futures through Citizen Science. *Citizen Science: Theory and Practice*, *7*(1).

eBird. (2022). eBird: An online database of bird distribution and abundance [web application]. eBird, Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org.

eBird FAQ. (2023). eBird Hotspot Frequently Asked Questions (FAQs). Modified 23 March 2023. https://support.ebird.org/en/support/solutions/articles/48001009443-ebird-hotspot-faqs. *Accessed October 17, 2023.*

Ellis-Soto, D., Chapman, M., and Locke, D. H. (2023). Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nature Human Behaviour*, 7(11), 1869-1877.

Fong, E., and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2), 489-496.

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525), 445-452.

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B. O., Olsen, K., Rahbek, C., and Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11), 1139-1149.

Grade, A. M., Chan, N. W., Gajbhiye, P., Perkins, D. J., and Warren, P. S. (2022). Evaluating the use of semi-structured crowdsourced data to quantify inequitable access to urban biodiversity: A case study with eBird. *PloS One* 17(11), e0277223.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25, 835-862.

Grove, J. M., Locke, D. H. and O'Neil-Dunne, J. P. M. (2014). An Ecology of Prestige in New York City: Examining the Relationships Among Population Density, Socioeconomic Status, Group Identity, and Residential Canopy Cover. *Environmental Management* 54, 402–419.

Haklay, M. E. (2016). Why is participation inequality important? Ubiquity Press.

Hensley, C. B., Trisos, C. H., Warren, P. S., MacFarland, J., Blumenshine, S., Reece, J., and Katti, M. (2019). Effects of urbanization on native bird species in three southwestern US Cities. *Frontiers in Ecology and Evolution*, 7, 71.

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Lamigueiro, O.P., Bevan, A., Racine, E.B., Shortridge, A., and Hijmans, M. R. J. (2023) raster: Geographic Data Analysis and Modeling. R package version 3.6-23, <https://CRAN.R-project.org/package=raster>.

Hijmans, R. J., Bivand, R., Forner, K., Ooms, J., Pebesma, E., and Sumner, M. D. (2023b). 'terra': Spatial Data Analysis. R package version 1.7-46. https://CRAN.R-project.org/package=terra

Hoover, F. A., and Lim, T. C. (2021). Examining privilege and power in US urban parks and open space during the double crises of antiblack racism and COVID-19. *Socio-Ecological Practice Research*, 3(1), 55-70.

Hope, D., Gries, C., Zhu, W., Fagan, W. F., Redman, C. L., Grimm, N. B., ... and Kinzig, A. (2003). Socioeconomics drive urban plant diversity. *Proceedings of the National Academy of Sciences*, 100(15), 8788-8792.

Jimenez, M. F., Pejchar, L., and Reed, S. E. (2021). Tradeoffs of using place-based community science for urban biodiversity monitoring. *Conservation Science and Practice*, *3*(2), e338.

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., Kelling, S.T., and Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, 27(7), 1265-1277.

Jones, L. (2020). Black People Don't Go Outside: Impact of Stereotypes on the Black and African American Relationship with Nature (Doctoral dissertation, Alaska Pacific University).

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., …, and Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11), 1366-1379.

Kuldna, P., Poltimäe, H., and Tuhkanen, H. (2020). Perceived importance of and satisfaction with nature observation activities in urban green areas. *Journal of Outdoor Recreation and Tourism*, 29, 100227.

Kurnia, I., Arief, H., Mardiastuti, A., and Hermawan, R. (2021). Urban landscape for birdwatching activities. In IOP Conference Series: Earth and Environmental Science (Vol. 879, No. 1, p. 012005). IOP Publishing.

Leong, M., Dunn, R. R., and Trautwein, M. D. (2018). Biodiversity and socioeconomics in the city: A review of the luxury effect. *Biology Letters*, 14(5), 20180082.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4), 423-498.

Lopez B, Minor E, Crooks A. (2020) Insights into human-wildlife interactions in cities from bird sightings recorded online. *Landscape and Urban Planning* 196: 103742.

Mahmoudi, D., Hawn, C. L., Henry, E. H., Perkins, D. J., Cooper, C. B., and Wilson, S. M. (2022). Mapping for whom? Communities of color and the citizen science gap. *UMBC Faculty Collection*.

Mair, L. and Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PloS One*, *11*(1), e0147796.

Montanari, M., Jacobs, L., Haklay, M., Donkor, F. K., and Mondardini, M. R. (2021). Agenda 2030s, "Leave no one behind" in citizen science? *Journal of Science Communication*, 20(06), A07-A07.

Martin, A. L., Adams, A. E., and Stein, T. V. (2023). Equity, identity, and representation in outdoor recreation: 'I am not an outdoors person'. *Leisure Studies*, 1-14.

Mullenbach, L. E., and Baker, B. L. (2020). Environmental justice, gentrification, and leisure: A systematic review and opportunities for the future. *Leisure Sciences*, 42(5-6), 430-447.

Nugent, J. (2018). iNaturalist: citizen science for 21st-century naturalists. *Science Scope*, *41*(7), 12-15.

Paleco, C., García Peter, S., Salas Seoane, N., Kaufmann, J., and Argyri, P. (2021). Inclusiveness and diversity in citizen science. *The Science of Citizen Science*, 261.

Pandya, R. E. (2012). A framework for engaging diverse communities in citizen science in the US. Frontiers in Ecology and the Environment, 10(6), 314-317.

Pateman, R. M., Dyke, A., and West, S. E. (2021). The diversity of participants in environmental citizen science. Citizen Science: Theory and Practice.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

Pebesma, E., and Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Peddinenikalva, N. R. (2023). Residual analysis of spatial point process models using Bayesian Methods. Inlabru vignettes. https://inlabru-org.github.io/inlabru/articles/2d_lgcp_residuals.html.

Perkins, D. J. (2020). *Blind Spots in Citizen Science Data: Implications of Volunteer Biases in eBird Data*. North Carolina State University.

Peter, M., Diekötter, T., Höffler, T., and Kremer, K. (2021). Biodiversity citizen science: Outcomes for the participating citizens. *People and Nature*, *3*(2), 294-311.

Pocock, M. J., Chandler, M., Bonney, R., Thornhill, I., Albin, A., August, T., Bachman, S., Brown, P.M., Cunha, D.G.F., Grez, A., Jackson, C., Peters, M., Rabarijaon, N.R., Roy, H.E., Zaviezo, T., and Danielsen, F. (2018). A vision for global biodiversity monitoring with citizen science. In Advances in Ecological Research (Vol. 59, pp. 169-223). Academic Press.

Rigolon, A. (2016). A complex landscape of inequity in access to urban parks: A literature review. *Landscape and Urban Planning*, 153, 160-169.

Rigolon, A., Keith, S. J., Harris, B., Mullenbach, L. E., Larson, L. R., and Rushing, J. (2020). More than" Just Green Enough": Helping Park Professionals Achieve Equitable Greening and Limit Environmental Gentrification. *Journal of Park and Recreation Administration*, 38(3).

Robinson, O. J., Ruiz-Gutierrez, V., Reynolds, M. D., Golet, G. H., Strimas-Mackey, M., and Fink, D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, 26(8), 976-986.

Rosenblatt, C. J., Dayer, A. A., Duberstein, J. N., Phillips, T. B., Harshaw, H. W., Fulton, D. C., Cole, N.W., Raedeke, A.H., Rutter, J.D., and Wood, C. L. (2022). Highly specialized recreationists contribute the most to the citizen science project eBird. *Ornithological Applications*, 124(2), duac008.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319-392.

Rutter, J. D., Dayer, A. A., Harshaw, H. W., Cole, N. W., Duberstein, J. N., Fulton, D. C., Raedeke, A.H., and Schuster, R. M. (2021). Racial, ethnic, and social patterns in the recreation specialization of birdwatchers: an analysis of United States eBird registrants. *Journal of Outdoor Recreation and Tourism*, 35, 100400.

Schell, C. J., Dyson, K., Fuentes, T. L., Des Roches, S., Harris, N. C., Miller, D. S., Woelfle-Erskine, C.A., and Lambert, M. R. (2020). The ecological and evolutionary consequences of systemic racism in urban environments. *Science*, 369(6510), eaay4497.

Schleder, B. W. (2010). Residential irrigation as a driver of urban bird community structure. California State University, Fresno.

Schwarz, K. et al. (2015). Trees Grow on Money: Urban Tree Canopy Cover and Environmental Justice. *PLoS One* 10, e0122051.

Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics*, *42*, 100446.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science,* 32(1).

Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D., and Rue, H. (2019). Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(3), 543-564.

Steen, V. A., Tingley, M. W., Paton, P. W., and Elphick, C. S. (2021). Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, *12*(2), 216-226.

Strimas-Mackey, M., Miller, E., and Hochachka, W. (2018) auk: eBird Data Extraction and Processing with AWK. *R package version 0.3.0.* https://cornelllabofornithology.github.io/auk/

Stuber, E. F., Robinson, O. J., Bjerre, E. R., Otto, M. C., Millsap, B. A., Zimmerman, G. S., Brasher, M.G., Ringelman, K.M., Fournier, A.M., Yetter, A., Isola, J.E., … and Ruiz-Gutierrez, V. (2022). The potential of semi-structured citizen science data as a supplement for conservation decision-making: validating the performance of eBird against targeted avian monitoring efforts. *Biological Conservation*, 270, 109556.

Sullivan, B.L., C.L. Wood, M.J. Iliff, R.E. Bonney, D. Fink, and S. Kelling. (2009). eBird: a citizen-based bird observation network in the biological sciences. *Biological Conservation* 142: 2282-2292.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... and Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31-40.

Sullivan, B. L., Phillips, T., Dayer, A. A., Wood, C. L., Farnsworth, A., Iliff, M. J., ... and Kelling, S. (2017). Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, 208, 5-14.

The Trust for Public Land (2023). ParkScore 2023. https://www.tpl.org/parkscore (Accessed October 2023)

Theobald, D. M. (2014). Development and applications of a comprehensive land use classification and map for the US. *PloS One*, *9*(4), e94628.

Tivadar, M. (2019). "OasisR: An R Package to Bring Some Order to the World of Segregation Measurement." Journal of Statistical Software, 89 (7), 1-39. https://doi.org/10.18637/jss.v089.i07

Turner, W. R. (2003). Citywide biological monitoring as a tool for ecology and conservation in urban landscapes: the case of the Tucson Bird Count. *Landscape and Urban Planning*, 65(3), 149-166.

U.S. Census Bureau. (2020). "2015-2019 American Community Survey 5-Year Estimates." Washington, D.C.: US Census Bureau.

U.S. Geological Survey. (2019) 3D Elevation Program 1-Meter Resolution Digital Elevation Model (published 20211006). https://www.usgs.gov/the-national-map-data-delivery. *Accessed 20 September 2022*.

Vasquez, A. V., and Wood, E. M. (2022). Urban parks are a refuge for birds in park-poor areas. *Frontiers in Ecology and Evolution*, *10*, 1048.

Walker, K., and Herman, M. (2023). Tidycensus: Load us census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames. R package, version *1*.4.4.

Warren, P. S., Ryan, R. L., Lerman, S. B., and Tooke, K. A. (2011). Social and institutional factors associated with land use and forest conservation along two urban gradients in Massachusetts. *Landscape and Urban Planning*, 102(2), 82-92.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.

World Atlas (2023). Fresno, California. https://www.worldatlas.com/cities/fresno-california.html. *Accessed 22 September 2023.*

Young, B. E., Dodge, N., Hunt, P. D., Ormes, M., Schlesinger, M. D., and Shaw, H. Y. (2019). Using citizen science data to support conservation in environmental regulatory contexts. *Biological Conservation*, 237, 57-62.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*(8), 1-25.

