Christian Stratton
christianstratton@montana.edu

Research statement

John Tukey once said, "The best thing about being a statistician is that you get to play in everyone's backyard." This quote is an adage that I have embraced through my research by developing statistical methodology to answer scientific questions posed by many different collaborators in the environmental and ecological sciences. Recently, as a post-doctoral researcher working with Dr. Kathryn Irvine at the U.S. Geological Survey, my research interests have primarily concerned the development of statistical methodology to support the North American Bat Monitoring Program (NABat). The NABat program is a multi-agency, multi-discipline, collaborative monitoring program that aims to assess the status and trend of bat populations in North America, with particular emphasis on evaluating the response of bat populations to stressors such as disease and wind energy. As one of two statisticians on the team of interdisciplinary scientists, my role entailed developing statistical methodology for species distribution modeling (SDM), developing methods for assessing the impact of disease on bat populations, and communicating and collaborating with domain experts in the environmental and ecological sciences.

Besides the modeling of species distributions, the work I have done with the NABat program during my post-doctoral appointment has led to additional lines of research under the umbrella of development of statistical methodology for the environmental and ecological sciences. For example, the data sets provided by collaborators through the NABat program are often spatially explicit and span large spatial domains such as the continental United States; analysis of these data sets requires scalable modeling frameworks for spatially explicit data. As a result, I have developed broad research interests in spatial statistics and Bayesian computation. Additionally, the data contributors to the NABat program often provide multiple different types of data, leading to an interest in data integration techniques for ecological data types.

Other lines of research in the environmental and ecological sciences predate my post-doctoral appointment with the NABat program. Prior to that appointment, I collaborated with environmental and ecological scientists working with plant community data, motivating interest in Bayesian nonparametrics to facilitate clustering of subgroups within the larger communities. In the sections below, I describe my contributions to each of the research areas described above and provide my plan for future research at [**SCHOOL**].

- **Species distribution modeling:** When monitoring birds, bats, insects, and frogs, we often rely upon acoustic recording devices to assess species presence; the audio recordings are then processed and assigned a species label via proprietary software packages. These software packages are prone to classification errors, which, if not acknowledged, result in biased estimates of species presence and abundance. My previous work investigated the impact of different validation strategies to account for classification errors in species distribution modeling; this work is published in *Methods in Ecology and Evolution*.

  My future research interests in this vein concern developing statistical methodology that better accounts for the idiosyncrasies of the acoustic processing pipeline. Current analytic frameworks remove audio recordings that are not assigned a single species label before analysis, resulting in a large volume of data being discarded before analysis. Future research will investigate statistical methodologies that can accommodate a larger portion of audio recordings, including those downgraded to multi-species labels.

- **Spatial statistics:** I am broadly interested in developing tools for spatially motivated data. Recently, my research in this area has concerned the development of a modeling framework to assess the impact of a wildlife disease called White-nose Syndrome (WNS) on North American bat populations. Before my involvement, this assessment was typically made by first modeling the disease spread to obtain some summary of disease occurrence across the landscape. Then, a measure of species distribution is regressed on that disease summary, and an estimate of the impact of the disease is obtained. However, this approach does not account for the uncertainty present in estimating the disease occurrence, resulting in overly precise estimates of the impact of the disease on the species distribution. A potential solution to this problem is to jointly model the two data sources in a single spatially misaligned regression model, allowing for appropriate uncertainty propagation throughout the model. This work is under review in the *Journal of*

Agricultural, Biological, and Environmental Statistics and was presented at the Joint Statistical Meetings in 2023.

Future work on this modeling effort is multi-faceted. First, I plan to explore optimal sampling designs for the joint modeling framework. Previous investigations into optimal designs for spatial modeling typically focus on guidance for selecting sample locations to measure one variable. For the joint modeling framework, it would be valuable to consider how to distribute effort for both the acoustic monitoring and disease surveillance data sources. Additional avenues for future work include adding complexity to the disease observation process to accomodate false-positive detections, adding temporal dynamics to the disease process, and exploring scalable options for the spatial component of the model.

- **Bayesian computation:** As increasingly large and complex data becomes more accessible, the need for scalable analytic methods increases. My previous work in this area focused on constructing a computationally efficient MCMC sampling procedure for multi-scale occupancy models that leveraged data augmentation strategies to afford Gibbs updates for all parameters in the model. This work is published in *Methods in Ecology and Evolution*, and provided to the public via the 'msocc' 'R' package.

