Comparing Methods for Modelling Tree Mortality

James W. Flewelling
Robert A. Monserud

Abstract—Our goal is to compare different methods for estimating the parameters of individual tree mortality models. We examine general methods including maximum likelihood and weighted nonlinear regression, and a few specialized methods including Proc LOGISTIC in *SAS*, and an implementation of the Walker-Duncan algorithm. For fixed period lengths, almost all methods for fitting the logistic mortality model should work. The LOGISTIC procedure in *SAS* is quite robust and the easiest to use. For unequal period lengths, either a weighted nonlinear least squares or a maximum likelihood formulation is needed to specify the annualized logistic mortality model: NLIN(wtLS) and NLIN(LOSS) in *SAS*. Of course, other statistical packages that mimic these procedures should give the same results.

Introduction

Mortality remains one of the least understood yet important components of growth and yield estimation (Hamilton 1986). Great success in modeling mortality is rare, perhaps because the focus is on modeling the occurrence of rare events. Realistically, mortality modelers mostly hope to capture the average rate of mortality and relate it to a few reliable and measurable size or site characteristics (Keister 1972; Hamilton and Edwards 1976; Monserud 1976; Hamilton 1994; Monserud and Sterba 1999).

A few preliminary observations provide some context, several of which conspire to increase the difficulty of mortality modeling in forestry. First, mortality is a discrete, rare event. A common rule of thumb for both temperate and boreal forests is that roughly 0.5 to 1.5 percent of the trees are expected to die in a given year (background mortality). It therefore follows that a large sample is needed to observe enough occurrences of mortality for modeling the process adequately. Using this rule of thumb, we expect that a sample of 10,000 trees/yr (or 2000 trees/5 yrs) would be needed to observe approximately 50 to 150 deaths, which is a relatively small sample of the event of interest. Second, observations of the same individual tree at two points in time are necessary to observe the survival or mortality status of a tree; this requires remeasured permanent plots. Estimates of the date of death from temporary plots are notoriously unreliable and should not be used. Third, we usually do not know tree age, so we cannot use Survival Analysis methods (Allison 1995) that have been successful in medical research. Furthermore, remeasured tree survival data are usually both left and right censored (Allison 1995; Meeker 1998): we only know that some of the trees died sometime during the sampling interval, and the others failed to die. Fourth, it is not uncommon to have widely varying period lengths, which precludes the proper use of some estimation programs that assume equal period lengths, such as Proc LOGISTIC in *SAS* (SAS Version 8.1 is a product of the SAS Institute, Cary, NC 27513).

Simulator architecture determines how mortality must be calculated and simulated. Mortality is *discrete* in a spatial simulator such as *TASS* (Mitchell 1975) or *FOREST* (Ek and Monserud 1974), with the tree either completely dead or alive. This mortality process is stochastic. Furthermore, the costs of misclassification are not equal in spatial models. Misclassifying a live tree as dead can never be corrected, but misclassifying a dead tree as live can be corrected in the future. Mortality can be *continuous* in a nonspatial simulator such as *FVS* (Stage 1973; Wykoff and others 1982; Hamilton 1994; Teck and others 1997) or *ORGANON* (Hann and others 1997), with the mortality rate smoothly reducing the number of trees each sample tree represents. Therefore, it is much easier to predict mortality rates over large areas with nonspatial models.

Our goal here is to compare several methods for estimating the parameters of individual tree mortality models. Note that our objective is not to find the best mortality model for a given data set. We will examine several methods: maximum likelihood, weighted nonlinear regression, and a few specialized programs (Proc LOGISTIC in *SAS*, and the Walker-Duncan algorithm of Hamilton 1974). Our motivation is that all too often authors do not provide enough details on methods to reproduce results. Although our immediate context is modeling in the *FVS* environment, our results apply to any other forest simulation system.

