JOINT OUTCOME MODELING IN ENVIRONMETRICS

CB DEAN [1] AND EM RENOUF [1,*]

A commonly used method in epidemiology links multiple outcomes such as survival and longitudinal data in a so-called *joint outcome model*. This analytical approach allows the simultaneous analysis of different types of correlated outcomes on an individual in a single model. Similarly, biologists and ecologists are often interested in the relationship between multiple types of outcomes such as, for example, species survival and breeding success, incorporating many types of correlations beyond individual-level associations. The use of joint models in ecology and environmetrics has pioneered new developments of more flexible, complex and creative modeling methods. In particular, researchers faced with different data types to be simultaneously modeled have eagerly embraced the joint modeling framework, exploring such new approaches. Beyond accounting for potential correlations in outcomes explicitly, advantages include the ability to quantify such correlations, as well as potential efficiencies in analysis. The approach also allows better understanding of related processes and how they influence each other. The aim of this special issue is to identify advantages of using the joint modeling framework in the environmental context.

1. History of Joint Modeling

The development of joint modeling techniques had its origins in the study of the informative dropout process in the analysis of repeated outcomes in biomedical studies. Reasons for termination of data collection in a longitudinal process may be informative due to death or study drop-out. For example, subjects with more serious illness, or those who feel the treatment may not be working, may drop out early, leading to the collection of fewer longitudinal measurements. To accommodate this, the time of dropout occurrence is modeled (as time-to-event data) simultaneously with the repeated longitudinal outcome. By including the time-to-dropout data in a joint distributional

[1]Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1
[*]Corresponding author
E-mail addresses: erenouf@uwaterloo.ca.

analysis with the longitudinal data, we incorporate information on informative dropout and so bias is reduced; this approach can also be used to reduce bias through interpolation of missing covariate values and by accounting for measurement error. Very often, a medical study will collect multiple types of data on a subject, such as repeated measurements of a longitudinal covariate like blood pressure, and time-to-event data such as survival. Both the longitudinal covariate and the survival outcome may be of interest, e.g. one may wish to track trends in a covariate over time while assessing its effect on survival. Early papers on joint modeling used this framework for analyses where the association between the longitudinal measurements and the survival time was of interest, e.g. Faucett and Thomas (1996); Wulfsohn and Tsiatis (1997). Further advancements grew from applications of the joint model in the area of HIV/AIDS, where survival was jointly modeled with a longitudinal laboratory value of interest, namely CD4 counts. A comprehensive summary of the early advancements is provided in Tsiatis and Davidian (2004). Meanwhile Furgal et al. (2019) have provided a comprehensive review of more recent advances in joint modeling for longitudinal and survival data, including available software.

1.1. **Joint Modeling in Environmetrics.** Biologists and ecologists have long recognized that different types of data are important in assessment of outcomes in many areas of environmetrics. The use of a joint model can be motivated by interest in the relationship between multiple types of outcomes for a single species (e.g., species survival and breeding success), and also by the desire to include important and relevant information from different species or from multiple environmental processes in a model. The joint modeling framework may offer an appropriate solution to these challenges. The application of joint models in ecology and environmetrics has pioneered new developments of the technique to allow more flexible constructions. With advances in computational power, joint models are able to handle larger and more complex datasets such as ones often used in environmetrics. Very commonly, a joint model may examine the same outcome across space and time, where the spatial dependence between these outcomes is modeled through random effects. What has not been considered broadly in environmetrics, as it has in health sciences and medical applications, is modeling the relationship among different types of related outcomes.

One of the earlier, more well known, applications of joint modeling in environmetrics occurred in population dynamics research, where a growing awareness of the importance of accounting for individual variation in the life history process led to adoption of the joint model by Wintrebert et al. (2005) to analyze the survival and breeding probabilities in a population of seabirds, the kittiwake. This paper highlights the difficulties of accounting for individual heterogeneity in assessments of population dynamics using traditional approaches. The authors carry out the joint model application using a frailty effect to link the survival and the breeding probability. Breeding probability is modeled logistically, and conditional on the common random effect, survival is modeled using a frailty model. This 2005 paper, while widely referenced, was not initially followed by an increase in joint modeling applications in ecological research, as was warranted by the novelty and utility of the approaches it provided. As pointed out by Sanderlin et al. (2019), there are many cases where the benefit gained from integrating more data into a joint model is outweighed by the cost of collecting it, particularly in a research context where 'additional' data collection may not necessarily occur in the course of answering the main objective. Miller et al. (2019) point to additional constraints such as difficulty in specifying a set of shared parameters, as well as computational feasibility. In some cases, preference remains for other types of flexible integrated distribution models, as summarized within Miller et al. (2019).