  Future work in this area concerns developing scalable MCMC sampling procedures for large spatial data sets. For example, by incorporating data augmentation strategies, the regression coefficients in the spatially misaligned regression model described above could be estimated via a Gibbs sampler. Additionally, the model is not well suited for large spatial domains due to the covariance structure of the Gaussian process. I plan to explore potential solutions to this problem, possibly considering locally approximated or nearest neighbor Gaussian processes.

- **Data integration:** Modern wildlife monitoring programs often rely on multiple data sources to infer species distribution and abundance. For example, when providing analytic support for listing three bat species under the Endangered Species Act, we were required to submit analyses of multiple data sources, including stationary acoustic data, mobile acoustic data, capture data, and roost count data. Recently, there has been interest in integrating multiple data sources under a single unified framework. Previous attempts at data integration in ecology typically rely on defining a multi-state observation process that accommodates the various data sources. My current research focuses on developing a single hierarchical model, motivated by a spatial point process, that can accommodate multiple ecological data sources. This model was recently implemented for a regulatory clearance process for tri-colored bats. The manuscript detailing this framework is currently in preparation.

  Following this effort, I plan to consider incorporation of other data sources, including the location of winter-time roosts. Currently, a summary of this information is typically included as a covariate on the occupancy portion of the model. This approach fails to acknowledge the uncertainty in locating winter-time roosts and assumes that all roosts are completely observed. By integrating the winter-time roost data in the model, potentially as a marked point process, the uncertainty in the discovery of winter-time roosts would be acknowledged.

- **Bayesian nonparametrics:** Scientists in the environmental and ecological sciences are often interested in clustering sample locations based on the distribution of species present. This problem usually manifests as a high-dimensional clustering problem, and the field of ordination has been developed to address this problem. Traditional ordination techniques rely upon a multi-step, distance-based process: first, the community data are converted to a dissimilarity matrix, where each entry describes the dissimilarity between sample locations based on some metric. Then, the distance matrix is projected into a lower dimension, often two, where the results are finally clustered based on their proximity in the lower dimension space. This multi-step process requires several subjective decisions (choice of dissimilarity index, projection technique, number of dimensions, number of clusters, etc.) that cannot be formally assessed, as the distance-based framework does not present a likelihood.

  My current research focuses on developing a hierarchical, model-based ordination framework that uses a Dirichlet process mixture to perform the clustering. This framework allows for probabilistic inferences about the number of clusters in the ecological community and provides tools for model comparison. This work is currently in review in *Methods in Ecology and Evolution*

and was presented at the Joint Statistical Meetings. In future work, I plan to explore the sensitivity of the modeling framework to the Dirichlet process priors and other potential clustering options, such as sparse finite mixture models.

- **Statistics education:** Given the availability of computational power and technology resources in the modern day, statistics education need not look the same as it did in the past. My current research interests focus on leveraging technology to improve statistics education and prepare students for the modern world. For example, during graduate school, I developed an 'R' Shiny web application to visualize the impact of multiple quantities (sample size, significance level, null values, etc.) on statistical power for various distributions. This web application has since been implemented in multiple statistical theory courses at Montana State University and Michigan State University at both undergraduate and graduate levels. A description of this work is published in *Technology Innovations in Statistics Education* and was presented at the Electronic Conference on Teaching Statistics in 2018, and an updated version of the application was presented at the United States Conference on Teaching Statistics in 2021.

  In the future, I will continue to explore how technology may be leveraged to improve instruction of statistics. For example, I will teach my courses using Quarto and integrate multiple coding languages, including 'R', Julia, and Python. Through this process, students gain valuable experience in technological literacy, a skill that translates to multiple career paths.

To John Tukey's point, my time in the backyard of ecologists has given me the opportunity to explore a number of different research topics, all unified under the development of statistical methodology for the environmental and ecological sciences. Beyond the environmental and ecological sciences, I have interest and experience in statistics education and sports analytics. However, as I look to begin the next chapter in my life as a faculty member at [**SCHOOL**], I am excited by the opportunity to forge new collaborative relationships and continue to explore backyards.