CHAPTER THREE

SOCIAL SAMPLING BIAS IN CROWDSOURCED DATA INHIBITS LUXURY EFFECT RESEARCH

Introduction

Urbanization is a defining feature of the modern age. Most of the global population currently resides in cities, and the proportion is projected to increase to 68% by 2050 (Habitat, 2016). Urbanized landcover exercises substantial influence over the global landscape, with the rapid expansion that has occurred in the past century projected to continue until at least 2100 (Seto et al., 2011; Chen et al., 2020). Thus, as cities expand and urban human populations grow, cities will continue to have diverse impacts on surrounding ecosystems and biodiversity, including habitat loss, fragmentation, and changes to ecological dynamics (Liu et al., 2016). At the same time, cities support a large amount of biodiversity (Aronson et al., 2014), and modern ecological urban planning can elevate cities to important refuges for biodiversity amidst widespread landscape degradation related to industrialization and extraction (Ahern, 2012; Piano et al., 2017; Salbitano et al., 2017). Thus, broad-scale biodiversity data across cities and regions are necessary to understand general trends, impacts of urbanization on ecological dynamics, and inform effective urban ecological planning and conservation.

Cities are distinctive social-ecological systems characterized by unique social-ecological phenomena that dictate the ecological structure and functioning of urban ecosystems (McHale et al., 2015; Frank et al., 2017; Andersson et al., 2021). A robust understanding of urban ecological processes requires an interdisciplinary approach that recognizes the fundamental interconnections between ecological and evolutionary processes and societal, cultural, and

economic systems and legacies (McHale et al., 2015; Pickett et al., 2016). Rather than focusing on biotic communities and ecological dynamics characteristic of other ecosystems, within an urbanized context, conceptualizing urban ecosystems as holistic social-ecological systems characterized by socioeconomic, cultural, and ecological landscapes and heterogeneity enables a deeper understanding of the complexities of these systems to inform development towards a more sustainable future for people as well as biodiversity (Pickett et al., 2016). Sustainable urban design that originates from this perspective has the potential to simultaneously support biodiversity while also progressing other societal goals, like creating sustainable and just food systems, developing climate resiliency, and advancing environmental justice (Millard, 2010; Heymans et al., 2019). On the other hand, a growing literature is revealing how inequitable societal systems like systemic racism impact ecological and evolutionary processes (Schell et al., 2020) and how legacies of historical development and cultural values shape contemporary patterns of biodiversity (Grove et al., 2020). For example, historical residential segregation practices are associated with current distributions of tree canopy and other environmental attributes, such that neighborhoods historically targeted for disinvestment under discriminatory lending practices ("redlined" neighborhoods) continue to have less tree cover, more urban heat islands, and greater pollution than neighborhoods viewed favorably ("greenlined") under these systems (Schell et al., 2020).

The luxury effect has emerged in the urban ecological literature as a prominent hypothesis describing a pattern of higher biodiversity associated with greater socioeconomic status observed in many cities around the world (Hope et al., 2003; Schell et al., 2020). The luxury effect hypothesis was first proposed to explain a recurring pattern of higher plant biodiversity in more affluent neighborhoods in Phoenix, Arizona, U.S.A., possibly due to

wealthier households spending money to influence plant species assemblages and resource availability (Hope et al., 2003). The luxury effect hypothesis has been extended to explain associations between wildlife (particularly birds) and affluence (Loss et al., 2009; Leong et al., 2018; Schell et al., 2020), and support for the hypothesis in both plants and animals has been shown in many cities across the globe (Strohbach et al., 2009; Chamberlain et al., 2019; Sultana et al., 2022; Hassell et al., 2021). Meta-analyses have concluded that evidence for the luxury effect is more pronounced in arid regions, and regions with greater urbanization, vegetation loss, and wider wealth gaps (Leong et al., 2018, Chamberlain et al., 2020).

However, support for the luxury effect is not universal, with many studies demonstrating negative or no associations between wealth and biodiversity, and in general, there is a lack of understanding and investigation of causal social and political mechanisms behind observed patterns (Kuras et al., 2020; Magle et al., 2021). In particular, correlation or aggregation of many socioeconomic factors and across multiple relevant scales has prevented disentangling underlying drivers of the luxury effect (Kuras et al., 2020; Schell et al., 2020). For example, greater avian biodiversity is associated with newer neighborhoods in Chicago, IL (Loss et al., 2009), but newer neighborhoods also tend to be inhabited by affluent residents (Grove et al., 2014). Thus, it difficult to tease apart effects of a possible inherent difference in habitat quality between old and new neighborhoods versus affluent residents increasing resources available on the landscape. Further, it is not clearly known whether observed patterns of greater plant biodiversity and cover in affluent neighborhoods are driven by residents fostering this richness by increasing resources available to plants, as was originally suggested by Hope et al. (2003), or by municipalities favoring these areas for park development and neighborhood improvement (Kuras et al., 2020).

Like many studies of urban ecological processes, studies assessing the luxury effect often use voluntarily collected or crowdsourced data (also known as citizen science, community science, or volunteered geographic information) to observe biodiversity (e.g., iNaturalist, GBIF, UWIN, and numerous small-scale programs like the city of Fort Collins' Nature in the City). Crowdsourced biodiversity data platforms are well-suited for urban ecological research because local volunteers can contribute data across much larger geographic areas than would be feasible by a limited research team (Sullivan et al., 2014; Peeters et al., 2022). Thus, crowdsourcing can be a low-cost method for collecting data across many taxa, especially with advances in technologies like photo recognition and portable noise and video monitors, in urban areas that historically have received less ecological monitoring and research attention (Knapp et al., 2021). For example, eBird data have been instrumental in filling spatial and temporal gaps in other avian monitoring data, like the Breeding Bird Survey (BBS) in the United States, which avoids urban areas in its sampling routes (Weiser et al., 2020).

However, crowdsourced biodiversity data collection also comes with challenges related to data quality and sampling bias. Crowdsourced biodiversity data are heavily influenced by sampling bias, because participants select observation locations and tend to favor convenient, accessible, and aesthetic locations for collecting and submitting data (Kuldna et al., 2020; Johnston et al., 2021). As a result, crowdsourced data tend to overrepresent spaces near populated or trafficked areas, like near cities or within national parks (Kolstoe and Cameron, 2017; Fink et al., 2020). At smaller scales, local features like roads and parking lots are often overrepresented in data collection, and sampling bias also reflects social landscape heterogeneity (Mair and Ruete, 2017; Fink et al., 2020).

In urban areas, evidence has accumulated that sampling bias in crowdsourced biodiversity data collection is also associated with social, economic, and culture landscapes. Data collection for many crowdsourced projects and platforms, including eBird, iNaturalist, and the Global Biodiversity Information Facility (GBIF), consistently under-samples urban areas that are characterized by predominantly low-income communities or communities of color, bias data towards predominantly white and affluent areas (Perkins, 2020; Grade et al, 2022; Mahmoudi et al., 2022; Ellis-Soto et al, 2023). In GBIF records, patterns of sampling bias are not only associated with current patterns of socioeconomics, but also with historical residential segregation (Ellis-Soto et al., 2023). Thus, legacies of historical social landscapes affect both current urban biodiversity distributions and our methods for observing them, linking these two distinct processes in observed data, because sampling bias is deeply linked to the biological process under investigation. The prevalence of this social sampling bias challenges the validity of any findings that rely heavily on urban crowdsourced data without considering biased data collection (Grade et al., 2022; Ellis-Soto et al., 2023). For example, data gaps in low-income communities may underlie some of the variation in conclusions surrounding the luxury effect across different cities, since representation of a substantial portion of observed incomes is missing from data.

Sampling bias, including observer preference and skill, variable detectability, and spatial bias, has long been acknowledged as a challenge of crowdsourced data (Hughes et al., 2021). Many methods have been proposed to mitigate these different sources of sampling bias after collection, including data filtering (i.e., spatial subsampling, balancing), the inclusion of sampling and effort variables in analyses, and many model specifications, to aim to balance the benefits of large amounts of data and the drawbacks of sampling bias (Fink et al., 2020; Johnston

et al, 2021; Tang et al., 2021). These methods can be very effective for mitigating bias to produce large scale predictive distributions, augment data coverage from systematic data, and estimate temporal dynamics of migratory animals (Fink et al., 2020; Stuber et al., 2022). However, most strategies for addressing sampling bias in largescale crowdsourced data occur at the landscape scale, and often do not address sources of bias specific to smaller, urban scales (Planillo et al., 2021). In particular, crowdsourced data collection has an inherently social dimension, reflecting the social values and contexts that data are collected under (Sieber and Haklay, 2015; Haklay 2016; Mahmoudi et al., 2022), and to our knowledge, there are no existing methods for mitigating this specific, yet crucial, source of sampling bias.

One form of sampling bias particularly relevant for voluntary crowdsourced data collection is preferential sampling, which describes the case when observers are selecting sampling locations, because they expect them to be favorable for the species or process they are sampling (Diggle, 2010; Pennino et al., 2019). Thus, in the case of biodiversity sampling, the resulting data are likely to be biased toward areas with high occurrence or abundance of the species of interest. When left unconsidered, this can lead to overestimates of occurrence throughout the region, because the data lack the variation contained in less-abundant areas, but this is not always the case (Grade et al., 2022). When sampling preferences parallel covariates of interest so that one or more underlying covariates influence both the process being observed and the sampling process, results can either over- or underestimate true relationships between the observations and the covariates (Knox et al., 2020; Grade et al., 2022).

Statistical techniques to address preferential sampling have been applied in ecology to create species distribution models of blue and red shrimp (*Aristeus antennatus*) from fishing data that reflect fisherman preferences for areas with high shrimp abundance (Pennino et al., 2019).

They have also been applied to account for increased observer activity near roads in presence-only crowdsourced data for moose (*Alces alces*) occurrence (Sicacha-Parada et al., 2021), and for spatial patterns in observer effort and skill in eBird data (Tang et al., 2021).

Despite the growing body of literature documenting social sampling bias and the potential harmful consequences of making conservation decisions based on inequitable data, research has yet to explore how socially driven sampling bias impacts inference and prediction informed by crowdsourced data, or if existing data pre-processing or analytical methods can effectively mitigate this bias. This is particularly important – and challenging – considering the parallel and intertwining relationships between the social-ecological phenomenon being observed (e.g., the luxury effect) and the drivers of urban sampling bias. To reliably learn about urban ecological systems with crowdsourced data, we must be confident we can disentangle relationships between patterns of biodiversity and demographics from socially driven sampling bias.

As the largest crowdsourced biodiversity project globally, eBird currently contains over 100 million observations and participation grows by over 20% each year (Sullivan et al., 2009; Sullivan et al., 2014; ebird.org). Like other crowdsourced data programs, eBird sampling has been shown to overrepresent affluent, predominantly white neighborhoods in several U.S. cities (Perkins, 2020; Grade et al., 2022), and the legacies of historical discriminatory housing practices are associated with eBird survey completeness in cities (Ellis-Soto et al., 2023). In fact, among multiple similar projects, eBird sampling was found to be more unevenly distributed across historically A and D graded neighborhoods than iNaturalist (Ellis-Soto et al., 2023). eBird data has been used in urban ecological studies to assess relationships between bird biodiversity and urban parks (Callaghan et al., 2019a; LaSorte et al., 2020; LaSorte et al., 2023), residential

yards (Lerman et al., 2021), and urbanization and fragmentation (Callaghan et al., 2021; Soifer et al., 2021). Additionally, eBird data are regularly applied to inform species management, habitat protection, urban planning, and policy decisions (Sullivan et al., 2017; Callaghan et al., 2019b; Planillo et al., 2021). Thus, eBird data are regularly being used to inform decisions that impact both people and wildlife, though the vast majority of these studies do not consider demographics, inequities, or the social context of either the urban landscape or crowdsourced data collection in their research, leaving the potential impacts of social sampling bias unacknowledged. This not only risks false understanding of true biological pattern and relationships, but risks decision making that is informed by and reinforces inequities.

We conducted a study to assess the impacts of not correcting for social sampling bias on inference related to the luxury effect, as a starting point for understanding the impacts of confounded urban social sampling bias on research outcomes and conservation initiatives more broadly. Simultaneously, we explored whether random spatial subsampling as recommended by Johnston et al. (2021) mitigated social sampling bias. Using spring eBird data from Raleigh-Durham, North Carolina, U.S.A. collected between 2015 and 2019, we first quantified the luxury effect without considering spatially biased sampling using a spatial regression model to explore associations between social factors and species richness. Then, we investigated which factors predict the sampling intensity of eBird locations. Finally, we jointly modeled richness and sampling intensity to investigate the impact of accounting for sampling on estimated relationships between demographics and avian species richness. On one hand, since affluent neighborhoods tend to be oversampled, we expected that sampling effects may lead to over-estimated species richness in these neighborhoods, thus artificially inflating the observed strength of luxury effect. Alternatively, failure of biased sampling to collect a representative

sample across all levels of socioeconomics in the region may lead to finding no evidence of a luxury effect because eBird data capture do not capture the full range of variation in our response.