We have a few caveats. We will not address:

- Correlation of trees within a plot. They will have the same stand density, and they could all be exposed to the same unobserved mortality agent.
- Plot size effects on model parameters (see Stage and Wykoff 1998).
- Simultaneous fitting of mortality and the growth equations (Hasenauer and others 1998; Cao 2000)

In: Crookston, Nicholas L.; Havis, Robert N., comps. 2002. Second Forest Vegetation Simulator Conference; 2002 February 12–14; Fort Collins, CO. Proc. RMRS-P-25. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

James W. Flewelling is a Consulting Biometrician, 26724 51st Place S., Kent, WA 98032. Fax: +1-253-520-6733. E-mail: jimflew@eskimo.com. Robert A. Monserud is Chief Biometrician, Pacific Northwest and Rocky Mountain Research Stations, located at 620 SW Main St., Suite 400, Portland, OR. Fax: +1-503-808-2020. E-mail: rmonserud@fs.fed.us

The Algebra of Mortality

Because a tree can only die once, mortality is not a Markov process. Survival is, however. This property requires that all algebra be mediated in terms of survival, not mortality. The probabilities of survival (P_s) and mortality (P_m) are connected by the standard identity:

$$\text{Survival} = 1 - \text{Mortality}: P_s = 1 - P_m$$

We must use the compound interest formula to convert survival to other period lengths:

1-year: $P_{s_1} = 1 - P_{m_1}$

n-year: $P_{s_n} = P_{s_1} P_{s_1} P_{s_1} \ldots P_{s_1} = (P_{s_1})^n = (1 - P_{m_1})^n$

(*The Markov property*)

1-year: $P_{m_1} = 1 - P_{s_n}{}^{1/n} = 1 - (1 - P_{m_n})^{1/n}$

Statistical Considerations

The modeling of tree mortality requires fundamentally different statistical methods than those used for most of the other components of forest growth models. This is because the dependent variable is dichotomous rather than continuous: 0 or 1 (live or dead). It then follows that the variance is Binomial and is not constant: Var = $P(1-P) = P_m P_s$. Maximum Likelihood (ML) methods are the proper estimation procedure. The likelihood of a sample of observations is the joint probability of that sample (Mendenhall and Shaeffer 1973). The ML estimation problem is to find the values of the unknown parameters that maximize this likelihood, which is a product of individual probabilities. The log of this product is used instead because it is a monotonically increasing function of the Likelihood and is much easier to differentiate and maximize (Mendenhall and Shaeffer 1973). Although seemingly complicated, ML estimates are quite intuitive. They have the desirable properties of consistency, asymptotic unbiasedness, and efficiency. The independence of the observations is an underlying assumption, which is violated if there are correlations between trees on the same permanent plot, resulting in the underestimation of confidence intervals for parameters. We do not examine the independence assumption.

A seminal paper by Grizzle and others (1969) demonstrated that Weighted Least Squares (with the weight $\propto$1/var) could be an equally appropriate estimation method for categorical data analysis (Agresti 1990). The method is widely available and is asymptotically equivalent to maximum likelihood (Agresti 1990). Fisher's Discriminant Analysis also works but quickly becomes unwieldy as the problem becomes increasingly multivariate.

To model the probability of mortality (or survival), all that is needed is a flexible function bounded by 0 and 1. The Logistic equation is most often used, although any sigmoid or cumulative distribution function would work. The logistic for the probability of survival (P_s) is:

$$P_s = \frac{1}{1+e^{b'X}} = \frac{1}{1+e^{b_0+b_1x_1+b_2x_2+\ldots}}$$

where X is a vector of independent variables, and $b'X$ is a linear combination of the n parameters b_i and the x_i independent variables. Invoking the identity $P_s = 1 - P_m$ and some algebraic manipulation leaves us with the closely related logistic form for the probability of mortality (P_m):

$$P_m = \frac{e^{b'X}}{1+e^{b'X}}$$

The inverse transform has a long history (McCullagh and Nelder 1983):

$$\log(P_m/(1-P_m)) = b'X$$

It is recognizable as the "logit" or the "log odds ratio." This equation provides the link between the unconstrained function $b'X$ and the logistic probability prediction, which is bounded by [0,1].

In the foregoing, P_m and P_s are interchangeable. The function to model P_m could be used equally well to model P_s; the only difference is that the signs in the coefficient vector b would be reversed. This amounts to switching the 0-1 coding on the dependent variable.

We always want to know how good the model is. Usual measures of residual variation (R^2) are useless for dichotomous variables. It does not matter how close a predictor is to 0 or 1 as long as it can accurately estimate expectations. We will simply use the sum of the log likelihoods, with the standard factor of –2 that results from maximizing the logarithmic linearization; we will label this –2LL. A closely related statistic for model comparison is Akaike's Information Criterion (AIC). It penalizes –2LL by the number of parameters in the model (Chatfield 1996). The Chi-square statistic is also appropriate, with the calculation based on dividing important predictor variables into classes. Hosmer and Lemeshow (2000) discuss these and other statistics.