However, the advantages offered by the joint modeling framework has helped drive adoption and expansion of the methodology. Over the last decade, researchers in the field of species abundance have enthusiastically adopted a joint modeling approach. Clark et al. (2014) address the need to account for the fact that species are limited by competition with other species and that article offers a detailed summary of ways in which approaches to species abundance studies can benefit from the use of this technique. In a comprehensive 2015 paper, Warton et al. (2015) details the use of hierarchical approaches such as multivariate generalized linear mixed models and latent variable models in a joint modeling approach for species abundance. The authors show how models based on the generalized linear modeling framework can be unwieldy with large unstructured variance-covariance matrices. There was also a desire to integrate data from multiple sources, for example, from environmental interactions or the effect of one species on another in terms of abundance. A comparison of possible approaches for joint species distribution modeling is offered by Wilkinson et al. (2019). In movement ecology, joint modeling allows the use of data with different time scales, thus removing the loss of information that comes with coarsening of the data, e.g. Adam et al.

(2019). Research in species distribution modeling using a joint modeling approach has been greatly expanded by the work of Gelfand and Shirota (2019), Clark et al. (2017), including accommodations for zero-inflated data (Clark and Gelfand (2019)).

Researchers using spatial and spatial-temporal data have also enthusiastically embraced the joint modeling framework. As noted in Feng and Dean (2012), the spatial structures of a collection of sites may be correlated and this common risk surface can be leveraged in a joint model. Feng and Dean (2012) include a review of early approaches for the joint modeling of spatial data. Spatial analysis for species distribution comparing individual and joint models has been covered comprehensively by Gelfand (2020). Climate science, forestry applications and wildland fire analyses are areas where joint models are now beginning to be applied extensively. Examples can be found in disease mapping studies such as tree growth trajectories and pine weevil infection (Nathoo (2010)), or in analyses of the duration and spread of wildfires in Canada (Xi et al. (2020)). Feng and Dean (2014) present an example of a joint model with zero-inflated count data for spatial outcomes, where the zero-inflated count data are spatially correlated, in an analysis of Commandra blister rust infection of lodgepole pine trees in British Columbia, Canada. The outcomes are the count of lesions on infected trees and the count of host plants within a grid surrounding each tree. A zero-inflated distribution is used for both the lesion counts and the count of host plants, linking across outcomes with a spatial random effect in a shared frailty model. This paper also provides useful discussion on the power of tests for a common spatial structure. This research showed that an approach using joint modeling of two zero-inflated outcomes has several important features. First, it accounts for the dependence across multiple outcomes, which can lead to less biased inference if not incorporated. Second, the joint models can model the probability of belonging to the mass of true zeros in both outcomes. The models also link the random count components in both outcomes.

Our aim in this special issue is to highlight some interesting work and motivate future applications of the joint model in environmetrics. In this issue, the applications span a wide range of topics. In climate science, Konzen et al. (2021) provide a framework for inference on storm peak significant wave heights at multiple locations in the North Sea. Xi et al. (2021) present an analysis of extreme wildland fire behaviour, modeling the duration and size of fires as joint outcomes. North et al. (2020) employ a joint statistical framework for an analysis of the minimum and maximum temperature cycles for the years 1996 to 2018 at almost one hundred thousand locations. Zhang et al. (2021) utilize a joint modeling framework for analysis of very large spatial datasets with over a million locations, illustrated by an application to a study of vegetation structure and its spatio-temporal variation. Finally, Niku et al. (2021) employ a latent variable model in an application of a joint model for species abundance, and also provide a validation of type 1 errors for interaction tests. Environmetrics data offer a wealth of opportunities for the application of a joint modeling framework. The range of the articles included in this issue, considering both the research topics and the statistical approaches, demonstrate the broad usefulness of the joint modeling framework. The special issue also motivates and identifies significant opportunities to address important gaps in joint modeling approaches to environmental research.