Methods

We conducted our study in the Raleigh-Durham ('Research Triangle') metropolitan area in North Carolina, U.S.A. We defined our study region as Wake and Durham counties, which encompass a total area of 2991 square kilometers in the Piedmont region of North Carolina. The region has a humid, subtropical climate characterized by a highly forested urban landscape. The predominantly deciduous forests support diverse animal communities. Wake and Durham counties make up the central, most densely populated area of the surrounding nine-county metropolitan region, that is one of the most rapidly growing and urbanizing regions in the United States. With this growth comes rapid gentrification, though the legacy effects of historical redlining and other racial segregation remain prominent in the housing landscape and distributions of environmental amenities (Perkins, 2020). As of the 2020 United States Census, Wake County has an overall population of 1,129,410 people, of which 57% are non-Hispanic white and 18% Black. The median household income in Wake County in was around $55,000 (U.S. Census Bureau, 2020). In contrast, Durham County has a total population of 324,833 people, of which 34% are Black and 41% are non-Hispanic white. The median household income in Durham County is $43,337 (U.S. Census Bureau, 2020). We chose this region of North Carolina for our study because previous work has investigated the relationship between socioeconomic status and eBird sampling in this area. Further, there is past and present work to systematically survey the avian communities across this urbanizing landscape. The Triangle Bird Count (TBC, Perkins, 2020) is an annual, volunteer-based systematic avian point count, modeled after the Tucson Bird Count, that has been conducted in the region since 2019 and offers

systematically sampled survey data for comparison to eBird data. Previous work has also described the impact of the urban-rural landscape gradient on the composition of bird communities in the area (Minor and Urban, 2010).

eBird data

We downloaded the eBird Basic Dataset (EBD) and sampling event data for Wake and Durham counties (eBird, 2022). We downloaded data for all species in these regions from 2015-2021, and used the 'auk' package in R to filter data prior to analysis (Strimas-Mackey et al., 2018). We filtered data to include only records from April 1 to June 30 across all years, to align with the TBC season and only included complete checklists with 'stationary' or 'traveling' protocols, durations less than 5 hours, distances less than 5 km, and up to 10 observers following eBird best practices outlined by Johnston et al (2021). We also removed all records for 2019 that were collected as part of the TBC, because these sampling locations were established as part of a systematic survey procedure, not chosen by eBird participants. We obtained TBC data from the TBC project manager on 14 July 2022 (trianglebirds.org) and removed associated checklists by cross-referencing recorded locations with Bird Count locations and filtering by eBird comments and location IDs. One assumption of our continuous point process modeling approach is that multiple observations cannot occur in identical locations (Banerjee et al., 2015). Thus, we applied a small random jitter to observation locations prior to analysis to remove duplicated locations. The magnitude of this jitter was smaller than the precision of our spatial data and the discretization mesh used to implement our models and thus had no impact on our results.

Additionally, we subsampled checklists prior to analysis to reflect the best practices outlined by Johnston et al. (2021). First, in instances in which a single observer submitted more than 10 checklists from the same location within a single year, we randomly retained 10

checklists for that observer-location-year combination. Second, we spatially subsampled the eBird data by gridding the study region using a 5 kilometer (km) hexagonal grid (R package 'sf', Pebesma, 2018). Next, we randomly retained one checklist per grid cell throughout the study region to create our analysis dataset. We similarly created subsampled data sets using 3km and 8km hexagonal grids to assess the impact of grid size on the resulting subsampled data.

Spatial and social data

As the emphasis was on the luxury effect, we did not consider an exhaustive set of environmental covariates. Instead, we included percent canopy cover and percent impervious surface cover from the National Landcover Dataset (NLCD) and a set of metrics derived from U.S. Census American Community Survey (ACS) 5-year estimates at the block group level for the period covering 2015-2019 (US Census Bureau, 2020) as possible covariates. We obtained percent impervious surface cover from the 2019 NLCD release and percent canopy cover from the 2016 NLCD release (Dewitz, 2021) using the 'FedData' R package (Bocinsky, 2023). We included percent canopy cover and percent impervious surface cover as the mean percent within a 1km buffer around observation locations. We conducted all spatial analyses in an Albers Equal Area projection in R using the 'sf,' 'raster,' and 'terra' packages (Pebesma, 2018; Pebesma and Bivand, 2023; Hijmans, 2023(a); Hijmans, 2023(b)).

The U.S. Census ACS 5-year estimates (https://www.census.gov/data/developers/data-sets/acs-5year.html) are obtained by analyzing all surveys collected over a 5-year time period (in our case, 2015-2019) to produce a single "period estimate," with associated uncertainty, rather than "point-in-time" estimate for each metric. Including samples over five years allows for a larger sample size and thus greater precision estimates, while corresponding to the 5 years of

eBird data we included. Census data were downloaded, visualized, and processed via the R package 'tidycensus' (Walker and Herman, 2023).

We included race/ethnicity, median household income, and median building age in our set of possible demographic variables. For race, we obtained estimates of the number of people within the block group that identified as each race and calculated the proportion of the population in each race by dividing the estimate by the estimate of total population within the block group. The U.S. Census reports the following races: Asian, Black, Hawaiian/Pacific Islander (HIPI), American Indian/Alaskan Native (Native), White, Single Other, and Two or More Races and Hispanic/Latino ethnicity (U.S. Census Bureau, 2020). From these, we disaggregated White into non-Hispanic White and Hispanic White. We included race in our model as two possible variables: the proportion of Black, Indigenous, and People of Color (BIPOC), defined as one minus the proportion of non-Hispanic White (NHW) residents, and the proportion of Black residents. Non-Hispanic White residents are the most represented racial group in our study area and historically have held the most power in residential zoning and municipal decision (Whittemore, 2018), so we considered proportion of BIPOC residents to capture all races and ethnic groups outside of this dominant group, and we considered the proportion of Black residents specifically because Durham is a historically Black city and the legacy effects of residential segregation policies disproportionately impact Black communities in this area (Whittemore, 2018). Additionally, we computed the localized racial diversity index and local Shannon diversity of race to reflect residential segregation using 'OasisR' (Tivadar, 2019) and included these as possible predictor variables. We centered and scaled all predictor variables to have a mean of 0 and a standard deviation of 1 prior to analysis to assist with model fitting and to enable comparison of estimated effect sizes across predictors.

Sampling intensity and spatial subsampling

To evaluate the efficacy of our spatial balancing procedure, we specified a point process model to estimate sampling intensity and test relationships between intensity and demographic covariates before and after spatial subsampling. We specified this model with the following log-Gaussian Cox Process (LGCP) intensity,

$$\log(\lambda(\boldsymbol{s})) = \beta_0 + \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \omega(\boldsymbol{s}), \quad (1)$$

where β_0 is an intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{x}(\mathbf{s})$ is a vector of spatial covariates at location $\boldsymbol{s}$, standardized to mean 0, standard deviation 1 for computational efficiency and ease of interpretation of importance across covariates. The Gaussian random field $\omega(\boldsymbol{s})$ captures spatial structure in the intensity of locations not described by the spatial covariates and was implemented as a spatial partial differential equation (SPDE) random effect, using a Matérn covariance function with penalized complexity priors on the hyperparameters in INLA (Fuglstad et al., 2016; Simpson et al., 2017). We fit all models to eBird data under an approximate Bayesian inferential framework using integrated nested Laplace approximation via the R-INLA and 'inlabru' R packages (Rue et al., 2009; Lindgren et al., 2011; Bachl et al., 2019).

We fit this model to the un-subsampled dataset and the subsampled data that was sampled on the 5km hexagonal grid. For each data set, we fit the model with an SPDE random effect and no fixed effects, and then as a set of single covariate models, including each of the following variables: race, median household income, median housing age, and local racial diversity index, and these with additive canopy or impervious cover. We used marginal log likelihood to compare model fit for each dataset and determine the most supported covariates (Fong and Holmes, 2020). We visually inspected predictive sampling intensities and compared the

estimated coefficients associated with the fixed effects to assess the impact of the spatial subsampling on estimated sampling intensity and social sampling bias. We also fit the model with no fixed effects and with the most supported fixed effects to data subsampled over the 3km and 8km hexagonal grids to investigate how our results varied across different subsampling resolutions.

Luxury effect and preferential sampling analysis

We next fit a spatial regression with an SPDE spatial random effect to our spatially subsampled data (5km grid) to explore associations between our biological response variable of interest, species richness, and our set of environmental and demographic predictor variables. We computed species richness as the total number of species observed for a checklist, using all species presences recorded for each checklist, and used a square root transformation of richness as the response variable, which we modeled with a Gaussian likelihood. We first constructed a model set to test additive combinations of the eBird sampling effort variables as suggested by eBird best practices (Johnston et al., 2021) and a spatial random effect to capture residual spatial structure in species richness that is not explained by the included covariates. We used the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) and the Watanabe Akaike Information Criterion (WAIC, Watanabe, 2010) to select the most supported combination of effort variables, opting not to include all variables together because of multicollinearity between effort variables.

After we established the most informative effort variables, we constructed a second model set to test the relationships between our environmental and demographic covariates and species richness, maintaining the spatial random effect and the selected effort variables throughout all models. We again used DIC and WAIC to establish the most supported model

among this set. We used the *predict* function in 'inlabru' to predict richness at all points across the discrete mesh, holding the effort variables constant (Bachl et al., 2019).

Finally, we combined the sampling point process model and the species richness regression under a single inferential model to jointly model sampling and avian biodiversity. We used a model similar to the preferential sampling models developed by Diggle (2010) and Pati et al. (2011) and implemented by Sicacha-Parada et al. (2021) and Pennino et al. (2019) to achieve this. In this model, each response variable (species richness and observation location) has its own likelihood, which are linked together via a shared spatial random effect and shared covariates. We specified this model as,

$$\boldsymbol{s}_i = \frac{\lambda(\boldsymbol{s}_i|\boldsymbol{\beta}_s)}{\tilde{\lambda}(S|\boldsymbol{\beta}_s)}, \quad i=1,\ldots,n, \tag{2}$$

$$\log(\lambda(\mathbf{s})) = \beta_0 + \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}_{\boldsymbol{s}} + \boldsymbol{\omega}(\boldsymbol{s}) + \boldsymbol{\nu}(\boldsymbol{s}), \tag{3}$$

$$\sqrt{y(\boldsymbol{s}_i)} \sim \mathrm{N}(\mu(\boldsymbol{s}_i), \tau), \tag{4}$$

$$\mu(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta}_{\boldsymbol{y}} + \alpha\boldsymbol{\omega}(\boldsymbol{s}), \tag{5}$$

$$\boldsymbol{\omega}(\boldsymbol{s}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \tag{6}$$

$$\boldsymbol{\nu}(\boldsymbol{s}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}}), \tag{7}$$

where $\lambda(\boldsymbol{s})$ is the intensity of the sampling locations $\boldsymbol{s_i}$ (i=1, …, n) and is modeled as the linear combination of an intercept, β_0, spatial covariates $\boldsymbol{x}(\mathbf{s})$ multiplied by a vector of regression coefficients, $\boldsymbol{\beta}$, and Gaussian random fields $\boldsymbol{\omega}(\boldsymbol{s})$ and $\boldsymbol{\nu}(\boldsymbol{s})$. The random field $\boldsymbol{\omega}(\boldsymbol{s})$ captures the spatial structure in the intensity not described by covariates that is shared by the sampling intensity and species richness. The field $\boldsymbol{\nu}(\boldsymbol{s})$ captures residual spatial structure in sampling intensity that is not shared with the estimated species richness process. The square root of species richness *y* at location $\mathbf{s_i}$ arises from a normal distribution with mean μ and precision τ. The mean, μ, is defined as a linear combination of the spatial covariates and associated regression

coefficients $\boldsymbol{\beta}_y$ (which are distinct from the coefficients $\boldsymbol{\beta}$ in the point process likelihood) and the shared spatial random field $\boldsymbol{\omega}(\boldsymbol{s})$. The parameter α is weight parameter that controls the strength of the effect of the sampling process on the response. Each of the Gaussian random fields was specified with a Matérn covariance function with penalized complexity priors on the hyperparameters (Fuglstad et al., 2016; Simpson et al., 2017). Hyperparameters were specified with non-overlapping distributions on the range parameters to aid identifiability between the random effects. τ was specified as arising from the default Gamma prior. Due to computational challenges and the scope of the study, we did not consider all covariates or combinations of covariates in the preferential sampling model. Instead, we fit the model with no covariates, with effort covariates in the richness likelihood only, and including the covariates that were supported in the separate models in the respective likelihoods. We assessed the sensitivity of the model to mesh and prior specification by refitting the model using a sequence of mesh sizes and hyperparameter values.

