FUNDAMENTAL RESEARCH

For. Sci. 60(4):613–622
http://dx.doi.org/10.5849/forsci.12-133

biometrics

Predicting Tree Mortality for European Beech in Southern Germany Using Spatially Explicit Competition Indices

Andreas Boeck, Jochen Dieler, Peter Biber, Hans Pretzsch, and Donna P. Ankerst

Individual tree mortality prediction is a key component of single tree-based stand simulators. However, accurate modeling of long-term research plot data is hampered by rare events, variable lengths of observation, and multiple sources of heterogeneity. This study makes use of a result from medicine that demonstrates the equivalence of logistic and Cox proportional hazards regression for modeling survival data in the case of large sample sizes, rare events, and variable interval periods of observation. Pooled logistic regression models are used to model tree mortality across multiple observation periods with random effects to account for heterogeneity due to plots and calendar year. The models are applied to data from 21,051 observation periods (each approximately 5 years) from 9,292 beech trees in a Bavarian long-term forest research plot network. Among the observation periods studied, 604 (2.9%) resulted in a mortality. Indices measuring competition from light, trees of the same species, conifer trees, and shading are significantly associated with mortality, whereas other variables, including dbh, fail to add additional predictive value. Analytic equations for predicting mortality in new trees are provided and yield an area underneath the receiver operating characteristic curve of 91.5%.

Keywords: *Fagus sylvatica* L., logistic regression, survival, competition index, dbh

Individual mortality prediction is an essential component of single tree-based forest growth models, including the simulator SILVA currently used in southern Germany (Pretzsch et al. 2002). There are data and statistical challenges to mortality modeling. Although data from large numbers of regularly monitored long-term research plots may be available, their information content is hampered by low numbers of events, in this case, deaths of trees. The situation is exacerbated for estimating the combined effects of risk factors on rare events. An additional challenge is that the duration of follow-up of individual trees often varies and comprises several periods. This raises the question of how to account for dependencies between multiple observations on the same tree. Different statistical methods have been used in the forest science literature for predicting individual tree mortality, the primary being logistic regression. To avoid the dependence issue, typically a cross-sectional approach is chosen, whereby a single period of observation is selected for analysis and the rest are discarded. The interval may be selected at random for each tree or nonrandomly as the one that ultimately resulted in mortality of the tree. The latter results in an overestimation of the mortality rate and is biased.

The Cox model for survival data is a ubiquitous method used in medicine for modeling time to mortality (Cox 1972). Rather than modeling the dichotomous event, alive versus dead, it models the time until mortality. Importantly, it correctly accounts for unobserved event times due to censoring. There are three types of censoring. Interval censoring refers to situations in which the time of the event is known to fall within a window of time but not known exactly when during that interval. Right-censoring occurs when an individual is known to not have experienced the event up until a certain time, but not known when afterwards, and left-censoring is the reverse. To use the Cox model, it is necessary to assume that the censoring mechanism is independent of the time until the event process. The Cox model relates the hazard or instantaneous rate of mortality at any time to covariates. Alternatively, logistic regression relates the probability of death in a single interval to covariates. Unfortunately, survival models are not ideal for situations in which there are long periods between monitoring and low event rates, as could be expected in this application.

This article outlines an individual tree mortality prediction strategy that is tailored for data from multiple periods of observations in

Manuscript received November 5, 2012; accepted July 1, 2013; published online August 8, 2013.

Affiliations: *Andreas Boeck (andreas.boeck@tum.de), Technical University Munich. Jochen Dieler, Technical University Munich. Peter Biber, Technical University Munich. Hans Pretzsch, Technical University Munich. Donna P. Ankerst (ankerst@tum.de), Technical University Munich, Mathematics, Garching, Germany.*

Acknowledgments: *The authors are grateful for the helpful comments from an associate editor and reviewers. The authors thank the Bavarian State Ministry for Nutrition, Agriculture, and Forestry for permanent support of the project W07: Long-term experimental plots for forest growth-and-yield research.*

Table 1. Definitions of variables and risk factors used in the mortality analysis.

Characteristic	Definition
PeriodOnset	First year of survey period
PeriodOffset	Last year of survey period
PeriodLength	Length of the period of observation in years
Dbh	Diameter at breast height (1.3 m) in cm
Height	Tree height in m
KKL[a]	Quantifies light competition by neighboring trees
CIIntra[a]	Competition from trees of the same species as the tree of interest
CIConifer[a]	Competition from conifer trees
CIOvershade[a]	Expresses to what extent a tree is over-shaded by other trees
CILateral[a]	Lateral competition of a tree
Dbhdom	Estimation of the dbh (in cm) a tree would have at its current height if predominant for its whole life [$= 0.6553 \times \text{height}^{1.327}$]
Reldbhdom	Ratio dbh/dbhdom that measures long-term competition
SiteIndex	Plot- and species-wise site index, expressed as stand height in m at age 40 (derived from standard yield tables)

[a] See Appendix.

long-term research plots. The approach is based on a result previously proven and applied in medicine. This useful result is that the Cox regression model is equivalent to the logistic regression model in the case of large data sets, interval monitoring, and rare events (Abbott 1985, D'Agostino et al. 1990). In this forest application, a pooled logistic random-effects model is fit to data derived from 21,051 periodic measurements obtained at approximately 5-year intervals on 9,292 beech trees in a network of Bavarian forests. The model-fitting process is evaluated using internal cross-validation. The models are applied to improve the individual-tree mortality component of the existing forest simulator SILVA (Pretzsch et al. 2002). However, the general methodology is applicable to the development of mortality risk prediction models based on commonly collected attributes, such as dbh. Unlike most studies, the data here comprise individual tree positions, which enable the inclusion of spatially explicit individual-tree competition indices as covariates.

Materials and Methods

Data

Data were collected from beech trees taken from 60 plots at 11 test sites in Bavaria, Germany, that were undergoing surveillance from 1954 until 2007. Individual trees were observed for between one and seven observation periods during these years, with observation periods ranging from 3 to 28 years (most were 5 years). Observation periods during which the tree experienced mortality through man-made thinning or natural disasters, such as storms, were excluded. Only individual trees that had information on the risk factors defined in Table 1 at the beginning of an observation period and mortality (yes versus no) at the end of the same observation period were included in the analysis. Risk factors considered in the prediction models comprised measures of the size of individual trees, a set of indices covering different aspects of competition, site quality information, calendar year, and length in years of the individual observation periods. In total, 21,051 single-tree observation periods comprising 9,292 beech trees from 60 plots were available for analysis. Of the 21,051 observation periods, 604 (2.9%) were associated with mortality.

For the model selection component of the analysis, 29 plots were included that had a minimal mortality of 1% for all observation periods. The model selection was performed on this subset to reduce convergence problems in estimation of plot random effects arising from plot observation periods that had negligible or no mortality. Once the final model form was selected, it was refit on the entire data set comprising all 60 plots.