Methods

We concentrate on a comparison of several methods for estimating parameters with a fixed logistic model. Five fitting techniques are described:

1. Logit Model for Proportions { REG } — This model, also referred to as the linear logistic model, is ascribed to several authors in the 1940s and 1950s (McCullagh and Nelder 1983). It requires that the data be grouped into classes of X values. Within each class, the predictions are identical for all observations. The observed proportion in each class should be between zero and one, and the expected numbers of events and nonevents in each class should be at least five. The procedure is to transform the independent variable by computing its logit and then regressing that transformation against X.

2. Least squares { NLIN (LS) } — Unweighted least squares minimizes the sums of squared errors, where the error is calculated as $\varepsilon = Y - P_m$, where Y equals one for mortality and zero otherwise, and P_m is the predicted probability of mortality. Because of the logistic link function, P_m is not a linear function of the parameters. Therefore, nonlinear regression is needed.

3. Walker-Duncan {RISK } — The Walker-Duncan algorithm (Walker and Duncan 1967) is a sequential approach to weighted least squares that was developed in an era of limited computer resources. It is claimed to be asymptotically equivalent to the maximum likelihood fit of the logistic

model. This routine was coded by Hamilton (1974) in the RISK program and extended to deal with variable period-length data.

4. Weighted Least Squares {NLIN (wtLS), LOGISTIC } — Weighted least squares minimizes the weighted sums of error squared, using weights that are inversely proportional to the estimated variance: Var = *P(1-P)*. McCullaugh and Nelder (1983) show that the minimization of this weighted sums of squares produces the maximum likelihood solution. That proof assumes that the link function is linear in the parameters, which is not true for the variable length period regressions to be discussed later. Nonetheless, in every test regression that we have examined, the weighted least squares method does produce the ML solution, a result supported by Agresti (1990). The weighting is iterative because the weights depend on the parameter vector *b*, which is being estimated. The iterative weighting can prevent some fitting techniques from fully converging. A conservative strategy is to hold the weights constant while iterating toward a least squares solution, then resetting the weights and proceeding to further rounds of minimization. Most LOGISTIC programs use iteratively reweighed least squares to obtain the ML solution.

5. Maximum Likelihood { NLIN (LOSS) } — Maximum likelihood solutions can be derived directly by iterating the model parameters until the likelihood is maximized, or equivalently until the negative log likelihood is minimized. This is done by writing the LL function shown in the appendix and incorporating that computation into an optimizer. The *SAS* procedure NLIN and equivalent procedures in other statistical packages can perform the optimization. In NLIN, the keyword _LOSS_ defines the objective criteria.

The above programs are all procedures within *SAS*, except for RISK, which was developed by Hamilton (1974). We use the notation "LS" to indicate the least squares objective function normally employed in nonlinear regression. The notation "LOSS" indicates where we employ an explicitly defined loss function.

Data Sets

Four data sets are used to illustrate various regression methods. The data sets are referred to as data A, B, C, and D; they are summarized or listed in their entirety in tables 1 through 4, respectively. Data sets A and B offer simple examples suitable for event estimation (tables 1 and 2); period length is not involved. Data set C is real data (64,121 observations) from young unthinned plantations of Douglas-fir in coastal Washington, Oregon, and British Columbia (summary in table 3). The growth rates in those data are from regressions of change in DBH versus initial DBH fit to the data one plot and one growth period at a time. Growth periods are from 1 to 4 years. Data set D is a simulated data set, which also has variable period lengths (table 4). It consists of four 0.5 acre plots, each with 10 arbitrary selected tree sizes, arbitrary frequencies, and arbitrary growth rates. The mortality was simulated for all trees in all years using the ML regression fit to data set C. Because data set D was selected to have high densities and long periods, the predicted mortality rates are high, allowing us to better see the differences due to the various fitting methods. As with data set C, the growth rates are best thought of as having been predicted by a diameter growth function.

Table 1—Data set A - categorized data distributed as an example for the RISK program.

