Dealing with incomplete data in survival analysis: A model based on probabilistic prior information on time of occurrence

R. Manso [a], M. Fortin [a], M. Pardos[b], R. Calama[b]

[a] INRA-AgroParisTech UMR 1092.
Laboratoire d'Etude de Ressources Forêt-Bois (LERFoB)
54280 Champenoux, France
ruben.manso@nancy.inra.fr
mathieu.fortin@agroparistech.fr

[b] Department of Silviculture and Forest Systems Management.
INIA-CIFOR
Ctra. de La Coruña km 7.5 28040 Madrid. Spain
pardos@inia.es
rcalama@inia.es

Keywords: individual-based models, survey design and analysis, proportional hazards, left-truncation

Abstract:
Efficient methods have been described to deal with incomplete data when modelling survival through lifetime analysis. However, some cases of left-truncated data may still be a source of bias in parameter estimation of these models. This is especially true in plant survival studies where the time of occurrence (i.e. establishment) t_i of an observed individual i is unknown. In the present work we propose a method to cope with this shortcoming based on prior probabilistic information on t_i. In order to illustrate it, a dataset of *Pinus pinea* seedling survival in Central Spain was used. The data consisted on interval- and right-censored observations. Left-truncation was also present because seedlings did not enter the experiment at germination time (t_i), which was also undefined. A proportional hazards model was fitted to these data, the likelihood function being defined accordingly as to consider all forms of data incompleteness. However, the likelihood for seedling i cannot be referred to a specific time of emergence, since there are as many values of t_i as days where germination was possible. Therefore, the likelihood for all possible values of t_i needed to be calculated, the resulting outputs being weighted by the corresponding germination probability obtained from a predicted germination probability mass function (pmf_i), and eventually summed over. In our case, pmf_i was derived from an existing germination model for the species but whatever sensitive prior probability mass function could have been used. Provided that the exact information on t_i is not available in most datasets based on field experiments, the proposed methodology may be of great interest for ecologists, overcoming the current limitations of survival analysis.