Tree size as a risk factor was expressed by the dbh. Age was not available for the trees in this study, and, furthermore, it is not always available to the forest manager. However, age inevitably correlates with tree size. Tree height was another measure for size, but because in these data height was only measured on a subsample of trees and estimated for the others, it was not preferred over dbh.

Competition was divided into two aspects: momentary competition and the long-term competition a tree has had in the past. Although the former can be strongly influenced by ad hoc thinnings, the latter expresses the typical competition a tree has undergone during its life. For quantifying momentary competition, KKL, a simple geometric competition index, and a set of indices derived from local vertical competition profiles were used (Pretzsch et al. 2002; see Appendix). The indices CIIntra and CIConifer were derived from a general competition index, called CICUM60, which is similar to KKL and designed to measure overall momentary competition. CIIntra was the component of CICUM60 attributable to trees that belong to the same species as the tree of interest. CIConifer represented the portion of CICUM60 that originates from conifer species. The basic concept of vertical competition profiles permitted separation of two other important aspects of momentary competition: overshading and lateral constriction, expressed by the indices CIOvershade and CILateral, respectively (Assmann 1961, Pretzsch 1992). Further details on the competition indices can be found in the Appendix.

For long-term competition, a different concept that compared actual tree size to a reference tree size was needed. If a given tree size was small compared with a reference tree size, the tree must have had strong competition in the past and vice versa. Because trees under competition show a reduction in diameter increment more than in height increment, the dbhdom measure was used as a reference. This measure was defined as the dbh a predominant (low long-term competition) tree had at a given height and was estimated as follows. From a subsample of the data, the allometric relationship, $\text{dbhdom} = 0.6553 \cdot \text{height}^{1.327}$, was estimated (with the units of m for height and cm for diameter) and used to estimate the dbhdom a tree could have achieved at its current height under very low competition during its life up until the present. Dividing the tree's current dbh by the estimated dbhdom yielded the measure reldbhdom. Low values of reldbhdom indicated that the tree had undergone stronger long-term competition, whereas larger values near or even exceeding 1 indicated the tree had not undergone much competition throughout its life. Finally, site quality (SiteIndex) was expressed through the expected mean stand height in m at age 40 years based on the yield table for European beech of Schober (1967).

Statistical Models

As an initial exploratory analysis, risk factors and observational characteristics were compared between tree observation periods with and without mortality using means, standard deviations (SDs), and ranges. Differences in risk factors between tree observation periods that resulted in mortalities versus nonmortalities were tested for statistical significance using the nonparametric Wilcoxon test. Multiple observation periods for individual trees were treated as

independent, whereas lengths of tree observation periods were ignored. The Wilcoxon statistic was transformed to the area underneath the receiver operating characteristic (ROC) curve (AUC) (Faraggi and Reiser, 2002). ROC curves and AUCs are statistics often used in diagnostic medicine to evaluate risk factors or computed probabilities (risks) of diseases. The concepts can be transferred for evaluating the predictive value of tree characteristics for mortality. Without loss of generality suppose that high values of a tree characteristic are associated with mortality, such as high values of a competition index X (for the reverse, where low values of a characteristic are associated with mortality, inequality signs in the following are reversed). A diagnostic test for mortality based on X might be defined as a tree testing positive for mortality whenever X exceeds a threshold c ($X > c$). The sensitivity of this threshold is defined as the proportion of all tree observation periods with a mortality where $X > c$, and the specificity as the proportion of all tree observation periods without mortality where $X \leq c$ (test negative). Sensitivities and specificities range from 0 to 100%; higher values of both imply a better test for distinguishing tree observation periods that result in mortalities from nonmortalities. The ROC curve is a plot of 100% − specificity on the x-axis (called the false-positive rate) versus sensitivity on the y-axis for all possible thresholds c. The AUC is the area underneath the ROC curve and ranges from 50 to 100%. Interestingly, the AUC has a partner mathematical definition: the AUC equals the probability that for a randomly chosen tree observation period that resulted in mortality and a randomly chosen tree observation period that did not result in mortality, the former has a higher value of X. An AUC close to 100% indicates good discrimination of X for mortality, whereas an AUC close to 50% indicates that the risk factor exhibits no better discriminating ability between observation periods with mortality versus nonmortality than random choice. The P value reported for the Wilcoxon test is also the P value for a test of the null hypothesis that the AUC equals 50% versus the alternative that the AUC exceeds 50%. The ROC and AUC also apply for the case where X is a predicted probability of mortality, such as that arriving from a model fit to a training set of trees. In this case, however, it is important that a set of trees (validation or test set) separate from that used to build the model for X be used to evaluate the ROC and AUC. This is necessary to avoid overoptimism from fitting and evaluating the model on the same data set.

The statistical model used for the association of risk factors collected at the beginning of an observation period with mortality by the end of the same period was adapted from an application in public health. Faced similarly with large sample sizes and rare events in their analysis of cardiovascular events in the Framingham Heart Study, Abbott (1985) and D'Agostino et al. (1990) demonstrated the asymptotic equivalence of the grouped Cox proportional hazards survival model (Cox 1972) to pooled logistic regression, where, in the latter, multiple observation periods per individual were treated as independent. In this analysis, to account for unequal length of observation periods, a fixed offset term, called the observation length, was applied as commonly performed in relative risk modeling in epidemiology. Risk factors were modeled as fixed effects and plots as random effects. The initial year of each observation, referred to as the calendar effect, was also included as an independent random effect to the plot effect, to serve as a proxy for long-term effects such as global warming. Plot and calendar year random effects were assumed to follow independent normal distributions centered at 0 with respective plot and calendar year SDs.

Selection of the optimal transformation of risk factors to include in the logistic regression was performed in two stages. The first stage (stage I) did not use the mortality endpoint but rather focused on transforming the individual risk factors so that their empirical distributions would be as close as possible to a normal distribution. Normality of the risk factors is not a requirement for logistic regression; however, a symmetric and compact design space can improve the fit of the model. Three of the risk factors, KKL, CILateral, and CIConifer, had a disproportionately large number equal to 0 (Table 1). The zero values were removed to achieve an optimal transformation. The transformations considered were power transformations for which power could range from 0.01 to 1. The Kolmogorov-Smirnov test for normality was used to find an optimal power transform with the transform corresponding to the smallest value of the Kolmogorov-Smirnov test statistic declared as optimal. The optimal power was rounded to the closest fraction, and the variable was transformed by this power for all further analyses.