Results

We included 10,253 complete checklists submitted to eBird in Wake and Durham counties in spring 2015-2019 in our analysis (Fig. 3.1). Random spatial subsampling resulted in the retention of 3,201 of these checklists. Average species richness across checklists was 18 species (range [1, 78]). The range of median household incomes represented by the locations of the eBird observations was [$22,000, $232,955] (Fig. 3.2). However, the overall range of median household incomes present in the study area was [$13,856, $232,955]. Similarly, the eBird data occur in neighborhoods that are composed of a maximum of 86% Black residents (Fig. 3.2). However, some block groups within the study area are up to 97% Black. Thus, the available

eBird data do not contain information about bird communities in the lowest income or predominantly Black neighborhoods (Fig. 3.3).

Spatial subsampling did not mitigate bias

We found evidence of a relationship between sampling intensity and median household income before and after spatially subsampling the data. Estimated sampling intensity predicted from both the un-subsampled and spatially subsampled data show substantial heterogeneity in sampling intensity across space (Fig. 3.4). The most supported model for sampling intensity for both the un-subsampled and subsampled data included median household income as the only fixed effect (Table 2.1). The model did not converge across a range hyper-parameter and initial values when fit with covariates to the full un-subsampled data. Therefore, we estimated sampling intensity of this full set using only the model with a spatial random effect and no fixed effects, and (non-spatially) randomly sampled n=3,201 observations to create un-subsampled data to fit to the models with covariates. The estimated coefficient associated with income for the model fit to this randomly sampled subset was $\hat{\beta} = 0.309$ (95% credible interval [0.208, 0.412], Fig. 2.4B). The estimated income coefficient for the spatially subsampled model was $\hat{\beta} = 0.441$ (95% credible interval [0.341, 0.544], Fig. 3.4A). Fitting the sampling model to data subsampled at different resolutions (3km, 8km) yielded similar results. Thus, spatially balancing the eBird dataset by randomly spatially subsampling the data following the analytical guidelines (Johnston et al., 2021) improved computational stability but did not reduce social sampling bias.

Evidence of inverse luxury effect in spatially sub-sampled data

Variables related to sampling effort were the most predictive of species richness, and we found minimal evidence that environmental or census variables explained variation in richness after accounting for effort. The square root of the duration of an eBird sampling event, the square

root of the distance traveled during a traveling checklist, and whether the checklist occurred at a hotspot or personal location were the most important effort variables associated with species richness ($\widehat{\beta_{dist}} = 0.510$, 95% CI (0.440, 0.579); $\widehat{\beta_{dur}}$ = 0.210, 95% CI (0.197, 0.224); $\widehat{\beta_{hot}} = 0.660$, 95% CI(0.550, 0.770)) and adding additional effort variables did not increase the amount of variation explained. There was evidence of a small negative association between income and species richness ($\hat{\beta}$ = -0.082; 95% CI(-0.136, -0.028)), which only slightly improved model fit based on WAIC and DIC (Table 3.2). The estimated range and standard deviation of the spatial random effect were $\hat{\rho} = 1.08$ (0.680, 1.60) and $\hat{\sigma}$ =1.16 (0.866, 1.54), respectively, indicating very localized residual spatial structure likely reflecting clusters of checklists at the same locations. The inclusion of additional environmental or census variables did not improve model fit based on WAIC and DIC, and the 95% credible intervals for estimated coefficients covered 0 for all other possible covariates (Fig. 3.5A).

However, limiting our analysis to stationary eBird checklists only and omitting effort variables from the model resulted in a larger magnitude negative relationship between income and species richness. Further, when extending our study period to also include data collected in 2020 and 2021, a negative relationship between the proportion of Black residents and species richness also emerged. We did not explore these patterns further, though note that they suggest there may be additional relationships between sampling effort, demographics, and species richness. In particular, we did not ultimately include data from 2020 or 2021 in our analysis because the COVID-19 pandemic led to anomalous data collection for both the U.S. Census sampling and eBird sampling (Crimmins et al., 2021; Hochachka et al., 2021; Velkoff and Hartley, 2022).

Preferential sampling model

Fitting the preferential sampling model to the eBird data without inclusion of any covariates showed evidence of preferential sampling (i.e., shared spatial structure) between eBird sampling and species richness ($\hat{\alpha} = 2.060, 95\%\ \mathrm{CI}(1.738, 2.405)$). Including effort variables in the richness likelihood reduced the strength of preferential sampling by about 25% ($\hat{\alpha} = 1.44$, 95% CI (1.096, 1.788)), suggesting some dependence between effort and the spatial distribution of sampling locations. Essentially, this suggests that sampling intensity is an important predictor of species richness, and more densely sampled regions are associated with higher species richness. The spatial random effect associated with preferential sampling had estimated range $\widehat{\rho_\omega} = 2.29$ (1.485, 3.351) and standard deviation $\widehat{\sigma_\omega} = 0.58$ (0.469, 0.708), reflecting a residual dependence between sampling and species richness at a very small spatial distance, likely due to heavily clustered observations. The spatial random effect capturing the remaining spatial structure in sampling that was independent of species richness had estimated parameters $\widehat{\rho_v} =$ 7.25 (5.775, 9.158) and standard deviation $\widehat{\sigma_v} = 1.24$ (1.070, 1.437), suggesting autocorrelation in sampling observations at a larger spatial scale than preferential sampling.

Including median household income in both the sampling and richness process models (Fig. 3.3, 3.5) indicated a positive relationship between income and sampling intensity ($\hat{\beta} = 1.101$, 95% CI (0.868, 1.435), Fig. 3.5B). There was no estimated relationship between income and species richness (95% CI (-0.066, 0.043)); however, including income as a fixed effect in both the sampling and richness likelihoods reduced the estimated weight of the shared spatial random effect by about 20% ($\hat{\alpha} = 1.123$, 95% CI (0.812, 1.098)). Therefore, median household income explains some of the spatial dependence between sampling intensity and richness, because including it as a covariate in the model reduced some of the unexplained shared

dependence, despite not being an informative predictor of species richness (Fig. 3.6). Including income as a covariate in the sampling and species richness likelihoods did not have a meaningful effect on the estimated spatial random effects ($\widehat{\rho_\omega} = 1.888$ (1.13, 2.87), $\widehat{\sigma_\omega} = 0.807$ (0.606, 1.03); $\widehat{\rho_\nu} = 6.202$ (5.01, 7.55), $\widehat{\sigma_\nu} = 1.747$ (1.489, 2.06)). Altogether, these results suggest that sampling intensity is positively associated with median household income, and when this is included in the model for species richness, there is no longer a relationship between richness and income that would suggest the presence of the luxury effect or its inverse in Raleigh-Durham, N.C.

Discussion

We found evidence of a small inverse luxury effect (i.e., income negatively associated with species richness) in spring eBird data in Raleigh-Durham, North Carolina between 2015-2019, before including the effects of sampling bias on observed richness. By combining models for sampling and species richness, we found spatial dependence between sampling and species richness that indicates that patterns of sampling intensity are positively associated with bird species richness. Sampling intensity was also positively associated with median household income, and there was no longer evidence of an inverse luxury effect when the relationship between sampling and income was reflected in the model for species richness.

Although not the primary focus of this study, our finding that random spatial subsampling did not mitigate socially biased sampling is notable because it underscores the importance of specific, tailored techniques for addressing social sampling bias analytically and moreover, the importance of correcting the biases in data collection that lead to socially biased data in the first place. Emphasizing subsampling data after collection to mitigate bias centers areas that are overrepresented in data, which assumes that there is adequate but more sparse data elsewhere and fails to address the implications and causes of where data is missing. As Steen et al. (2021)

point out, using subsampling techniques in instances like this where there are a mix of common and more rare observations, it is important to consider retaining rare observations. Otherwise, subsampling schemes can be ineffective or can influence results (Steen et al., 2021).

In cases like urban eBird data where sampling reflects social geography, a focus on mitigating oversampled areas rather than addressing sampling gaps centers privileged neighborhoods, and reflects the idea that contributory crowdsourced biodiversity data and platforms are created by *and for* affluent, white people who have the financial means, background, and leisure time to collect bird observations (Sieber and Haklay, 2015; Mahmoudi et al., 2022). The inefficacy of subsampling for reducing sampling bias highlights the importance of considering what areas and demographics are missing from the dataset and the impacts of the absence of sampling in some neighborhoods, which cannot be remedied via subsampling the data.

The absence of a luxury effect in our results may reflect the true pattern of avian biodiversity in Raleigh-Durham, or may be due to biased data collection. Studies of the luxury effect often occur in arid climates, where supplemental watering may play a large role in observed patterns of biodiversity and trends towards higher biodiversity in areas with greater financial resources (Kuras et al., 2020). Thus, the humid climate and high forest cover in Raleigh-Durham may mediate the impacts of income on vegetation cover and subsequent avian biodiversity. Alternatively, or in combination, the area's unique history and patterns of development may explain the observed pattern between species richness and income, as was found in the distribution of trees in Baltimore, Maryland. Shifting priorities among powerful communities over time have resulted in heterogeneity in canopy cover that reflects the opposite of what would be expected under the luxury effect (Grove et al., 2020). Previous work in the

Research Triangle region found that bird communities were more strongly influenced by landscape level environmental factors than local factors, and composition but not richness varied between predominantly urban and rural landscapes (Minor and Urban 2010). Thus, future work could investigate luxury effect using data from the systematically sampled Triangle Bird Count and explore avenues to integrate TBC and eBird data to investigate luxury effect.

However, the portions of the cities represented in the eBird data have a narrower range of median household incomes than is present in the study area overall, leading to limited representation of the true heterogeneity in income across the landscape. Thus, the luxury effect (or its inverse) may be more pronounced in Raleigh-Durham than we observe with crowdsourced data, because existing crowdsourced data do not capture the lower tail of the income gradient. Comparing the range and distribution of incomes reflected in the eBird data (before and after spatial subsampling) with the distribution of income in the region overall illustrates the importance of considering the implications of participation inequality and under-sampled neighborhoods, as our results are not reflective of Raleigh-Durham overall (Fig. 3.3). Thus, findings related to the luxury effect informed by eBird data must be interpreted with caution, and continued work in this same study area should explore the impact of integrating systematic Triangle Bird Count data to fill in data gaps and explore alternative subsampling schemes that mitigate oversampling while retaining all observations from under-sampled neighborhoods (Steen et al., 2021).

In contrast to the evidence of a negative relationship between median household income and species richness reflected in the spatial regression (no sampling model), there was no evidence for a relationship between income and species richness when richness and sampling were modeled jointly in the preferential sampling specification. However, observations are more

likely in areas with greater species richness (i.e., preferential sampling), and the unexplained spatial dependence between sampling intensity and richness decreased when income was included in both likelihoods. Thus, we found that eBird sampling in the Research Triangle is biased toward areas with higher median household incomes, and this biased sampling subsequently influences the estimated relationship between income and species richness. In our case, the estimated inverse luxury effect disappeared when accounting for sampling inequity, corroborating concerns raised by Grade et al. (2022), Mahmoudi et al. (2022), and Ellis-Soto et al. (2023) regarding the validity of results from crowdsourced biodiversity data that do not account for inequitable sampling.

