Title: An individual-based Bayesian survival model accounting for variable exposure times, repeated surveys, and covariates that change across survey intervals

Authors: Ana Miller-ter Kuile, Kiona Ogle, Jamie Sanderlin

Baby:

Survival analyses have broad application across x, y, z subject areas and are important for understanding many processes, from human health to forest management. Individual-based survival models help us understand \_\_\_.

Zoomed in baby:

There are different analyses created for different types of survival analyses. Many traditional analyses assess survival across one survey interval, usually with one “treatment” applied to a group throughout that entire interval and either outcomes or time to outcomes are measured for different individuals/data points. Others, such as in wildlife biology and forest ecology, try to link survival to finer-scale covariate data that is likely to change throughout the survey period. To do this, most studies employ a logistic exposure model to survey data that takes place over one or multiple sub-intervals throughout the survey period for each individual being tracked (Schaffer 2004, Schmidt et al. 2010). Then, these models link daily survival values to survey period survival by exponentiating daily survival to the length of a survey interval. These methods are appealing for applications in ecological contexts because they account for uncertainty about when an individual entered the population and when within a later survey interval they die or change states (such as when a nest produces fledglings in a nest survival study) (Williams 2003, Mayfield 1975). They also allow for covariates to be finer scale than the entire lifetime of the individual. So, for example, a nest being monitored for nest fate may experience lower temperatures early in the season and warmer temperatures later in the season. Modeling survival with an interval-level temperature variable rather than an average temperature for the overall nesting period may give a better understanding of how temperature shapes nesting ecology.

Werewolf:

While methods that allow covariates to change across survey intervals are broadly used

and account for some error in these interval survival processes, these models are often defined such that each interval for each individual is a data point in a Bernoulli process. As a result, these datasets can have an exaggerated number of survival (success) values (1) since individuals that die (fail) are usually not monitored after death (failure). While this may be less of an issue when the number of survey intervals is low or when there is a relatively equal number of surviving and dead individuals in the dataset, as survey interval number increases or the survival-failure numbers are skewed, these models may not properly predict the data distribution of the data generating the model. Thus, using these models as predictive tools may be limited and biased.

Silver Bullet:

In this study, we present a novel individual-based Bayesian survival model that models final 1-0 fate data as a Bernoulli process dependent on a probability of survival that is dependent on surviving a previous set of survey intervals. In this way, covariates that change throughout these survey intervals are allowed to feed into the final survival probability of an individual. We compare this new model to two other approaches common in individual-based survival analyses. The first approach we compare our model to is a “full exposure model” that is a logistic exposure model that does not allow covariate values to vary throughout the survey period and where data are the final 1-0 fate data. The second model we compare our model to is a “interval-data exposure model” where covariates can vary throughout the survey period and data are 1-0 values at each survey interval. We present our examples in a Bayesian framework to account for parameter uncertainty and to aid in fitting custom probabilities.

We first fit these three models to simulated datasets with \_\_\_\_

We then compare the models for three empirical datasets from different applications tracking individual-based survival, including one dataset analyzing nest survival (white-headed woodpeckers in Oregon and Idaho, USA; Miller-ter Kuile et al. in review) and two datasets analyzing plant and tree survival (giant kelp in Santa Barbara, CA, USA (Emery et al. in review) and ponderosa pine trees in Arizona, USA (Rodman et al. in review)). Each of these applications varies in the number of individuals that survive to the end of the survey period, the number of survey intervals per individual, and the number of covariates that vary per survey interval.

**Methods**

*1. Explanation of the three models*

All three models are logistic regressions that account for different exposure lengths (Schaffer 2004, Schmidt et al. 2010).

*1.1 Total exposure model*

The total exposure model uses the final 1-0 fate data and covariates for each individual that are averaged across the entire survey period. In this model, survival probability is assumed to be constant across all days in the survey period (e.g., Mayfield 1975). Because the datasets we used were originally generated or collected at interval levels, in this case, the model is less prone to the biases of failure/success date uncertainty than in studies where individuals aren’t repeatedly surveyed (Shaffer 2004).

The final fate data for individual, *i,* (*yi*) are modeled with a Bernoulli process with total survey length survival probability, *pi*. Survival probability for the total survey interval is based on daily survival, *psi,* exponentiated to the length of the total survey length, *ti*. Daily survival (*psi*) is predicted from a logistic regression of *K* covariates that remain fixed for each individual throughout the total survey length.

*1.2 Interval data model*

The interval data model follows the specifications of nest survival methods (Shaffer 2004, Schmidt et al. 2010, Kozma et al. 2017) in which individuals are repeatedly surveyed throughout the total survey period, and these survey intervals are of varying lengths. In this model, covariates can be incorporated into the model as varying between individuals or across survey intervals for an individual. For example, these models can incorporate local environmental variables, such as canopy cover, that don’t change across the survey period while also incorporating variables, such as temperature, that do change across survey periods for an individual.

Each survey interval fate for each individual in each interval, *yi,j*, is modeled with a Bernoulli process with each interval’s survival probability, *pinti,j*. Survival probability for each interval is based on daily survival probability, *psi,j*, in that interval exponentiated to the total length of the survey interval *ti,j*. Daily survival within each survey interval (*psi,j*) is predicted from a logistic regression of *K* covariates that can either be fixed for each individual (*xi,k*) and *M* covariates that vary across survey intervals for an individual (*xi,j,m*).