In the second stage (stage II), cubic B-splines with a second-order penalty were used to find the optimal relationship of transformed risk factors to mortality (Eilers and Marx 1996). This stage was performed by fitting one risk factor at a time in a logistic regression model. For risk factors with point masses at 0 (KKL, CILateral, and CIConifer), two alternative risk relationships were considered: the first, an odds ratio (OR) for the risk factor equal to 0 versus greater than 0, and the second, the same OR as in the first, along with an additional OR for a unit increase in the risk factor if the risk factor was greater than 0. The optimal risk relationship was selected based on goodness-of-fit assessments as described below. Empirical curves showing the relationship of risk of mortality to the risk factors in the data were constructed using smoothed lowess curves. Then multiple B-spline fits were overlaid on the empirical risk curves to visually inspect the goodness of fit across all values of the risk factors. B-splines have a complicated representation depending on the choice of base functions. In this application B-splines were used only to find the closest polynomial transformation that would optimize the fit of the mortality model to each of the individual risk factors, for example, to decide whether a quadratic relationship of dbh to mortality was required, implying estimation of ORs for dbh and dbh^2. The Bayesian information criterion (BIC) was used for the final determination of the optimal transformation, such as a linear versus quadratic transformation. The BIC equals $-2 \times$ log-likelihood + $k \times (\log n)$, where k is the number of parameters in the model and n is the sample size. Models with the smallest BIC were chosen as optimal. The BIC chooses models with large maximized likelihoods and penalizes models with many parameters. The models with the smallest BIC are most likely to externally validate.

In sum, the steps above resulted in a generalized linear mixed model with multiple risk factors, relating the probability π_{ijk} of mortality for tree j in plot i during observation period k to risk factors measured at the beginning of the observation period as follows

$$\text{Logit}_{ijk} = \log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \beta_0 + \text{offset}\left(\log\left(\frac{\text{observation length}_{ik}}{5}\right)\right) + \beta_1' t(\text{dbh}_{ij}) + \beta_2' t(\text{Height}_{ij}) + \ldots + \beta_{10}' t(\text{SiteIndex}_i) + \gamma_k + \gamma_i,$$

where β_0 is the global intercept of the model, β_1 is a vector of regression parameters corresponding to the chosen transform of dbh [e.g., β_1 would comprise two terms if the quadratic transformation $t(\text{dbh}) = (\text{dbh}, \text{dbh}^2)$ was used] and similarly for β_2 through β_{10}.

The offset function is the identity function, offset(x) = x. A 5-year observation period receives an offset of 0, so that the remaining parameters in the model hold for this period of observation. The variable γ_k is a normally distributed random effect with mean of 0 and SD σ_{year} [$N(0, \sigma_{\text{year}})$] that marks the calendar year of the start of the observation period (one of 1954, 1959, 1969, 1982, 1984, 1985, 1987, 1989, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1999, or 2000). The random plot effect γ_i is similarly distributed as $N(0, \sigma_{\text{plot}})$. The expression above implies that all 10 of the risk factors (Table 1) appeared in the final model, but reduction was performed to arrive at a parsimonious one that is more likely to be accurate on external validation. This process is described next.

Selection of Risk Factors

Internal K-fold cross-validation was used to select the optimal set of risk factors to include in the mortality prediction model. Because a large number of models needed to be evaluated and to ease convergence issues, only 29 of the 60 plots with mortality rates exceeding 1% over all of their respective observation periods were used for this portion of the analysis. Furthermore, because cross-validation was used here for fitting the model and not for final model validation, $K = 29$ corresponding to the 29 plots represented in the data, was chosen. Each of the 29 plots served in turn as a single test data set with the remaining 28 plots combined as a training set. The cross-validation scheme reflected the purpose of the model, which was to predict 5-year mortality for a new tree in a new plot. Following Skrondal and Rabe-Hesketh (2009), median prediction (which involves setting random effects to their prior median of 0) rather than mean prediction (which involves averaging over the distribution of the random effects) was used. One reason for favoring median over mean prediction was that it is more robust to outliers that may have influenced the fitting of the normal distribution assumed for the random effects. A second reason was that the objective function used to evaluate goodness of prediction for the internal validation was the AUC criterion, which like the median is a rank-based measure. A final advantage of median prediction was that the predicted mortality had an analytical expression that did not require numerical or simulation methods for evaluation. Secondary analyses (not shown) revealed near negligible differences in predictions and AUCs from the use of mean versus median prediction at this step of the analysis. For each training set, a set of candidate models were fit, and parameter estimates were used to predict the mortality for trees in the corresponding validation set. To reduce the influence of multicollinearity among the risk factors on stability of the model selection process, Spearman correlation coefficients among the transformed risk factors were computed. Models containing two risk factors with correlations exceeding 0.75 in absolute value were dropped from further consideration.

The predicted probability of mortality for a new tree was set equal to $\hat{\pi}_{ijk} = \exp\{\text{logit}_{ijk}\}/[1 + \exp\{\text{logit}_{ijk}\}]$, where logit_{ijk} is the right-hand side of the logistic model equation above with plot and calendar year random effects set to 0 and fixed-effect parameters estimated from a fit of the candidate model to the training set. These predicted probabilities were compared against the actual mortality outcomes in the test set using the AUC. For each validation set, the optimal model was chosen as that which achieved the highest AUC averaged over the 29 training-test set pairs. Alternative validation criteria that are combined measures of discrimination and calibration (closeness of predicted to observed risks), such as the Brier score and pseudo R^2, were also considered (Steyerberg 2010, Nagelkerke 1991). However, these generally led to the same selection of optimal models as the AUC and consequently they were not reported.

Based on the average of cross-validation results from the subset of 29 plots, an optimal model was selected. The mortality risk prediction from the final model was given by $\hat{\pi}_{ijk} = \exp\{\text{logit}_{ijk}\}/[\{1 + \text{logit}_{ijk}\}]$, with estimated parameters from the fit of the model to the entire data set of 60 plots. Confidence intervals for the prediction were determined using the delta rule applied to the variance-covariance matrix of all parameters estimated as part of the logistic regression; see the Appendix for specific details. Because transformations and polynomial risk relationships complicate interpretation of ORs, examples of risk plots with pointwise 95% confidence intervals are displayed.

All statistical computations were performed with the R Statistical Package version 2.14.2 (R Development Core Team 2012), including the packages mgcv (Wood 2011), ggplot2 (Wickham 2009), and ROCR (Sing et al. 2009). All statistical tests were performed at the $\alpha = 0.05$ level of statistical significance.