Although this study has extended the work on social sampling bias and its implications in crowdsourced biodiversity data, considerable additional work is needed to explore alternative modeling approaches and try to resolve ongoing identifiability, confounding, and computational issues, both specific to this study and related to preferential sampling and complex crowdsourced sampling processes more generally. Spatial confounding between the fixed and random spatial effects remains a challenge, especially when combined with the extent of covariation between our two primary processes of interest and in our spatial covariates. Several methods have been proposed to mitigate spatial confounding, including restricted spatial regression or transformed Gaussian Markov Random field models (Hanks et al., 2015; Urdangarin et al., 2022; Fink et al., 2023), and continuing to explore these approaches in the INLA context is worthwhile. Another potential approach for alleviating some of the computational and identifiability issues is respecifying the model as a discretized LGCP model (Sørbye et al., 2019) or a conditional autoregressive model (Banerjee et al., 2015). A discretized approach may facilitate more straightforward interpretation because it reflects the areal structure of the demographic data, and

may be better suited to alleviating confounding via prior specification in INLA than the continuous case (Sørbye et al., 2019). The drawbacks of a discrete approach are primarily concerned with the influence of the discretization scheme and alignment of different spatial data on the results, because given the resolution of the eBird data and the irregular census block group areal units.

We chose to model sampling intensity as a continuous process because aligning the several spatial datasets would be challenging. eBird observation locations are reported as point locations, but the precision of these locations may vary and many checklists reflect observations collected while the observer is walking (up to 5km) though are reported as a single location. Further, census block groups are delineated such that demographic variation within a block group is minimized and population size is relatively consistent across block groups (U.S. Census Bureau, 2020). Therefore, block groups vary substantially in area; in our study region, block group area ranged from 0.13 km^2 to 87.6 km^2 (median 4.92 km^2), and block group area was correlated with other demographic variables. As a result, some block groups are smaller than the maximum possible distance encompassed by a single eBird checklist, and the overlap between block groups and the hexagonal grid used to subsample the data is not consistent throughout the study area. We made a simplifying assumption that these data could be combined at the precision available, though acknowledge that scale and precision misalignment may influence results.

Further, there are several extensions to our current modeling approach that will be important for gaining a fuller understanding of the relationships between sampling dynamics and resulting data and inference. First, given evidence that hotspots and personal locations are driven by different underlying sampling intensities and factors, our future work will include an investigation of how relationships between estimated species richness and sampling vary at these

two types of eBird locations. Additionally, using species richness as our response variable was a useful starting point for beginning to understand the dynamics between sampling and eBird observations, but considering alternate response variables, such as richness of native species only and single-species occupancy, may reduce some noise in the dataset and refine the relationships between avian biodiversity and environmental factors that we seek to understand. Next steps also include working to integrate eBird data with systematic survey data as available, including in Raleigh-Durham. Finally, given the wide variability in eBird participation and findings related to the luxury effect, future work should explore a larger set of cities with different demographics and climates, including extending work beyond cities in the United States.

While continuing to investigate analytical approaches to alleviate social sampling bias in existing crowdsourced biodiversity data is worthwhile to a point, addressing inequitable participation and systemic under-sampling in program design and data collection is ultimately the most important future direction for crowdsourced biodiversity data. It has long been acknowledged that sound sampling design is preferable to analytically accounting for sampling effects across a range of sampling methodologies (Cochran, 1977; Albert et al., 2010), because results can still be biased by sampling even after statistically accounting for sampling (Edwards et al., 2006; Irvine et al., 2018). Though there are a wealth of quantitative techniques for accounting for sampling and observational processes in ecological data, minimizing complexity through design is optimal wherever possible (Conn et al., 2017; Williams and Brown, 2019). Given increasing prevalence and reliance on contributory science platforms and crowdsourced data for biodiversity monitoring, conservation, and management, it is irresponsible and counterproductive to continue to ignore participation inequality and its implications for social justice and data quality (Haklay, 2016; Mahmoudi et al., 2022). Ultimately, advanced statistical

techniques cannot help us in cases where we have consistent patterns of missing data, especially when these patterns mirror the social-ecological dynamics we seek to understand (Grade et al., 2022).

Going forward, it is critical that we understand the consequences of societal inequity, for people and wildlife, so that we can understand how systems that oppress and disenfranchise people also work against sustainability and biodiversity conservation (Schell et al., 2020). However, we cannot gain this understanding when sampling reflects similar inequity and fails to capture the full range of outcomes and relationships present on the landscape. Then, sampling bias reinforces inequity by biasing knowledge creation and informed conservation decisions toward privileged neighborhoods, resulting in a feedback loop that perpetuates environmental injustice and inequitable access to environmental amenities and ecosystem services (Haklay, 2016; Montanari et al., 2021). Assuming participant identity does not influence data overlooks meaningful social processes in contributory science in all settings, but is especially limiting in urban landscapes where histories of systemic segregation and inequity strongly influence ecological landscapes and processes (Schell et al., 2020). Improving equity in our research methods is vital for understanding urban ecological processes, conserving urban biodiversity, and addressing the distributional and environmental inequities that are realized in our physical landscapes.

Tables and figures

Table 3.1. Marginal log-likelihoods (MLL) for LGCP point process model of sampling intensity fit to a spatially subsampled set of eBird checklists (n=3,201) collected between April-June 2015-2019 in Wake and Durham counties, N.C. All models included a spatial random effect. The most supported model also included a fixed effect of median household income.

Model fixed effect(s)	Marginal log likelihood (MLL)
Median household income	-23077.29
Spatial random effect only	-23109.19
Localized diversity index	-23109.2
Prop. impervious cover	-23110.93
Med. housing age	-23113.84
Prop. canopy cover	-23115.05
Prop. Black residents	-23115.62
Prop. non-Hispanic white residents	-23115.68
Income + impervious cover	-23116.53

Table 3.2. Deviance information criterion (DIC), Watanabe Akaike information criterion (WAIC), the associated estimates of effective numbers of parameters (Eff. Pars.), and marginal log-likelihood (MLL) for spatial regression models of avian species richness fit in INLA to spatially subsampled eBird data collected between April-June 2015-2019 in Wake and Durham counties, N.C. 'Effort' includes the fixed effects of the duration of sampling (minutes) and the distance traveled (km) during sampling. All models also include a spatial random effect to account for spatial autocorrelation. The most supported model included the fixed effect of median household income, the effort variables, and the spatial random effect.

	DIC	Eff. Pars.	WAIC	Eff. Pars.	MLL
Income + effort	8645.64	340.36	8645.26	299.5	-4504.55
Loc. div. index + effort	8648.03	342.31	8647.6	300.95	-4506.98
Prop. impervious + effort	8649.12	342.93	8648.23	301.07	-4507.67
Effort only	8649.47	342.11	8648.87	300.66	-4501.99
Prop. canopy + effort	8649.55	342.83	8648.93	301.19	-4507.54
Prop. non-Hispanic white residents + effort	8649.99	342.7	8649.38	301.12	-4508.45
Prop. Black residents + effort	8650.26	342.93	8649.65	301.28	-4508.71
Median housing age	8650.45	342.8	8649.83	301.18	-4509.02
Spatial random effect only	10231.12	407.09	10244.23	360.91	-5344.91

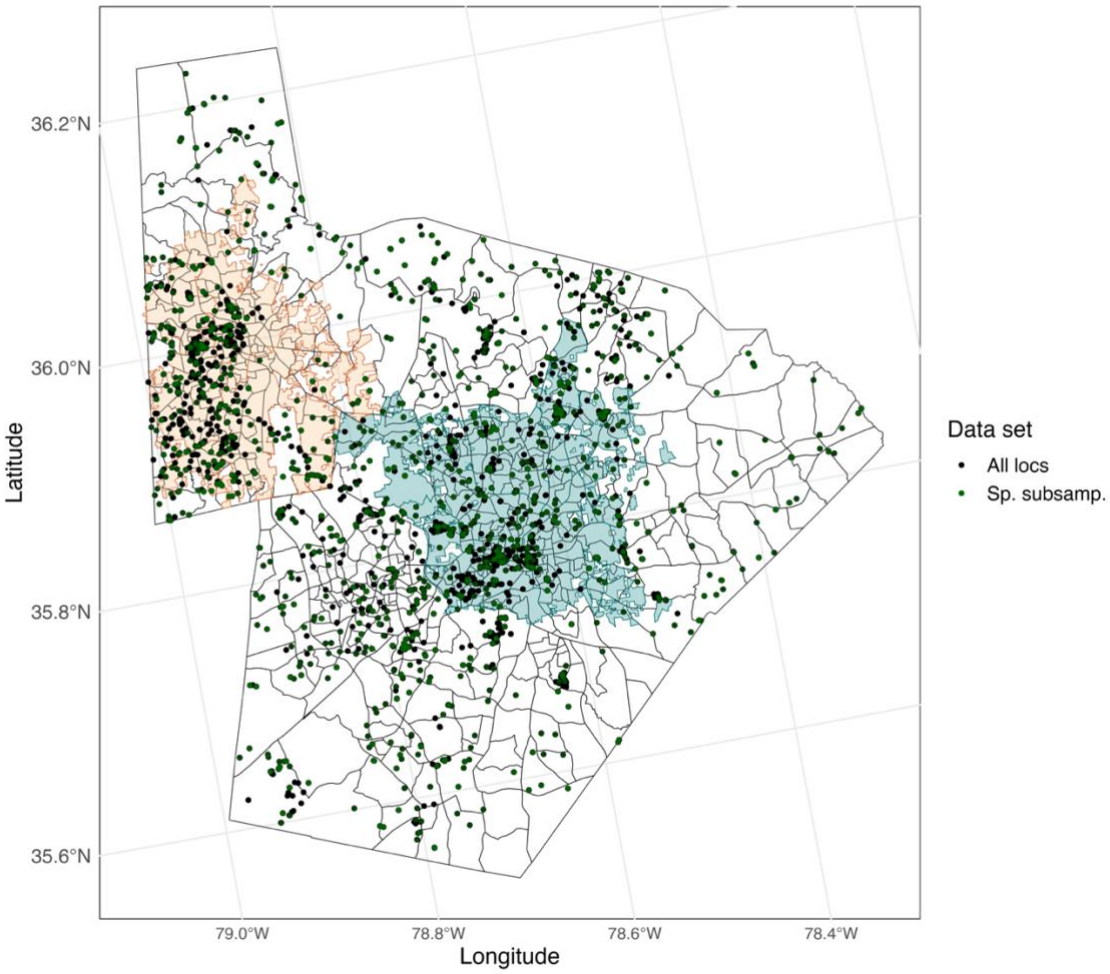


Figure 3.1. Map of study area encompassing Wake and Durham counties, North Carolina. Black outlined polygons are U.S. Census block groups. Shaded areas are the city limits of Durham (orange) and Raleigh (blue). All complete eBird checklists (n=10,253) and spatially subsampled checklists (n=3,201) collected in Wake and Durham counties, North Carolina, USA in April-June 2015-2019 are shown in black and green, respectively.