*1.3 Custom model*

Like the total exposure model, the final fate data for individual, *yi*, are modeled with a Bernoulli process with total survey length survival probability, *pi*. Total survival probability is then dependent on the number of survey intervals for individual, *i*, and differs between individuals that were surveyed in one survey interval and greater than one survey interval. For individuals that were only resurveyed once, survival probability, *p1i*, is equivalent to the total exposure model, and the survival probability for that interval, *pinti,1,*is equivalent to daily survival probability in that interval, *psi,1*, exponentiated to the length of that one survey interval, *ti,1*. When each individual is surveyed more than once, total survey length survival probability (*pi*) is equal to *p2i*, which is a normalized probability that is dependent on the unnormalized probability of success (*q1i*) over the unnormalized probability of success (*q1i*) plus the unnormalized probability of failure (*q0i*). The unnormalized probability of success, *q1i*, is the product of all the interval survival probabilities for individual *i*. The unnormalized probability of failure, *q0i*, is the product of all the interval survival probabilities for all but the final interval times the probability of failure for the final interval. Like in the interval data model, all interval survival probabilities, *pinti,j*, is based on daily survival probability, *psi,j*, in that interval exponentiated to the total length of the survey interval *ti,j*. Daily survival within each survey interval (*psi,j*) is predicted from a logistic regression of *K* covariates that can either be fixed for each individual (*xi,k*) and *M* covariates that vary across survey intervals for an individual (*xi,j,m*).

*2. Applying the models to simulated data*

- only interval level predictor data

- vary the variability in these data

- high, med, low

- make sure variable has effect through time and also has different amounts of variability

- struggling: how to simulate a set of 1s and then a 0/1 for each individual while linked up to a covariate??

To examine each model on each simulated dataset, we calculated an AUC value from a set of \_\_\_ iterations from an updated version of the converged model (using the update() function in the jagsUI package ()). We then determined the predictive accuracy of each model in predicting each fate class (1-0) based on a cutoff of survival probability of 50% (CITE). Finally, we qualitatively assessed the difference in covariate effects using covariate p-values for each covariate in each model.

*3. Applying the models to real datasets*

Finally, we fit each model to three real datasets from a broad set of ecological scenarios. These datasets included a dataset of tracked 1) nests of white-headed woodpeckers in Oregon and Idaho, USA (Miller-ter Kuile et al. in review); 2) giant kelp plants from Santa Barbara, California, USA (Emery et al. in review); and 3) ponderosa pine trees from Arizona, USA (Rodman et al. in review). For each of these datasets, we fit each model to the data with the same set of environmental covariates. For both the nest and kelp datasets, *psi* or *psi,j*, represents daily survival; for the tree dataset, *psi* or *psi,j* represent yearly survival. For the total exposure model for all datasets, we averaged any interval-level environmental covariates across the total survey period and used the total length of the survey period for each individual as the exposure value. All models for each dataset included the same hierarchical random effects structure that ensured identifiability of random effects (Ogle and Barber).

To examine each model, we performed the same steps as for the simulated datasets. Specifically, we calculated an AUC value from a set of \_\_\_ iterations from an updated version of the converged model. We then determined the predictive accuracy of each model in predicting each fate class (1-0) basd on a cutoff of survival probability of 50% (CITE). Finally, we qualitatively assessed the difference in covariate effects using covariate p-values for each covariate in each model.

We ran each model in R (cite) and JAGS (cite) using the jagsUI package (cite). We prepared data using the here () and tidyverse () packages. We validated model fit using the coda () and mcmcplots () packages. All figures were made with the ggplot2 () and patchwork () packages. We ran all models on the computing cluster, Monsoon, at Northern Arizona University. We ran each model on three MCMC chains for an initial run of 4000 iterations per chain. We used the raftery.diag() function in the coda () package to calculate the Raftery diagnostic (cite), which determines the sufficient burn-in and iteration number for convergence and re-ran each model accordingly.

**Results**

**Discussion**

**References**

**Figures and Tables**

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 1: The distribution of AUC values for each model for each dataset. A, C, and E are results from the nest survival dataset, B, D, and F are results of the kelp survival dataset. A and B are AUC values for the total exposure models. C and D are the AUC values for the interval data model (darker purple = AUC of model when only considering last interval data; lighter purple = AUC of model when considering data from all intervals). E and F are the AUC values for the custom model that uses data from the last survey interval but accounts for changes in covariates throughout different survey intervals.

A comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of

Description automatically generated with medium confidence

Figure 2: Distributions of AUC values for the total exposure and custom models for (A) the nest survival dataset and (B) the kelp survival dataset.

A picture containing text, diagram, plan, parallel

Description automatically generated

Figure 3: Predictive accuracy of each model for each dataset, determined by the relationship between predicted survival probability and observed fate (0 = dead, 1 = alive). Dashed lines in each figure represent 50% survival probability. Results from the nest survival dataset for the total exposure model (A), interval data model (B), and the custom model (C) that incorporates data from the final survey interval but allows for covariates to vary across survey intervals. Results from the kelp survival dataset for the total exposure model (D), interval data model (E), and the custom model (F) that incorporates data from the final survey interval but allows for covariates to vary across survey intervals.

A picture containing text, diagram, line, screenshot

Description automatically generated

Figure 4: Covariate Bayesian p-values and directions for the nest survival (A) and kelp survival (B) datasets. Colors and shapes represent different models (green circle = total exposure model, purple triangle = interval data model, and orange square = custom model with final survey data and covariates that vary through different survey intervals). The grey rectangles represent areas in which p-values are not significant at p < 0.05. Any p-values to the left of the middle line represent negative covariate effects; p-values to the right of the middle line represent positive covariate effects.