Results

Descriptive Analyses

For all risk factors, there was a statistically significant difference between mortality and nonmortality observation periods (all AUC, $P < 0.001$ except for CIConifer, $P = 0.52$) (Table 2). The average dbh of trees that experienced mortality at the end of an observation period was 7.2 ± 4.5 cm (mean ± SD) and was significantly less than that of observation periods that did not result in mortality (19.7 ± 13.4 cm). Thus, the discriminatory power of dbh alone for the prediction of tree mortality was high, with an overall AUC of 83.3%. The AUC of dbh was also consistently high among individual plots, with the lowest plot AUC equal to 67.9%. Similarly, height was lower among mortality compared with nonmortality tree observation periods (10.7 ± 4.5 m versus 18.9 ± 8.3 m), but it had lower discriminatory ability than dbh (overall AUC, 80.0%; minimum plot AUC, 58.8%). The other two dbh measures were also reduced in mortality observation periods and had AUCs very similar to that of dbh. The two competition indices, KKL (AUC, 82.5%; minimum plot AUC, 61.8%) and CIOvershade (AUC, 84.7%; minimum plot AUC, 55.3%), had the highest overall AUCs of all risk factors, although their plot-specific AUCs dipped below that for dbh (Table 2). The observed differences indicated that, as expected, smaller-sized trees that experienced more competition for light from neighboring trees or were more overshaded by other trees were at increased risk of mortality. The remaining three competition indices, CILateral, CIIntra, and CIConifer, had the lowest AUCS (77.3, 71.1, and 50.7%, respectively) and for some plots had little to no improved discriminatory ability for predicting mortality over random chance (minimum plot AUCs, 51.0, 50.7, and 50.0%). The SiteIndex was lower among nonmortality than mortality periods (AUC 63.5%). The length of observation periods was slightly higher for periods associated with mortality than nonmortality (AUC $P = 0.001$). Observation periods more recent in date were more likely to be associated with mortality than earlier ones (AUC, 61.2; $P < 0.001$).

The optimal normality transformation method (stage I) resulted in the following power transformations of the risk factors for subsequent use in the risk models: $\text{dbh}^{3/20}$, $\text{height}^{2/3}$, $\text{KKL}^{1/3}$, $\text{dbhdom}^{1/2}$, $\text{CILateral}^{3/20}$, $\text{reldbhdom}^{2/3}$, $\text{CIOvershade}^{1/2}$, $\text{CIIntra}^{1/2}$, and $\text{CIConifer}^{1/3}$.

Table 2. Characteristics of trees in observation periods associated with mortality versus no mortality.

	Nonmortality periods, mean (SD) [range] (N = 20,447)	Mortality periods, mean (SD) [range] (N = 604)	AUC (%) (P value) (N = 21,051)	Plot-specific AUCs (%), minimum, median, maximum (N = 14,239)[a]
Dbh	19.7 (13.4) [0.8–90.9]	7.2 (4.5) [0.9–37.9]	83.3 (<0.001)	67.9, 86.6, 97.2
Height	18.9 (8.3) [1.4–43.9]	10.7 (4.5) [1.4–27.1]	80.0 (<0.001)	58.8, 83.3, 100.0
KKL[b]	3.2 (4.9) [0.0–60.5]	9.5 (9.2) [0.3–65.5]	82.5 (<0.001)	61.8, 82.1, 99.0
CIIntra	128.4 (73.4) [5.9–517.6]	184.6 (85.3) [14.6–444.4]	71.1 (<0.001)	50.7, 57.8, 82.7
CIConifer	13.7 (22.0) [0.0–200.4]	13.4 (20.4) [0.0–120.2]	50.7 (0.52)	50.0, 54.5, 80.9
CIOvershade	86.3 (71.4) [0.0–505.9]	188.7 (81.7) [21.9–461.4]	84.7 (<0.001)	55.3, 80.1, 97.4
CILateral	54.4 (60.6) [0.0–436.9]	10.9 (31.6) [0.0–257.6]	77.3 (<0.001)	51.0, 79.8, 100.0
dbhdom[c]	40.9 (22.8) [1.3–117.7]	19.5 (10.6) [1.3–62.8]	80.0 (<0.001)	58.8, 83.3, 100.0
Reldbhdom	0.5 (0.1) [0.2–1.1]	0.4 (0.1) [0.2–1.0]	74.8 (<0.001)	58.0, 78.3, 96.8
SiteIndex	14.7 (3.7) [5.5; 22.5]	16.4 (4.6) [5.5; 22.5]	63.5 (<0.001)	
Observation length	5.5 (2.2) [3–28]	5.3 (1.6) [3–10]	50.6 (0.001)	50.0, 57.7, 80.1
Year of period onset	1993 (8.1) [1954–2000]	1995.8 (4.4) [1985–2000]	61.2 (<0.001)	50.0, 62.6, 80.1

AUC, the area underneath the operating characteristic curve, is the marginal effect of the risk factor for discriminating trees that would die by the end of an observation period versus not and ranges from 50% (no better than flipping a coin) to 100% perfect prediction. P values for the AUC test the null hypothesis that the AUC = 50%.
[a] Only plots with a minimum mortality rate of 1% across observation periods were used to accurately estimate the AUC.
[b] One observation with KKL = 120.13 removed as outlier.
[c] Fourteen observations with reldbhdom > 1.15 removed as outliers.

Table 3. Estimates and significance results from the optimal prediction model for tree mortality.

	Log OR (SD)	OR (95% CI)	P value
Intercept	−17.01 (1.46)	0.00 (0.00–0.00)	<0.001
KKL			
$KKL^{1/3}$	2.74 (0.53)	15.48 (5.52–43.39)	<0.001
$KKL^{2/3}$	−0.38 (0.12)	0.68 (0.54–0.85)	<0.001
CIOvershade			
$CIOvershade^{1/2}$	1.30 (0.16)	3.65 (2.67–4.99)	<0.001
CIOvershade	−0.03 (0.005)	0.97 (0.96–0.98)	<0.001
CIIntra			
$CIIntra^{1/2}$	−0.21 (0.05)	0.81 (0.74–0.89)	<0.001
CIConifer			
$CIConifer^{1/3}$	1.80 (0.51)	6.05 (2.23–16.42)	<0.001
$CIConifer^{2/3}$	−0.38 (0.09)	0.68 (0.58–0.81)	<0.001
I(CIConifer = 0)	0.57 (0.77)	1.78 (0.39–8.02)	0.75
Random effects	SD	95% CI	
Plot	0.89	0.25–3.24	
Calendar year	2.18	0.20–23.14	

CI, confidence interval; $I(X)$, effect for X versus not X.

Individual Mortality Prediction Model

The optimal model chosen by the selection procedure (stage II) was fit to data from all 60 plots with resulting estimates shown in Table 3. Only the four competition indices, KKL, CIOvershade, CIIntra, and CIConifer, appeared in the final model. These indices are derived measures that use the geometric relationship of neighboring trees in addition to tree size. Together they outweighed the crude predictor dbh. Multiple entries of the same predictor, such as $KKL^{1/3}$ and $KKL^{2/3}$, reflect the optimal transformation (power ⅓) from stage I along with the optimal polynomial (degree 2) from stage II. These multiple entries make interpretation of the ORs complicated. Alternatively, interpretation of the effects of the four predictors on risk can be most easily visualized in Figure 1, which shows the combined effect of each predictor on risk after adjustment for the effects of the other predictors on risk. Risk of mortality increased with increasing KKL and increasing CIOvershade but decreased with increasing CIIntra and displayed nonmonotonic behavior with increasing CIConifer. Trends, however, must be interpreted with caution. The number of events (mortality) is low with the result that pointwise uncertainty bands are wide. The confidence bands additionally widen in areas of the predictor space (x-axis) that had few observations in the data set. Finally, variation due to calendar year of the observation period (random-effects SD = 2.18) was estimated twice as large as the variation due to plot (SD = 0.89) (Table 3).