Y	X_1	X_2
0	1	2
1	2	1
1	1	2
1	2	2
0	1	2
0	1	1
0	2	1
0	2	1
1	2	2
1	2	2
1	2	2
1	2	1
1	1	2
0	2	1
1	1	1
1	2	2
1	2	2
1	1	2
1	2	1
1	1	2
1	1	2
1	2	2
1	1	1
0	2	2
0	2	2
1	2	1
1	1	1
0	1	1
1	1	2
0	1	1
1	1	2
0	2	1
1	1	1
1	1	2
0	2	2
0	1	1
0	2	1
1	1	1
1	2	1
1	1	1

An alternative formulation of the same data set shows N observation on each line, with NY1 of them indicating an event (Y=1):

X_1	X_2	N	NY1
1	1	10	6
1	2	10	8
2	1	10	5
2	2	10	7

Table 2—Data set B. A simple data set illustrating low-probability events. The data set is shown in its classed formulation. It could also be shown as 100 observations, with 11 observations having Y = 1 and 89 observations having Y = 0. All models using these data use the expanded form (100 observations) unless stated otherwise.

X_1	N	NY1
0.0	10	3
0.1	10	1
0.2	10	2
0.3	10	1
0.4	10	1
0.5	10	2
0.6	10	0
0.7	10	0
0.9	10	1
1.0	10	0

Table 3—Summary of data set C, remeasurement data from fixed-area plots in young Douglas-fir plantations. The data set contains 64,121 observations.

Variable	Mean	Minimum	Maximum
DBH (in)	4	0	14
Height (ft)	22	2	74
BA (ft^2/ac)	62	3	172
Age	14	5	22
Period length (years)	3.4	1	4
Mortality proportion	0.014		

Results

Fitting with Fixed Period Length

The examples presented use techniques that are suitable for predicting mortality in a period of fixed length, where all of the data are from remeasurement periods of that same length. In these circumstances, the length of the period does not enter into the prediction process.

Data set A is used to illustrate the fitting of a simple logistic regression. Code for three methods is shown in the appendix. Coefficients and -2 LL's are in table 5; predictions for all X values are in table 6. RISK, REG, and LOGISTIC solutions all have nearly identical likelihoods, even though there are minor differences in the coefficients. RISK predicts there will be 25.90 events, whereas the other methods have prediction closer to the observed 26 events; this offers a hint that the RISK solution may not be the true maximum likelihood solution. The results from RISK may vary a bit if the order of observations is changed. The conclusion is that for classed data, with event (or nonevent) expectations as low as 2, all of the fitting methods produce results that are statistically indistinguishable from one another.

Data set B (table 2) is another simple data set. It is included here to show some lower probability events, as are typical in mortality data sets. Code is presented for four regression procedures in the appendix. Unweighted least squares is implemented with NLIN (LS); weighted least squares is implemented by NLIN (wtLS) and LOGISTIC; direct maximum likelihood is implemented with NLIN (LOSS). In the NLIN(wtLS) code, note how the iterative reweighting is implemented in two passes: first, an unweighted regression returns preliminary estimates of event probabilities; second, a weighted regression then uses the earlier estimated probabilities in assigning weights. The solutions are nearly identical for all methods (table 7). The trivial differences in -2LL are due to incomplete convergence for NLIN(LOSS) and to the fundamental consideration that NLIN(LS) is not using the theoretically correct weighting.

Fitting with Variable Period Length

One complication in mortality modeling is unequal period lengths. In the Algebra subsection, we saw how the compound interest formula (Markov property) can bring mortality or survival rates for any period length to an annual basis. It is important to have this flexibility to simulate any reasonable period length. However, not all estimation procedures allow for variable period lengths. The LOGISTIC procedure in *SAS* is a case in point. Beyond the mechanical difficulties of fitting, there may be serious differences between how the regression predictions are applied to the data set, and how they will be made within a simulator.

Data set C has variable period lengths of 1 to 4 years (table 3). An annual mortality model is fit for these data. The independent variables generally take on values as assigned at the start of the growth period. One variant on the regression process uses midpoint values for the independent variables, a technique that is useful for building annual simulators from multiyear data (Hyink and other 1985). The linear combination used in the Logit is:

$$b'X = b_0 + b_1 DBH + b_2 BAL/BA + b_3 BA$$

where BA is the total basal area per acre, and BAL is the basal area of larger trees. Although these variables were a good set of predictors for this particular data set, they are not a recommended set for other applications.