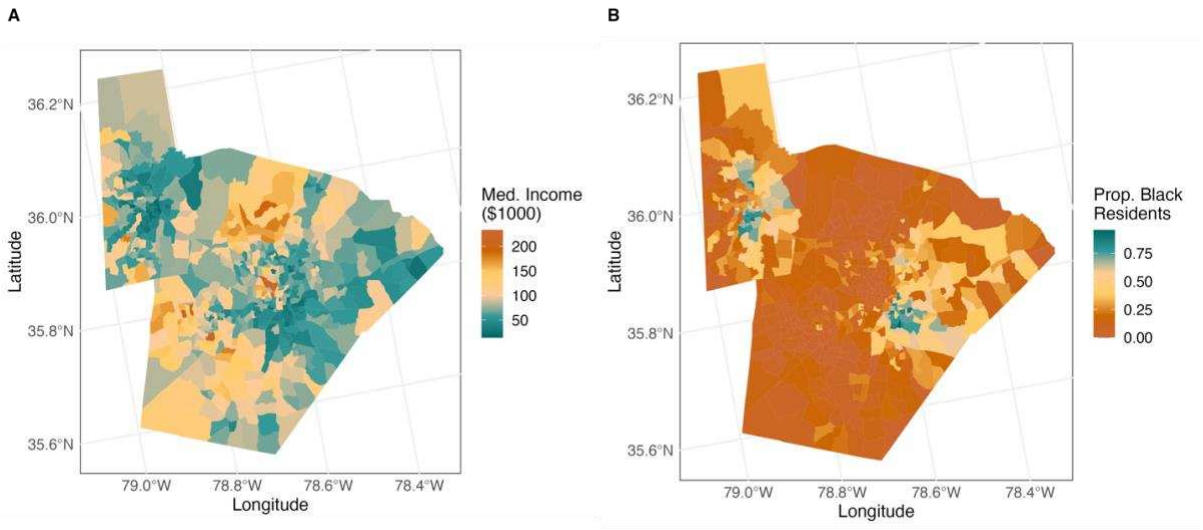


Figure 3.2. U.S. Census American Community Survey 5-year estimates of (A) Median household income (thousands of dollars) and (B) proportion of Black residents per U.S. Census block group in Wake and Durham counties, North Carolina, USA from 2015-2019.

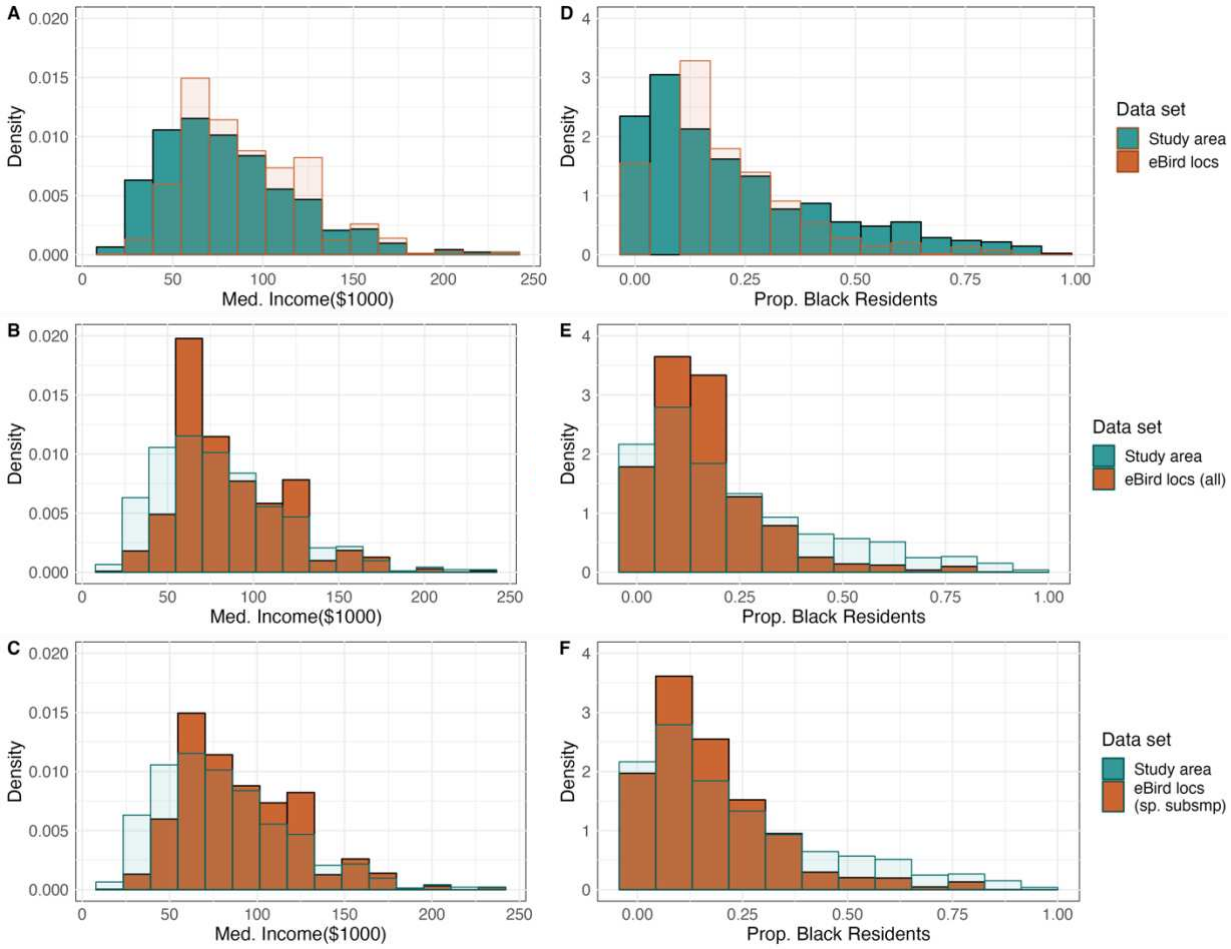


Figure 3.3. Histograms of (A) median household income and (D) the proportion of Black residents per US Census block group across the study region encompassing Wake and Durham counties, North Carolina, versus (B, C) the median household incomes and (E, F) proportions of Black residents associated with the locations of eBird locations from April-June 2015-2019. Top panel (A, D) shows demographics for full study area; middle panel (B, E) shows demographics associated with all eBird locations (n=10, 253); and bottom panel (C, F) shows demographics associated with eBird locations following spatial subsampling on a 5km hexagonal grid of the study area (n=3,201). Relative to the study area overall, median household income is skewed toward higher incomes and proportion of Black residents is skewed towards lower proportions for both spatially subsampled and un-subsampled eBird data.

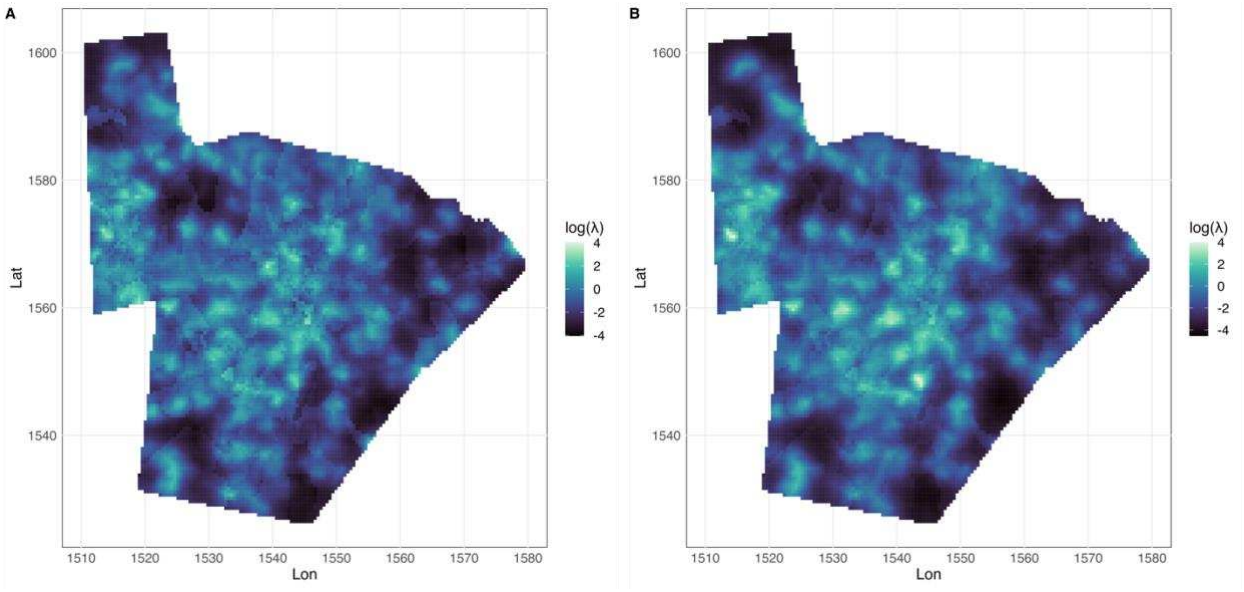


Figure 3.4. Sampling log-intensity predicted from log-Gaussian Cox Process (LGCP) model for eBird sampling based on complete checklists collected between April-June 2015-2019 in Wake and Durham counties, N.C., using (A) n=3,201 observations spatially sub-sampled using a 5km hexagonal grid, with one location per grid cell per week, and (B) n=3,201 randomly sampled (nonspatial) observations. For both datasets, sampling intensity is positively associated with median household income. Thus, spatially subsampling eBird observations prior to analysis did not mitigate social sampling bias.

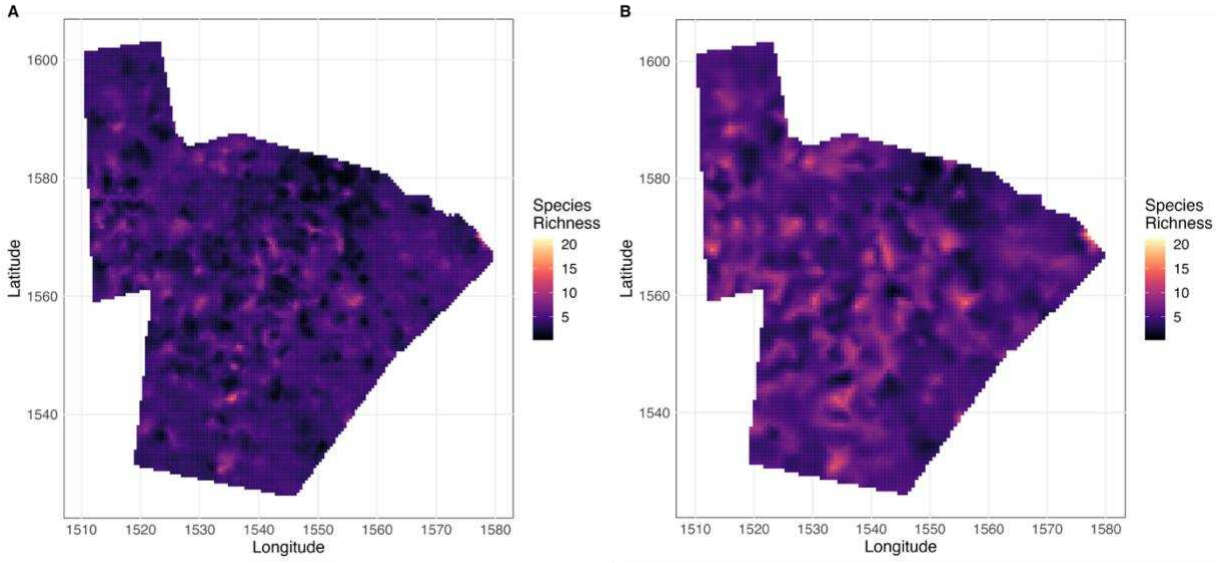


Figure 3.5. Bird species richness predicted from two models investigating the relationship between species richness and median household income in the Research Triangle metropolitan area, North Carolina. Both models were fit to n=3,201 spatially subsampled eBird checklists collected between April-June 2015-2019. (A) Estimated species richness from a spatial regression with median household income, checklist duration, and checklist distance traveled as fixed effects and an SPDE spatial random effect. Estimates reflect a small negative association between median household income and species richness. (B) Estimated species richness from a joint sampling intensity-species richness model with a fixed effect of income in both sampling and richness, checklist duration and checklist distance traveled in richness, and a shared spatial random effect capturing the relationship between sampling intensity and richness. Estimates do not reflect a relationship between income and richness, beyond the influence of sampling bias on richness estimates. Accounting for sampling via the joint model results in a smoother estimate of species richness across the study area that is larger, on average, than the estimates from the model that does not account for sampling.