Mortality Predictions for New Trees

A fit of the overall model to all observations in the dataset yields the following prediction for the probability of mortality during the next 5 years for a new tree

$$\begin{aligned}\text{Logit} = &-17.01 + 2.74\,KKL^{1/3} - 0.38\,KKL^{2/3} + 1.30\,CIOvershade^{1/2} \\ &- 0.03\,CIOvershade - 0.21\,CIIntra^{1/2} + 1.80\,CIConifer^{1/3} \\ &- 0.38\,CIConifer^{2/3} + 0.57\,I(CIConifer = 0),\end{aligned}$$

where I(CIConifer = 0) equals 1 if CIConifer has the value 0 and equals 0 otherwise. ROC curves for this model applied to the entire data set comprising 60 plots and to the individual 29 plots that had a minimum mortality of 1% are shown in Figure 2. The overall AUC of the prediction model was 91.5% and 6.8 percentage points higher than the AUC of the top individual risk factor, CIOvershade (AUC, 84.7%) (Table 2). However, this increase is optimistic because the same data were used to train the prediction model as to evaluate it. AUCs for the 29 plots ranged from 69.1 to 100%.

Discussion

Using comprehensive data from a series of long-term research plots, in this study, we have performed a detailed statistical analysis, resulting in a prediction model for individual tree mortality. Although the specific application was for the development of a prediction model for use in the SILVA simulator, the modeling concepts are general and could be applied to other species and prediction applications in forestry.

There have been many individual tree mortality models developed for different species of trees. Table 4 lists a small set of contemporary models that included European or American beech as one of the targeted species. All mortality models included dbh or some measure of basal area, and all except one performed logistic regression. The remaining model was based on the complementary

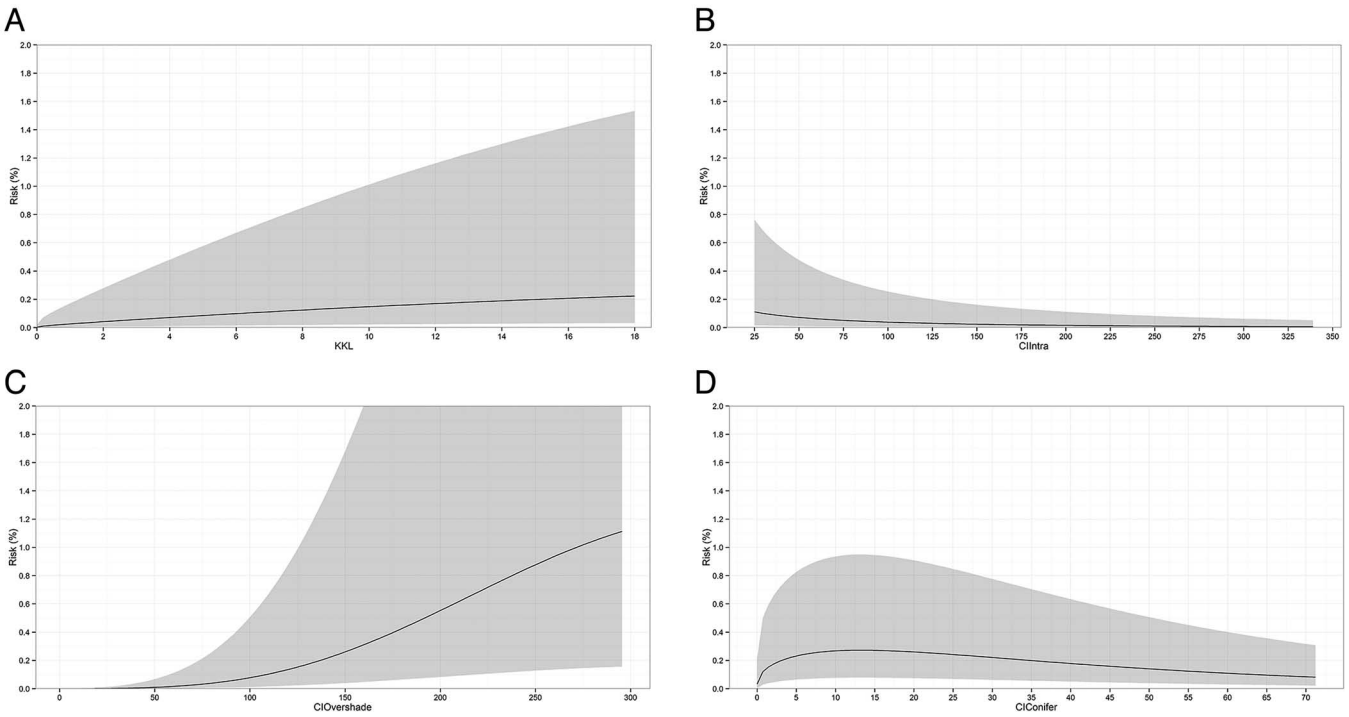

Figure 1. Risk of mortality in the next 5 years (solid lines) according to KKL (A), CIIntra (B), CIOvershade (C), and CIConifer (D) with pointwise 95% confidence intervals (shaded regions). Plot A corresponds to an individual tree with a CIIntra value of 116.1, CIOvershade value of 75.4, and CIConifer value of 0. Plot B corresponds to a KKL value of 1.42, CIOvershade value of 75.4, and CIConifer value of 0. Plot C corresponds to a KKL value of 1.42, CIIntra value of 116.1, and CIConifer value of 0. Plot D corresponds to a KKL value of 1.42, CIIntra value of 116.1, and CIOvershade value of 75.4.

log-log link, which is quite similar to the logistic link. Similar to the final model here, all models were based on a small handful of predictors proven to provide independent predictive information on mortality. Parsimonious models are better protected against overfitting and more likely to have better external validation performance. The mortality models in Table 4 are more generally applicable than those presented here because they are based on more easily calculated predictors, such as crown ratio or basal area of larger trees.