References:

Timo Adam, Christopher A Griffiths, Vianey Leos-Barajas, Emily N Meese, Christopher G Lowe, Paul G Blackwell, David Righton, and Roland Langrock. Joint modelling of multi-scale animal movement data using hierarchical hidden markov models. *Methods in Ecology and Evolution*, 10 (9):1536–1550, 2019.

James S Clark and Alan E Gelfand. Accommodating so many zeros: univariate and multivariate data. In *Handbook of Environmental and Ecological Statistics*, pages 211–239. Chapman and Hall Boca Raton, 2019.

James S Clark, Alan E Gelfand, Christopher W Woodall, and Kai Zhu. More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24 (5):990–999, 2014.

James S Clark, Diana Nemergut, Bijan Seyednasrollah, Phillip J Turner, and Stacy Zhang. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56, 2017.

Cheryl L Faucett and Duncan C Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15(15): 1663–1685, 1996.

C. X. Feng and C. B. Dean. Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. *Environmetrics*, 23(6):493–508, 2012.

C. X. Feng and C. B. Dean. Spatial and temporal pattern analysis of disease data. In D. M. Chen, B. Moulin, and J. H. Wu, editors, *Spatial and Temporal Dynamics of Infectious Diseases*. Wiley, 2014.

Allison KC Furgal, Ananda Sen, and Jeremy MG Taylor. Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review*, 87 (2):393–418, 2019.

Alan E Gelfand. Statistical challenges in spatial analysis of plant ecology data. *Spatial Statistics*, page 100418, 2020.

Alan E Gelfand and Shinichiro Shirota. Clarifying species dependence under joint species distribution modeling. *arXiv preprint arXiv:1908.09410*, 2019.

E Konzen, C Neves, Jonathan, and P. Modeling nonstationary extremes of storm severity: Comparing parametric and semiparametric inference. *Environmetrics*, 32:e2667, 2021.

David AW Miller, Krishna Pacifici, Jamie S Sanderlin, and Brian J Reich. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37, 2019.

Farouk S Nathoo. Joint spatial modeling of recurrent infection and growth with processes under intermittent observation. *Biometrics*, 66(2):336–346, 2010.

J Niku, Hui, Fkc, S Taskinen, and D Warton. Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics*, 2021. doi: 10.1002/env.2683.

J North, E Schliep, and C Wikle. On the spatial and temporal shift in the archetypal seasonal temperature cycle as driven by annual and semi-annual harmonics. *Environmetrics*, 2020. doi: 10.1002/env.2665.

Jamie S Sanderlin, William M Block, Brenda E Strohmeyer, Victoria A Saab, and Joseph L Ganey. Precision gain versus effort with joint models using detection/non-detection and banding data. *Ecology and evolution*, 9(2):804–817, 2019.

Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

David I Warton, F Guillaume Blanchet, Robert B O'Hara, Otso Ovaskainen, Sara Taskinen, Steven C Walker, and Francis KC Hui. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.

David P Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A McCarthy. A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2):198–211, 2019.

Claire MA Wintrebert, AH Zwinderman, Emmanuelle Cam, Roger Pradel, and JC Van Houwelingen. Joint modelling of breeding and survival in the kittiwake using frailty models. *Ecological modelling*, 181(2-3):203–213, 2005.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

Dexen DZ Xi, CB Dean, and Stephen W Taylor. Modeling the duration and size of extended attack wildfires as dependent outcomes. *Environmetrics*, page e2619, 2020.

Dexen DZ Xi, CB Dean, and Stephen W Taylor. Modeling the duration and size of wildfires using joint mixture models. *Environmetrics*, 2021. doi: 10.1002/env.2685.

L Zhang, S Banerjee, and A Finley. High-dimensional multivariate geostatistics: A Bayesian matrix-normal approach. *Environmetrics*, 32:e2675, 2021. doi: 10.1002/env.2675.