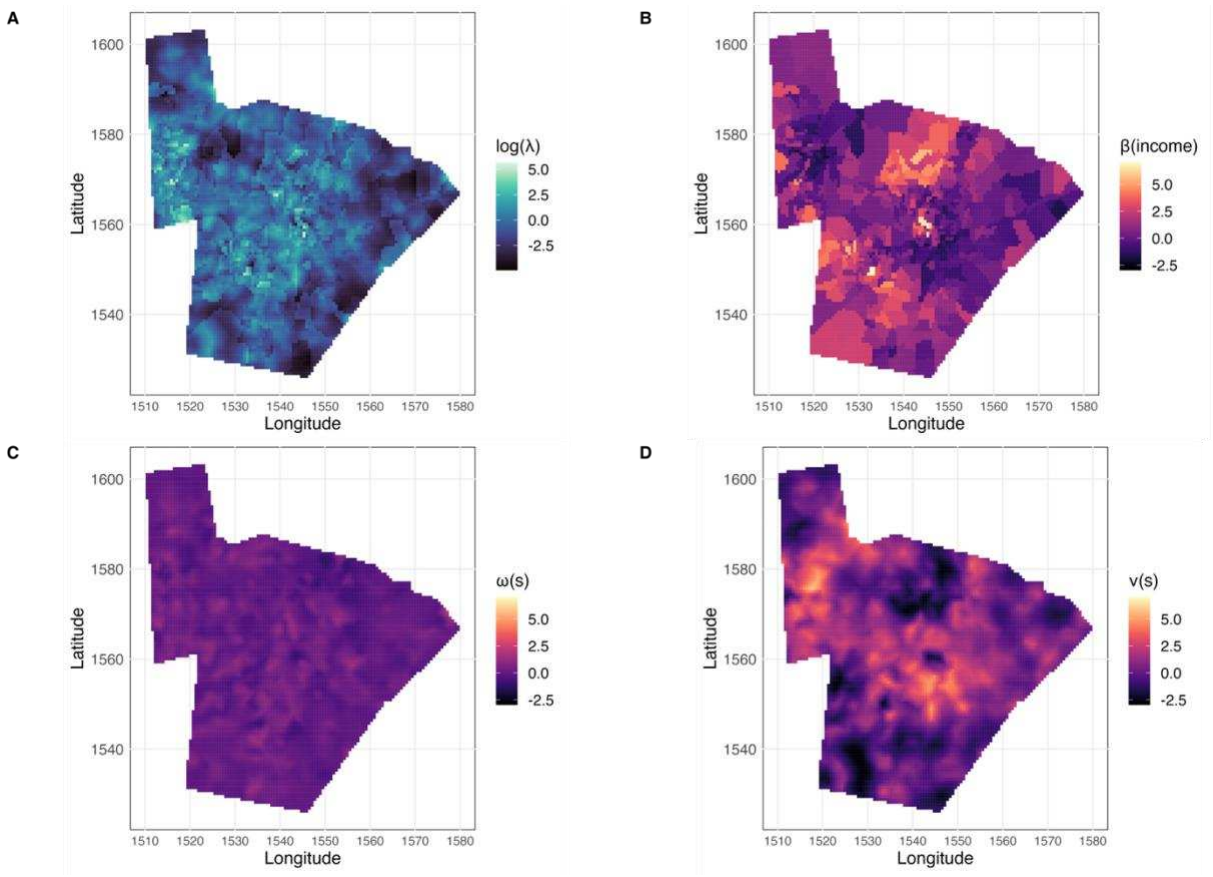


Figure 3.6. Estimates from joint sampling-bird species richness with fixed effect of income and shared spatial random effect, fit to n=3,201spatially subsampled checklists collected in the Research Triangle, N.C. between April-June 2015-2019. (A) Estimated sampling log-intensity for eBird checklists. (B) Estimated relationship between median household income and sampling intensity. Sampling intensity was positively associated with income. (C) Estimated spatial dependence between sampling intensity and species richness that is not explained by income. (D) Estimated residual spatial structure in sampling intensity that is not explained by median household income or shared with species richness.

REFERENCES

Ahern, J. (2013). Urban landscape sustainability and resilience: the promise and challenges of integrating ecology with urban planning and design. *Landscape Ecology*, 28, 1203-1212.

Albert, C. H., Yoccoz, N. G., Edwards Jr, T. C., Graham, C. H., Zimmermann, N. E., and Thuiller, W. (2010). Sampling in ecology and evolution–bridging the gap between theory and practice. *Ecography*, 33(6), 1028-1037.

Andersson, E., Haase, D., Anderson, P., Cortinovis, C., Goodness, J., Kendal, D., Lausch, A., McPHearson, T., Sikorska, D., and Wellmann, T. (2021). What are the traits of a social-ecological system: Towards a framework in support of urban sustainability. *Urban Sustainability*, 1(1), 14.

Aronson, M. F. J. et al. (2014) A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers. *Proceedings of the Royal. Society Series B: Biological Sciences* 281, 20133330.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6), 760-766.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). Hierarchical Modeling and Analysis for Spatial Data, second edition. Chapman and Hall/CRC.

Bocinsky, R. K., Beaudette, D., Chamberlain, S., and Bocinsky, M. R. K. (2015). 'FedData': Functions to Automate Downloading Geospatial Data Available from Several Federated Data Sources. R package version 3.0.4, <https://CRAN.R-project.org/package=FedData>.

Callaghan, C. T., Bino, G., Major, R. E., Martin, J. M., Lyons, M. B., and Kingsford, R. T. (2019a). Heterogeneous urban green areas are bird diversity hotspots: insights using continental-scale citizen science data. *Landscape Ecology*, 34, 1231-1246.

Callaghan, C. T., Major, R. E., Lyons, M. B., Martin, J. M., Wilshire, J. H., Kingsford, R. T., and Cornwell, W. K. (2019b). Using citizen science data to define and track restoration targets in urban areas. *Journal of Applied Ecology*, *56*(8), 1998-2006.

Callaghan, C. T., Sayol, F., Benedetti, Y., Morelli, F., and Sol, D. (2021). Validation of a globally-applicable method to measure urban tolerance of birds using citizen science data. *Ecological Indicators*, *120*, 106905.

Chamberlain, D. E., Henry, D. A., Reynolds, C., Caprio, E., and Amar, A. (2019). The relationship between wealth and biodiversity: A test of the Luxury Effect on bird species richness in the developing world. *Global Change Biology*, *25*(9), 3045-3055.

Chamberlain, D., Reynolds, C., Amar, A., Henry, D., Caprio, E., and Batáry, P. (2020). Wealth, water and wildlife: Landscape aridity intensifies the urban luxury effect. *Global Ecology and Biogeography*, 29(9), 1595-1605.

Chen, G., Li, X., Liu, X., Chen, Y., Liang, X., Leng, J., Xu, X., Liao, W., Qiu, Y.A., Wu, Q., and Huang, K. (2020). Global projections of future urban land expansion under shared socioeconomic pathways. *Nature Communications*, 11(1), 537.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons.

Conn, P. B., Thorson, J. T., and Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11), 1535-1546.

Crimmins, T. M., Posthumus, E., Schaffer, S., and Prudic, K. L. (2021). COVID-19 impacts on participation in large scale biodiversity-themed community science projects in the United States. *Biological Conservation*, 256, 109017.

Dewitz, J., and U.S. Geological Survey, (2021). National Land Cover Database (NLCD) 2019 Products (ver. 2.0, June 2021): U.S. Geological Survey data release, https://doi.org/10.5066/P9KZCM54. *Accessed 16 December 2022.*

Diggle, P. J., Menezes, R., and Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(2), 191-232.

Duncan, O. D., and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210-217.

eBird. (2022). eBird: An online database of bird distribution and abundance [web application]. eBird, Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org. April 2022 data release, accessed 09 June 2022

Edwards Jr, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., and Moisen, G. G. (2006). Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, 199(2), 132-141.

Ellis-Soto, D., Chapman, M., and Locke, D. H. (2023). Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States. *Nature Human Behaviour*, 1-9.

Fink, D., T. Auer, A. Johnston, V. Ruiz-Gutierrez, W. M. Hochachka, and S. Kelling. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* 30: e02056.

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W. M., Ligocki, S., ... and Rodewald, A. D. (2023). A double machine learning trend model for citizen science data. *Methods in Ecology and Evolution*, *14*(9), 2435-2448.

Fong, E., and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2), 489-496.

Frank, B., Delano, D., and Caniglia, B. S. (2017). Urban systems: A socio-ecological system perspective. *Sociology International Journal,* 1(1), 1-8.

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445-452.

Grade, A. M., Chan, N. W., Gajbhiye, P., Perkins, D. J., and Warren, P. S. (2022). Evaluating the use of semi-structured crowdsourced data to quantify inequitable access to urban biodiversity: A case study with eBird. *PLoS One*, *17*(11), e0277223.

Grove, J. M., Locke, D. H. and O'Neil-Dunne, J. P. M. (2014). An Ecology of Prestige in New York City: Examining the Relationships Among Population Density, Socioeconomic Status, Group Identity, and Residential Canopy Cover. *Environmental Management* 54, 402–419.

Grove, M., Ogden, L., Pickett, S., Boone, C., Buckley, G., Locke, D. H., Lord, C., and Hall, B. (2020). The legacy effect: Understanding how segregation and environmental injustice unfold over time in Baltimore. In *Social Justice and the City* (pp. 224-237). Routledge.

Habitat, U. N. (2016) World cities report 2016: Urbanization and development–emerging futures. Publ. UN-Habitat.

Haklay, M. E. (2016). Why is Participation Inequality Important? Ubiquity Press.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*(4), 243-254.

Hassell, J. M., Bettridge, J. M., Ward, M. J., Ogendo, A., Imboma, T., Muloi, D., ... and Fèvre, E. M. (2021). Socio-ecological drivers of vertebrate biodiversity and human-animal interfaces across an urban landscape. *Global Change Biology*, *27*(4), 781-792.

Heymans, A., Breadsell, J., Morrison, G. M., Byrne, J. J., and Eon, C. (2019). Ecological urban planning and design: A systematic literature review. *Sustainability*, 11(13), 3723.

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., ... and Hijmans, M. R. J. (2023) raster: Geographic Data Analysis and Modeling. R package version 3.6-23, <https://CRAN.R-project.org/package=raster>.

Hijmans, R. J., Bivand, R., Forner, K., Ooms, J., Pebesma, E., and Sumner, M. D. (2023b). 'terra': Spatial Data Analysis. R package version 1.7-46. https://CRAN.R-project.org/package=terra

Hochachka, W. M., Alonso, H., Gutiérrez-Expósito, C., Miller, E., and Johnston, A. (2021). Regional variation in the impacts of the COVID-19 pandemic on the quantity and quality of data collected by the project eBird. *Biological Conservation*, *254*, 108974.

Hope, D., Gries, C., Zhu, W., Fagan, W. F., Redman, C. L., Grimm, N. B., Nelson, C.M., and Kinzig, A. (2003). Socioeconomics drive urban plant diversity. *Proceedings of the National Academy of Sciences*, 100(15), 8788-8792.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., and Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, 44(9), 1259-1269.

Irvine, K. M., Rodhouse, T. J., Wright, W. J., and Olsen, A. R. (2018). Occupancy modeling species–environment relationships with non-ignorable survey designs. *Ecological Applications*, *28*(6), 1616-1625.

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., Kelling, S.T., and Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265-1277.

Kolstoe, S., and Cameron, T. A. (2017). The non-market value of birding sites and the marginal value of additional species: biodiversity in a random utility model of site choice by eBird members. *Ecological Economics*, *137*, 1-12.

Knapp, S., Aronson, M. F., Carpenter, E., Herrera-Montes, A., Jung, K., Kotze, D. J., LaSorte, F.A., Lepczyk, C.A., MacGregor-Fors, I., MacIvor, J.S., Moretti, M., Nilon, C.H., Piana, M.R., Rega-Brodsky, C.C., Salisbury, A., Threlfall, C.G., Trisos, C., Williams, N.S.G., and Hahs, A. K. (2021). A research agenda for urban biodiversity in the global extinction crisis. *BioScience*, 71(3), 268-279.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3), 619-637.

Kuldna, P., Poltimäe, H., and Tuhkanen, H. (2020). Perceived importance of and satisfaction with nature observation activities in urban green areas. *Journal of Outdoor Recreation and Tourism*, 29, 100227.

Kuras, E. R., Warren, P. S., Zinda, J. A., Aronson, M. F., Cilliers, S., Goddard, M. A., Nilon, C.H., and Winkler, R. (2020). Urban socioeconomic inequality and biodiversity often

converge, but not always: A global meta-analysis. *Landscape and Urban Planning*, 198, 103799.