The initial mortality model in SILVA was presented by Pretzsch et al. (2002) and was based on a small subset of the same data as for this application (526 individual tree periods compared to 21,051 here). They similarly used logistic regression, but instead of using all observation periods, they selected an equal-sized series of observation periods from trees that had survived to observation periods where trees had died. This procedure mimics the efficient case control designs used in medicine for rare diseases. Their mortality model indicated an increased risk of mortality for trees with smaller dbh, with lower ratios of heights to dbh, with larger values of a site index (estimated stand top height at age 50 years), and for larger ratios of estimated tree basal area growth over the next 5 years to dbh. In contrast to the model developed here, the original mortality model required at least two observation periods per tree so that a growth model could be used to project 5-year basal area growth. Because our intent was to develop a single-tree single-observation period model, we did not perform a comparison to the original SILVA simulator.

Monserud and Sterba (1999) used logistic regression to develop individual tree mortality models for the six major forest species of Austria, one being European beech. They used a single 5-year remeasurement period of a permanent plot network of the Austrian National Forest Inventory. In addition, for use in an individual tree stand growth simulator, their aim was to provide a general mortality model to replace outdated yield tables that were still being used in Austria. Their inventory recorded an overall 5-year mortality rate for European beech of 4.3%, which is close to what was observed in this study (2.9%), and they similarly elucidated the obstacles for accurate modeling of rare events. To make their model generally applicable in Austria, where they argued that most stands failed to meet the definition of even-aged, they intentionally excluded site index and age of individual trees from consideration, arguing that tree size is already an integrated response to these factors. They hypothesized that a hyperbolic dbh^{-1} transformation would more accurately track the high mortality rates for small trees and gradually decreasing mortality rates for larger trees. Thus, they implemented a more subject-driven approach to transformation of risk factors than the automated spline approach used here. Their transformation was highly statistically significant, indicating the hypothesized nonlinear relationship of dbh to mortality. The ratio of crown length to height of the tree, as a measure of tree vigor, worked in tandem with a measure, basal area in larger trees (BAL), which counted the stand's basal area from trees with a larger diameter than the individual tree under consideration. Mortality increased as BAL in larger trees increased and as crown ratio decreased.

Using permanent plot data from a mountainous region in Switzerland, Wunder et al. (2007) focused on prediction models for European beech that distinguished between growth-dependent and growth-independent mortality. The growth-dependent models used as risk factors the relative basal area increment between two measurement periods divided by the basal area at the second measurement period. Location site and dbh were included as

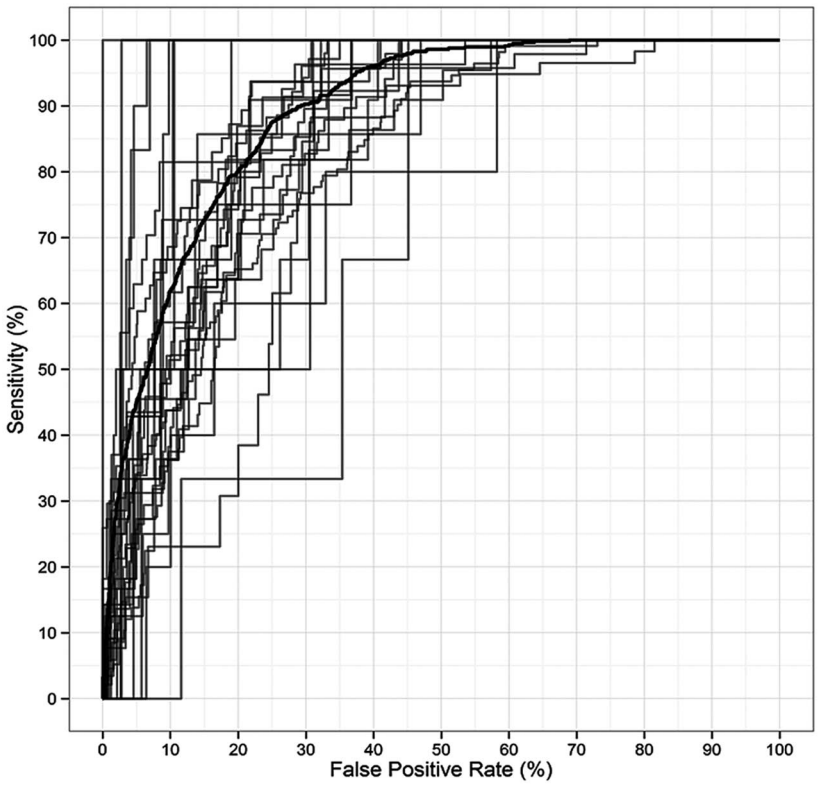

Figure 2. ROC curves of the optimal risk prediction model based on observations from all 60 plots (thick black line, AUC, 91.5%) and for each of the 29 plots with at least 1% mortality (thin lines, range between 69.3 and 100.0% for a plot with only one observed mortality).

Table 4. Previously published individual tree mortality models for European and American beech trees.

Reference	Tree species	Method	Outcome	Covariates
Monserud and Sterba (1999)	Norway spruce White fir European larch Scots pine European beech Oak	Logistic regression	5-year mortality	1/dbh CR BAL
Pretzsch et al. (2002)	Norway spruce Silver fir Scots pine Common beech Sessile oak	Logistic regression	5-year mortality	dbh Height Basal area growth[a] Site index[b]
Wunder et al. (2007)	Deciduous trees, Conifer	Logistic regression	Status at end of intervals Lengths were 5 yr and more	Log(dbh) Relative basal area of a tree (to dbh)
Fortin et al. (2008)	American beech Yellow birch Red maple Sugar maple Balsam fir	Binomial GLMM with complementary log-log link Plot random effects	5-year mortality	dbh (+ dbh^2) Treatment BA (BA^2) Tree vigor
Kiernan et al. (2009)	Sugar maple American beech White ash Bellow birch Striped maple Mixed conifers	Logistic regression GEE modeling for intratree correlation	Different period lengths, length (in years) used as factor variable	dbh/BAL No. of trees in plot Length of observation

BA, stand basal area; GLMM, generalized linear mixed model; BAL, basal area of larger trees; CR, crown ratio (crown length/tree height).
[a] Expected basal area growth over the next 5 years.
[b] Estimated stand top height at age 50 years.

growth-independent risk factors. Their data showed that trees that died experienced lower relative growths in the period before death than comparable time periods among trees that survived. A spline fit for the relationship of relative growth to survival revealed a nonlinear relationship. Among trees with smaller relative growth, the impact of growth on survival was stronger than that among trees with higher relative growth. At both sites in the study, trees with larger dbh had a higher chance of survival. Their prediction model obtained an AUC of 89.6% using bootstrapping on the same sample, a procedure similar to that for the AUC reported here. Their AUC was close to the AUC of the optimal prediction model obtained here (91.5%).

