Target Journal: Methods in Ecology and Evolution (7000-8000 words including table/figure captions and references)

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 Kuile1,2, Kiona Ogle1, Jamie Sanderlin2

1 School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona, USA

2 USDA Forest Service Rocky Mountain Research Station, Flagstaff, Arizona, USA

Corresponding Author:

Ana Miller-ter Kuile

School of Informatics, Computing, and Cyber Systems

Northern Arizona University

1295 Knoles Dr

Flagstaff, AZ 86011

[Ana.miller-ter-kuile@nau.edu](mailto:Ana.miller-ter-kuile@nau.edu); [ana.kuile@usda.gov](mailto:ana.kuile@usda.gov)

Running headline: Individual-based survival model

**Abstract**

1: set the context for and purpose of the work;

2: indicate the approach and methods;

3: outline the main results;

4: identify the conclusions and the wider implications.

**Keywords**

**Introduction**

Baby:

Survival analyses have broad application across many biological fields, from human medicine to ecology (CITE). Survival analyses are important for understanding many processes, from understanding how quickly people recover from diseases like cancer (“time to event” analyses; CITE), to how environmental variables shape the survival of trees following forest management (CITE paper from Kyle). Given the importance to understanding survival (and mortality) dynamics across ecological disciplines to a variety of pressing challenges, including anthropogenic change drivers, it is imperative to have models that represent the ways in which fields collect data and how survival changes across individuals and across time (Mantgem et al. 2009).

Zoomed in baby:

There are different analyses created for different types of survival analyses. While many studies examine survival probabilities for larger groups of individuals (e.g. stand-level tree mortality; Sheil et al. 1995, Lutz and Halpern 2006), individual survival models provide important mechanistic insights relating to the effects of individual traits and local-scale habitat features on individual survival (Zens and Peart 2003). Many traditional individual-based analyses assess survival across one survey interval, usually with one “treatment” applied to each individual throughout that entire interval and either outcomes or time to outcomes are measured for different individuals/data points (CITE, look for papers on logistic exposure models too). Others, such as in wildlife biology and forest ecology, try to link individual-based survival to finer-scale covariate data that is likely to change throughout the survey period (CITE). To do this, many studies employ a modified logistic regression model to survey data that takes place over one or multiple sub-intervals throughout the survey period for each individual being tracked (Mayfield 1975, Schaffer 2004, Schmidt et al. 2010). These models link daily survival values to survey period survival by exponentiating daily survival to the length of a survey interval (sometimes called “logistic exposure models”). 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) (“right-censored data”; Lindsey and Ryan 1998, 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 individual-based survival models that allow covariates to change across survey intervals are broadly used and account for some error in these interval survival processes, the two common ways these models are often employed have some shortcomings. First, models that consider survival to be constant throughout the survey period and exponentiate the logistic regression to a total (potentially relatively long, e.g. a lifetime) survey period do not account for potential changes in individual traits or environmental covariates throughout the total survey period, potentially missing important effects of these covariates. These models may only be appropriate in contexts where covariates are fairly constant throughout the entire survey period. Second, models that take into account for changes in environmental and individual traits throughout a set of survey intervals for each individual are 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, either total exposure or interval-data exposure models may provide inaccurate information about the effects of covariates on survival of individuals and may poorly fit the models used to define them, thus limiting or biasing any predictive use of these models to new datasets and contexts.

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 with a survival probability that is dependent on surviving a previous set of survey intervals. In this way, covariates that change throughout these survey intervals are allowed to contribute to the final survival probability of an individual. We compare this new model to two other logistic exposure 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 (CITE). 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 (Shaffer 2004, Schmidt et al. 2010, tree citation here). 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 three sets of simulated datasets where survival was dependent on a covariate which varied across individuals in the dataset. For each of the three sets of simulated datasets, we varied the degree to which the covariate varied within an individual across survey intervals (“low”, “medium”, and “high”). We then compare the three models for three empirical datasets from different applications tracking individual-based survival, including one dataset analyzing daily nest survival (white-headed woodpeckers, *Dryobates albolarvatus*, in Oregon and Idaho, USA; Miller-ter Kuile et al. in review), one analyzing daily plant survival (giant kelp, *Macrocystis pyrifera*, in Santa Barbara, CA, USA; Emery et al. in review), and one analyzing yearly tree survival (ponderosa pine, *Pinus ponderosa*, 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, the number of covariates that vary per survey interval, and the relative effects and variation in covariates across individuals and sampling surveys. [Some sentence about what we can learn from this approach.]