La Sorte, F. A., Aronson, M. F., Lepczyk, C. A., and Horton, K. G. (2020). Area is the primary correlate of annual and seasonal patterns of avian species richness in urban green spaces. *Landscape and Urban Planning*, 203, 103892.

La Sorte, F. A., Clark, J. A., Lepczyk, C. A., and Aronson, M. F. (2023). Collections of small urban parks consistently support higher species richness but not higher phylogenetic or functional diversity. *Proceedings of the Royal Society B*, 290(2006), 20231424.

Leong, M., Dunn, R. R., and Trautwein, M. D. (2018). Biodiversity and socioeconomics in the city: a review of the luxury effect. *Biology Letters*, *14*(5), 20180082.

Lerman, S. B., Narango, D. L., Avolio, M. L., Bratt, A. R., Engebretson, J. M., Groffman, P. M., Hall, S.J., Heffernan, J.B., Hobbie, S.E., Larson, K.L., Locke, D.H., Neill, C., Nelson, K.C., Padullés Cubino, J., and Trammell, T. L. (2021). Residential yard management and landscape cover affect urban bird community diversity across the continental USA. *Ecological Applications*, 31(8), e02455.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(4), 423-498.

Liu, Z., He, C. and Wu, J. (2016) The Relationship between Habitat Loss and Fragmentation during Urbanization: An Empirical Evaluation from 16 World Cities. *PLoS One* 11, e0154613.

Loss, S. R., Ruiz, M. O., and Brawn, J. D. (2009). Relationships between avian diversity, neighborhood age, income, and environmental characteristics of an urban landscape. *Biological Conservation*, *142*(11), 2578-2585.

Magle, S. B., Fidino, M., Sander, H. A., Rohnke, A. T., Larson, K. L., Gallo, T., Kay, C.A.M., Lehrer, E.W., Murray, M.H., Adalsteinsson, S.A., Ahlers, A.A., …, and Schell, C. J. (2021). Wealth and urbanization shape medium and large terrestrial mammal communities. *Global Change Biology*, 27(21), 5446-5459.

Mahmoudi, D., Hawn, C. L., Henry, E. H., Perkins, D. J., Cooper, C. B., and Wilson, S. M. (2022). Mapping for whom? Communities of color and the citizen science gap. *UMBC Faculty Collection*.

Mair, L., and Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PloS One*, *11*(1), e0147796.

McHale, M. R., Pickett, S. T., Barbosa, O., Bunn, D. N., Cadenasso, M. L., Childers, D. L., ... and Zhou, W. (2015). The new global urban realm: complex, connected, diffuse, and diverse social-ecological systems. *Sustainability*, 7(5), 5211-5240.

Millard, A. (2010). Cultural aspects of urban biodiversity. *Urban Biodiversity and Design.* Wiley-Blackwell, Oxford, 56-80.

Minor, E., and Urban, D. (2010). Forest bird communities across a gradient of urban development. *Urban Ecosystems*, *13*, 51-71.

Pati, D., Reich, B. J., and Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1), 35-48.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

Pebesma, E., and Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Peeters, E. T. H. M. et al. (2022) Monitoring biological water quality by volunteers complements professional assessments. *PLOS ONE* 17, e0263899.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., and Conesa, D. (2019). Accounting for preferential sampling in species distribution models. *Ecology and Evolution*, 9(1), 653-663.

Perkins, D. J. (2020). Blind Spots in Citizen Science Data: Implications of Volunteer Biases in eBird Data. North Carolina State University.

Piano, E., Isaia, M., Falasco, E., La Morgia, V., Soldato, G., and Bona, F. (2017). Local versus landscape spatial influence on biodiversity: A case study across five European industrialized areas. *Environmental Monitoring and Assessment*, *189*, 1-12.

Pickett, S. T., Cadenasso, M. L., Childers, D. L., McDonnell, M. J., and Zhou, W. (2016). Evolution and future of urban ecological science: ecology in, of, and for the city. *Ecosystem Health and Sustainability*, 2(7), e01229.

Planillo, A., Fiechter, L., Sturm, U., Voigt-Heucke, S., and Kramer-Schadt, S. (2021). Citizen science data for urban planning: comparing different sampling schemes for modelling urban bird distribution. *Landscape and Urban Planning*, *211*, 104098.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319-392.

Salbitano, F., Borelli, S., Conigliaro, M., Yahya, N. A., Sanesi, G., Chen, Y., and Corzo, G. T. (2017). Urban forest benefits in developing and industrializing countries. In *Routledge Handbook of Urban Forestry* (pp. 136-151). Routledge.

Schell, C. J., Dyson, K., Fuentes, T. L., Des Roches, S., Harris, N. C., Miller, D. S., Woelfle-Erskine, C.A., and Lambert, M. R. (2020). The ecological and evolutionary consequences of systemic racism in urban environments. *Science*, 369(6510), eaay4497.

Seto, K. C., Fragkias, M., Güneralp, B. and Reilly, M. K. A. (2011) Meta-Analysis of Global Urban Land Expansion. *PLoS One* 6, e23777.

Sicacha-Parada, J., Steinsland, I., Cretois, B., and Borgelt, J. (2021). Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics*, *42*, 100446.

Sieber, Renée E, and Muki Haklay. (2015). "The Epistemology(s) of Volunteered Geographic Information: A Critique: The Epistemology(s) of VGI: A Critique." *Geo: Geography and Environment* 2 (2): 12236. https://doi.org/10.1002/geo2.10.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science,* 32(1).

Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D., and Rue, H. (2019). Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. *Journal of the Royal Statistical Society Series C: Applied Statistics,* 68(3), 543-564.

Soifer, L. G., Donovan, S. K., Brentjens, E. T., and Bratt, A. R. (2021). Piecing together cities to support bird diversity: Development and forest edge density affect bird richness in urban environments. *Landscape and Urban Planning*, 213, 104122.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4), 583-639.

Steen, V. A., Tingley, M. W., Paton, P. W., and Elphick, C. S. (2021). Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution*, 12(2), 216-226.

Strimas-Mackey, M., Miller, E., and Hochachka, W. (2018) auk: eBird Data Extraction and Processing with AWK. *R package version 0.3.0.* https://cornelllabofornithology.github.io/auk/

Strohbach, M. W., Haase, D., and Kabisch, N. (2009). Birds and the city: urban biodiversity, land use, and socioeconomics. *Ecology and Society*, 14(2).

Stuber, E. F., Robinson, O. J., Bjerre, E. R., Otto, M. C., Millsap, B. A., Zimmerman, G. S., Brasher, M.G., Ringelman, K.M., Fournier, A.M., Yetter, A., Isola, J.E., ... and Ruiz-Gutierrez, V. (2022). The potential of semi-structured citizen science data as a supplement for conservation decision-making: validating the performance of eBird against targeted avian monitoring efforts. *Biological Conservation*, 270, 109556.

Sullivan, B.L., C.L. Wood, M.J. Iliff, R.E. Bonney, D. Fink, and S. Kelling. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142: 2282-2292.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... and Kelling, S. (2014) The eBird enterprise: An integrated approach to development and application of citizen science. *Biological conservation*, *169*, 31-40.

Sullivan, B. L., Phillips, T., Dayer, A. A., Wood, C. L., Farnsworth, A., Iliff, M. J., ... and Kelling, S. (2017). Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, 208, 5-14.

Sultana, M., Storch, I., Naser, M. N., and Uddin, M. (2022). Land cover and socioeconomic factors explain avian diversity in a tropical megacity. *Ecology and Society*, 27(1).

Tang, B., Clark, J. S., and Gelfand, A. E. (2021). Modeling spatially biased citizen science effort through the eBird database. *Environmental and Ecological Statistics*, 28(3), 609-630.

Tivadar, M. (2019). "OasisR: An R Package to Bring Some Order to the World of Segregation Measurement." *Journal of Statistical Software*, 89 (7), 1-39. https://doi.org/10.18637/jss.v089.i07

Urdangarin, A., Goicoa, T., and Ugarte, M. D. (2023). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 36(2), 333-360.

U.S. Census Bureau. (2020). "2015-2019 American Community Survey 5-Year Estimates." Washington, D.C.: US Census Bureau.

Velkoff, V., and Hartley, C. (2022). Moving forward with the US Census Bureau's annual population estimates post-2020. *Harvard Data Science Review,* 4(4).

Walker, K., and Herman, M. (2023). Tidycensus: Load us census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames. R package, version *1*.4.4.

Watanabe, S., and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).

Weiser, E. L., Diffendorfer, J. E., Lopez-Hoffman, L., Semmens, D., and Thogmartin, W. E. (2020). Challenges for leveraging citizen science to support statistically robust monitoring programs. *Biological Conservation*, *242*, 108411.

Whittemore, A. H. (2018). The role of racial bias in exclusionary zoning: The case of Durham, North Carolina, 1945–2014. *Environment and Planning A: Economy and Space*, *50*(4), 826-847.

Williams, B. K., and Brown, E. D. (2019). Sampling and analysis frameworks for inference in ecology. *Methods in Ecology and Evolution*, *10*(11), 1832-1842.

CONCLUSION

Quantitative ecology can help us do more with data by understanding the complex processes influencing and underlying those data. Unifying this dissertation is the relationship between quantitative techniques and the inextricable interactions between the observer, the observation process, and the observed. Although we often strive for generalizability in our methods, each chapter of this dissertation explores how tailoring statistical or computing methods to specific observational contexts is important for advancing ecological knowledge. In the first chapter, the computational technique that we present parallels the process of data collection – first considering each individual separately, then combining information across individuals in a second stage of model fitting – to speed computation while promoting interpretation at both individual and population levels. This technique can facilitate inference not only related to variation in individual behavior, but can also be used to account for distinct sources of observational noise that vary across individuals.

While Chapter 1 highlights an example of a quantitative technique to extend what we can do with large, multi-faceted datasets, Chapters 2 and 3 explore sampling bias in crowdsourced ecological data and highlight many future directions and open questions, including what we *should* do with ecological data in some cases. eBird and other crowdsourced data are a valuable tool for ecological research and a major facet of big data ecology. However, these chapters highlight the critical importance of understanding and remedying the sampling bias that influences crowdsourced data, bringing together quantitative techniques to seek solutions to the systemic inequity woven throughout the design and fabric of these programs.

Though I set out to analytically mitigate social sampling bias in crowdsourced data to a greater extent, the limitations I encountered underscore the importance of addressing these issues in the design and implementation of crowdsourced projects. Crowdsourced and contributory science projects must prioritize more equitable participation in their platforms, both to move toward more ethical, equitable practice and because current participation inequity negatively impacts data quality and project goals. Although analytical mitigation of sampling bias is helpful, it is a single component of a much larger, multifaceted effort to reduce the inequity and bias in ecological data. Chapters 2 and 3 lay the foundation for additional quantitative work related to social sampling bias in eBird data, including data integration with spatially balanced data, the exploration of additional modeling frameworks, and the investigation of relationships between urban ecological hypotheses and sampling bias across more cities. However, none of this additional research can mitigate missing data in low-income communities and communities of color, and addressing these data gaps must be a priority for crowdsourced biodiversity science programs.

The societal context we all work in informs how we interact with and observe ecological processes, despite narratives from the western scientific academy that attempt to separate humans from ecological processes. We are trained to believe that through quantitation and mathematics, we can remove ourselves from the systems we study to act as objective observers. However, we are all participants actively shaping the ecological processes we observe, and the actions, approaches, and assumptions used in our research reflect societal systems and biases. Data are never objective, and it is dangerous and false to assume that quantitative techniques can take data out of the contexts in which they were collected. Instead, quantitative frameworks that embrace, reflect, and seek to improve the ways in which social and observational contexts inform

what is observed can elevate analytical techniques to tools towards more just, inclusive, and transparent ecological research and conservation. We must improve equity in our research methods, participation, and scope because we cannot gain the knowledge we need to develop just solutions to climate change and biodiversity loss from biased datasets and inequitable systems. In the current context of technology and automation, including increases in computing, artificial intelligence, and remote sensing, it is more important than ever that we maintain a connection to the people, places, and organisms that data represent.