The above prediction models did not incorporate random effects to account for variable results among plots. In their prediction models for northern hardwood stands, which included American beech in Quebec, Canada, Fortin et al. (2008) stressed the importance of accounting for risk differences among plots that could not be explained by measured individual tree risk factors, such as soil and weather conditions. They also stressed the need to adjust for different intervals of measurement to account for changing climate conditions. Both their interval and plot random effects were significant, with SD estimates of 0.33 and 0.22, respectively. The analysis here also revealed a bigger impact of calendar year (SD, 2.18) than plot (SD, 0.89) contribution to unexplained variability. Magnitudes of the SDs of random effects depend on the amount of successful adjustment by the fixed effects in the model so that comparisons across studies and models are compromised. Nevertheless, the fact that both the Fortin et al. (2008) study and this study found larger variability due to time than to plot provides evidence that global changes due to climate may have a bigger impact than differences in plots due to soil and water conditions. In terms of fixed effects, they found that tree vigor, dbh, and basal area had an impact on survival, with the effects of dbh and basal area being nonlinear in nature. Their model entertained some common distance-independent competition indices, including the sum of basal area for all trees with dbh greater than that of the tree of interest, the relative position of the tree in the cumulative basal area distribution and the ratio between dbh and plot mean quadratic diameter. None of these had a significant impact on mortality. One possible reason for the lack of statistical significance of their competition indices is that they were distance-independent, in contrast to those used in this study and hence were not sensitive enough to detect competition. Another is the large plot size (0.5 ha). In addition, their model included many species, whereas this study focused only on European beech. Instead of logistic regression, they used the closely related complementary log link regression model that, like the model here, included a fixed offset term to account for variable lengths of observation periods.

In their modeling of tree mortality after selection in upstate New York for a multitude of species, including American beech, Kiernan et al. (2009) contrasted ordinary logistic regression with a generalized estimating equation (GEE) approach that accounted for dependencies between observation periods on the same tree. Both models showed that mortality increased with the ratio of basal area to dbh, with time of observation, and with number of trees in the plot and gave similar predictions. The GEE approach had slightly lower prediction error, in particular for smaller trees with dbh less than 15 cm. By accounting for the dependence between observation intervals rather than treating multiple observation periods from the same tree as independent, the SEs of parameters estimated through the GEE approach were larger, which the authors suggested yielded more accurate statistical significance results. We have argued that, because pooled logistic regression with rare events is asymptotically equivalent to grouped Cox regression, one need not additionally adjust for dependence between multiple observation periods on the same tree. The Cox regression likelihood accounts for this form of dependence. On the other hand, our model explicitly adjusted for spatial dependence via the plot random effect, whereas the model of Kiernan et al. (2009) did not.

The brief review of the literature combined with the results of this study show that a variety of statistical methods have effectively been used for modeling the rare event of forest mortality. Mortality models are designed with specific objectives in mind; these objectives determine the risk factors used in the model. In contrast with the other models, mortality models in this study were specifically designed to capitalize on the many geometrical and distance-based competition indices that are calculated with detailed forest inventory data through the SILVA simulator. Thus, these distance-dependent competition indices outweighed the effect of the crude predictor dbh or other predictors of tree size. The mortality model in this report was limited to European beech, one of the largest of two species currently under observation as part of the Bavarian forest network.

The median-based approach to prediction used in this application (by setting random effects equal to their prior median of 0) gave 5-year mortality forecasts similar to those of the more commonly used mean predictions in terms of multiple criteria, including discrimination of mortality from nonmortality observation periods (AUC) and squared error (Brier score). Median predictions provide a nice alternative to mean predictions because they require no additional computation beyond fitting the model. However, median predictions may not outperform mean predictions for longer forecasts, such as prediction of mortality over the next 25 years. The equations developed in this report only apply to short-term mortality predictions.

Exploratory data analyses, including B-splines, along with graphical techniques were used to arrive at the specific transformations to optimize the logistic regression fit. It is worth remarking that B-splines, which are complicated in form, were not an essential ingredient but rather a choice for the model building. They were used here as a nonparametric smoothing device to suggest the optimal transformation (such as a quadratic transformation) of risk factors in the logistic regression equation. Simpler techniques, such as the Box-Cox transformation could have been implemented to produce the same effect (Box and Cox 1964). The high specificity of a data-driven approach to transformations incurs the risk that the same model may not apply to other species or to the same species in forests in other geographical areas. Of 10 potential predictors measuring size and competition, our final model included only four competition indices, raising a concern over multicollinearity. As described in the Materials and Methods section, these indices had correlations of less than 0.75 or they would not have been simultaneously allowed to enter the final model. The risk curves in Figure 1 demonstrate that the model relationships with the four competition indices are as predicted. Nevertheless, there is a concern that any amount of multicollinearity may affect external validation. Therefore, we are currently repeating the modeling approach in Douglas fir, another common species among the research plots used in this analysis. We hope that the analytical formulas for the proposed model will facilitate external validation among beech trees in other forests that similarly record individual tree positions.

Literature Cited

ABBOTT, R.D. 1985. Logistic regression in survival analysis. *Am. J. Epidemiol.* 121:465–471.

ASSMANN, E. 1961. *Waldertragskunde. Organische Produktion, Struktur, Zuwachs und Ertrag von Waldbeständen.* BLV Verlagsgesellschaft, München, Germany. 490 p.

BIGING, G.S., AND M. DOBBERTIN. 1992. A comparison of distance-dependent competition measures for height and basal area growth of individual conifer trees. *For. Sci.* 38:695–720.

BOX, G.E.P., AND D.R. COX. 1964. An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26:211–252.