**Materials and Methods**

*1. Explanation of the three models*

All three models are based on logistic regression. They are all formulations of the logistic exposure model that account for different exposure lengths across individuals or survey intervals (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 use in our examples were originally generated or collected at interval levels, 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 and across survey intervals for an individual. For example, these models can incorporate local environmental variables, such as habitat type, that may not 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 are fixed for each individual (*xi,k*) and *M* covariates that vary across survey intervals for an individual (*xi,j,m*).

*1.3 Custom model*

The custom model combines aspects of the total exposure and interval data models by modeling final fates only but incorporating interval-specific covariates into the calculation of the total survey survival probability for each individual.

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 for only one survey interval and individuals with 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 an 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*) and 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*, are 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 are fixed for each individual (*xi,k*) and *M* covariates that vary across survey intervals for an individual (*xi,j,m*).

*2. Software and model running*

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 for real datasets 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. We then ran each model for sufficient iterations for convergence and verified convergence with the Gelman diagnostic ().

To examine each model on each dataset (both simulated and real datasets), 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 simulated data*

We first fit all three models to simulated datasets where the effect of a centered and scaled covariate, *x*, on survival probability, was pre-defined based on a logistic regression with an intercept and covariate parameter.

In this model, *p* represents daily survival probability, beta0 is an intercept and beta1 is the effect of parameter x on daily survival probability. L is the length of time of the survey interval, which we defined as 1 unit for all intervals when simulating the data.We predefined beta0 and beta1 to ensure that individuals had a generally high likelihood of survival. (beta0 = 1.38, beta1 = 0.5). We generated threecovariate datasets with varying degrees of variability among survey intervals centered around an individual-level mean value for the covariate. The means for each individual were broadly normally distributed on the scale of the data (N(mean = 0, sd = 1)). These three datasets represent data where the covariate, *x*, has a) low variability across survey intervals for an individual N(mean = meani, sd = 0.01), b) medium variability across survey intervals for an individual N(mean = meani, sd = 0.5; roughly spanning one unit of x on either side of the mean), and c) high variability across survey intervals for an individual, such that survey-to-survey variability within an individual is comparable in magnitude to the difference among individuals (N(mean = meani, sd = 1)).

From these three covariate datasets, we used the logistic regression equation in Equation (X) to generate survival probabilities, p, for each individual in each survey interval. We then predicted binomial 1-0 survival data from these survival probabilities based on Bernoulli probability (Yi ~ Bern(pi)). We removed any survey intervals for an individual following a simulated 0 value. We repeated this process for a total of 100 datasets of y data for each level of variation in the x covariate.

We fit each of the three survival models to each of these sets of datasets and assessed fit by determining [what??] how often the model predicted the correct values for b0 and b1?? How to deal with 100 median estimates + BCIs of each b0 and b1? How to assess and/or visualize?

When models predicted similar parameter estimates – do we repeat the AUC/predictive assessments as those done for the real datasets??

*4. 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 (*Dryobates albolarvatus*) in Oregon and Idaho, USA (Miller-ter Kuile et al. in review); 2) giant kelp plants (*Macrocystis pyrifera* ) from Santa Barbara, California, USA (Emery et al. in review); and 3) ponderosa pine trees (*Pinus ponderosa*) 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* in Eqns [X], represents daily survival; for the tree dataset, *psi* or *psi,j* in Eqns [X] 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).

**Results**

**Discussion**

**Acknowledgements**

**Conflict of Interest Statement**

**Author Contributions**

**Data Availability**

**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. For ease of interpretation, we only included covariates with p-values <= 0.05 for at least one of the three models. 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.

Figure Ideas for simulation dataset

How often are b0 and b1 within the credible interval of the model for the dataset?

* How to show something where each dataset’s median and 95% CI is represented? Or is it okay to just have a distribution of medians? Hmm…

AUC/Predictive accuracy – how to break up by “dataset”?