COX, D.R. 1972. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* 34:187–220.
D'AGOSTINO, R.B., M.L. LEE, A.J. BELANGER, L.A. CUPPLES, K. ANDERSON, AND W.B. KANNEL. 1990. Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat. Med.* 9:1501–1515.
EILERS, P.H., AND B.D. MARX. 1996. Flexible smoothing with B-splines and penalties. *Stat. Sci.* 11:89–121.
FARAGGI, D., AND B. REISER. 2002. Estimation of the area under the ROC curve. *Stat. Med.* 21:3093–3106.
FORTIN, M., S. BÉDARD, J. DEBLOIS, AND S. MEUNIER. 2008. Predicting individual tree mortality in northern hardwood stands under uneven-aged management in southern Québec, Canada. *Ann. For. Sci.* 65:205p1–205p12.
KIERNAN, D., E. BEVILACQUA, R. NYLAND, AND L. ZHANG. 2009. Modeling tree mortality in low- to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations. *For. Sci.* 55:343–351.
MONSERUD, R.A., AND H. STERBA. 1999. Modeling individual tree mortality for Austrian forest species. *For. Ecol. Manage.* 113:109–123.
NAGELKERKE, N.J.D. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78:691–692.
PRETZSCH, H. 1992. Modellierung der Kronenkonkurrenz von Fichte und Buche in Rein- und Mischbeständen. *Allg. Forst. Jagdzeit.* 163(11/12):203–213.
PRETZSCH, H. 2001. *Modellierung des Waldwachstums.* Blackwell, Berlin. 664 p.
PRETZSCH, H., P. BIBER, AND J. DURSKY. 2002. The single tree-based stand simulator SILVA: Construction, application and evaluation. *For. Ecol. Manage.* 162:3–21.
SCHOBER, R. 1967. Buchen-Ertragstafel für mäßige und starke Durchforstung. In *Die Rotbuche 1971.* Sauerländer's Verlag, Frankfurt am Main, 1972; Schriften aus der Forstlichen Fakultät der Universität Göttingen und der Niedersächsischen Forstlichen Versuchsanstalt 43/44. 333 p.
SING, T., O. SANDER, N. BEERENWINKEL, AND T. LENGAUER. 2009. *ROCR: Visualizing the performance of scoring classifiers*. R package version 1.0-4. R Foundation for Statistical Computing, Vienna, Austria.
SKRONDAL, A., AND S. RABE-HESKETH. 2009. Prediction in multilevel generalized linear models. *J. R. Stat. Soc. Ser. A.* 172:659–687.
STEYERBERG, E.W. 2010. *Clinical prediction models.* Springer, New York. 497 p.
WICKHAM, H. 2009. *ggplot2: Elegant graphics for data analysis.* Use R! series. Springer, New York. 213 p.
WOOD, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B* 73:3–36.
WUNDER, J., B. REINEKING, J.F. MATTER, C. BIGLER, AND H. BUGMANN. 2007. Predicting tree death for *Fagus sylvatica* and *Abies alba* using permanent plot data. *J. Veget. Sci.* 18:525–534.

Appendix

Competition Indices Derived from Vertical Competition Profiles

Many different types of competition indices have been proposed previously, for example, those in Biging and Dobbertin (1992). The SILVA simulator uses an additional set of indices that are rooted in well-proven concepts and are based on vertical competition profiles. The basis of the competition indices CICUM60, CIIntra, CIConifer, CIOvershade, and CILateral is a procedure that is visualized in Figure A1. The space around a tree of interest (shaded in gray) is stacked with horizontal planes spaced at distances 1/20 of the tree of interest's height. An upturned cone with an opening angle of 60° is placed with its tip in the footprint of the tree of interest. The intersection areas of the cone and the horizontal planes form a series of circles that become larger with increasing distance from the forest floor. Any neighbor tree that touches that cone is considered a competitor. Thus, the left tree in the Figure A1 is not a competitor, whereas the right tree is.

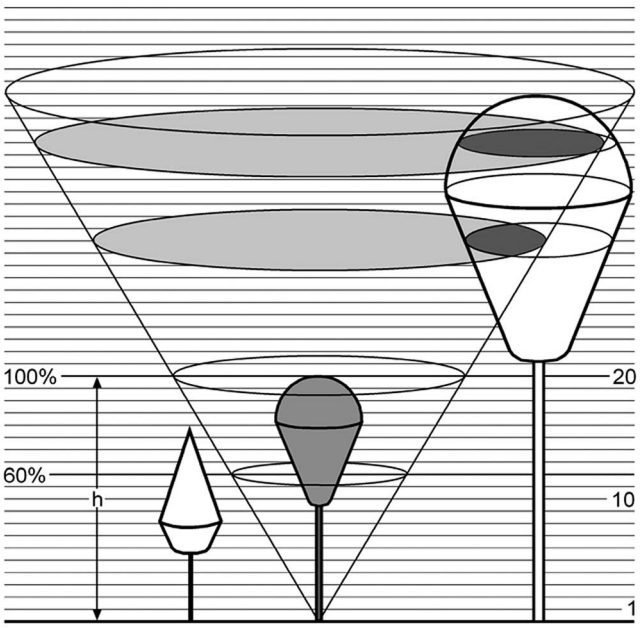

Figure A1. The principle for determining vertical competition profiles.

The three-dimensional crown models of Pretzsch (2001) are applied to measure the overlapped area (shown in dark gray in Figure A1) of each competitor's crown with the respective cone-intersection circle (shown in light gray). The relative proportions of the overlapped areas to the cone-intersection circles are summed up plane-wise, and then the profiles are stepwise integrated from their topmost point down to the forest floor. The resulting integrals are multiplied by 1/20 (one step width relative to the tree of interest's height). The integral value obtained at 60% of the tree of interest's height is the competition index CICUM60, a general measure of competition. CIIntra is the component of CICUM60 that comes from trees that belong to the same species as the tree of interest, whereas CIConifer is the component resulting from coniferous competitors, such as Norway spruce (*Picea abies* [L.] H. Karst) and Scots pine (*Pinus sylvestris* L.).

To divide competition into the ecologically different aspects of overshading and lateral constriction (Assmann 1961, Pretzsch 1992), the integral value at the tree of interest's top is assigned to the measure CIOvershade, because it comes from tree crowns above the tree's top, which cause overshading. The difference CILateral = CICUM60 − CI Overshade is used as a measure for lateral competition, because large values mean that competition increases downwards from the top along the tree of interest's crown, expressing competition that does not come from overshading.

The final competition index KKL is described in detail in Pretzsch et al. (2002). A virtual cone is placed within a given tree with axis equal to the tree axis and vertex in the crown of the tree. Any tree whose top is inside this virtual cone is regarded as a competitor. For any competitor tree, the angle β between the insertion point of the cone and the top of the competitor tree is determined (see Figure 4

of Pretzsch et al. 2002). This angle is weighted by the relation between the crown cross-sectional areas of the competitor and tree of interest. These areas are calculated according to crown models and multiplied by species-specific light transmission coefficients from Pretzsch (1992). The competition index is defined as the sum of all competitor contributions

$$\mathrm{KKL}_i = \sum_{j=1}^{n} \beta_j \frac{\mathrm{CCA}_j}{\mathrm{CCA}_i} \mathrm{TM}_i,$$

where KKL_i is the competition index for tree i, β_j is the angle between the cone vertex and top of competitor j, CCA_j and CCA_i are the crown cross-sectional areas of trees j and i, respectively, TM_j is the species-specific light transmission coefficient for tree j, and n is the number of competitors of tree i.

Calculation of Confidence Intervals for Mortality Predictions

Let $\pi(x'\beta) = \exp(x'\beta)/\{1 + \exp(x'\beta)\}$ be the estimated probability of mortality based on the logistic regression model with fixed-effects covariate vector x and vector of log ORs for fixed effects β. Let $V(\beta)$ be the estimated variance-covariance matrix of β that is output of standard logistic regression software. Then a 95% confidence interval for $\pi(\beta)$ is given by

$$\left(\pi(x'\beta - 1.96\sqrt{x'\mathrm{V}(\beta)x}), \qquad \pi(x'\beta + 1.96\sqrt{x'\mathrm{V}(\beta)x}) \right).$$