The procedures applied to data set C include RISK, LOGISTIC, and NLIN(LOSS). The latter procedure is run once with start-of-period X values and once with midpoint values. Results are in table 8. The results from NLIN(LOSS) are the best (the lowest -2LL), and RISK is the poorest. Surprisingly, the LOGISTIC results are almost the best, even though they rely on a statistically incorrect model. In this application, each period of YIP years is treated as YIP independent observations, one of which may result in mortality. For situations where mortality rates are very low (as in data set C), the LOGISTIC procedure can produce good estimates of parameters. However, all the fit statistics (including -2LL) and confidence intervals returned by the LOGISTIC regression program are incorrect. The NLIN (LOSS) regression using midpoint values is significantly poorer than those using start-of-period values. The use of midpoint values was expected to be an improvement; the lack of improvement might indicate a deficiency in our choice of independent

Table 4—Data set D. Simulated data set based on models from data set C.

PLOT	SIZE	YIP	DBH	DBHRATE	NTREES	GROUP	BAL	BA	TPA	NSURV	BA_FINAL	TPAFINAL
1	0.5	8	2.8	0.80	40	1	1.710	17.191	800	37	220.410	688
1	0.5	8	2.6	0.78	40	2	4.896	17.191	800	35	220.410	688
1	0.5	8	2.4	0.76	40	3	7.627	17.191	800	35	220.410	688
1	0.5	8	2.2	0.74	40	4	9.940	17.191	800	36	220.410	688
1	0.5	8	2.0	0.72	40	5	11.868	17.191	800	36	220.410	688
1	0.5	8	1.8	0.70	40	6	13.448	17.191	800	33	220.410	688
1	0.5	8	1.6	0.68	40	7	14.713	17.191	800	37	220.410	688
1	0.5	8	1.4	0.64	40	8	15.699	17.191	800	35	220.410	688
1	0.5	8	1.2	0.62	40	9	16.441	17.191	800	31	220.410	688
1	0.5	8	1.0	0.60	40	10	16.973	17.191	800	29	220.410	688
2	0.5	12	2.8	0.80	40	1	1.710	17.191	800	34	296.201	456
2	0.5	12	2.6	0.78	40	2	4.896	17.191	800	29	296.201	456
2	0.5	12	2.4	0.76	40	3	7.627	17.191	800	27	296.201	456
2	0.5	12	2.2	0.74	40	4	9.940	17.191	800	31	296.201	456
2	0.5	12	2.0	0.72	40	5	11.868	17.191	800	30	296.201	456
2	0.5	12	1.8	0.70	40	6	13.448	17.191	800	26	296.201	456
2	0.5	12	1.6	0.68	40	7	14.713	17.191	800	22	296.201	456
2	0.5	12	1.4	0.64	40	8	15.699	17.191	800	12	296.201	456
2	0.5	12	1.2	0.62	40	9	16.441	17.191	800	8	296.201	456
2	0.5	12	1.0	0.60	40	10	16.973	17.191	800	9	296.201	456
3	0.5	8	7.8	0.50	30	1	9.955	156.883	600	24	283.609	434
3	0.5	8	7.6	0.50	30	2	29.361	156.883	600	24	283.609	434
3	0.5	8	7.4	0.50	30	3	47.772	156.883	600	20	283.609	434
3	0.5	8	7.2	0.50	30	4	65.214	156.883	600	22	283.609	434
3	0.5	8	7.0	0.50	30	5	81.714	156.883	600	23	283.609	434
3	0.5	8	6.8	0.50	30	6	97.298	156.883	600	22	283.609	434
3	0.5	8	6.6	0.50	30	7	111.991	156.883	600	24	283.609	434
3	0.5	8	6.4	0.50	30	8	125.821	156.883	600	17	283.609	434
3	0.5	8	6.2	0.50	30	9	138.813	156.883	600	21	283.609	434
3	0.5	8	6.0	0.50	30	10	150.993	156.883	600	20	283.609	434
4	0.5	12	7.8	0.50	30	1	9.955	156.883	600	20	326.410	356
4	0.5	12	7.6	0.50	30	2	29.361	156.883	600	23	326.410	356
4	0.5	12	7.4	0.50	30	3	47.772	156.883	600	15	326.410	356
4	0.5	12	7.2	0.50	30	4	65.214	156.883	600	20	326.410	356
4	0.5	12	7.0	0.50	30	5	81.714	156.883	600	18	326.410	356
4	0.5	12	6.8	0.50	30	6	97.298	156.883	600	18	326.410	356
4	0.5	12	6.6	0.50	30	7	111.991	156.883	600	17	326.410	356
4	0.5	12	6.4	0.50	30	8	125.821	156.883	600	16	326.410	356
4	0.5	12	6.2	0.50	30	9	138.813	156.883	600	14	326.410	356
4	0.5	12	6.0	0.50	30	10	150.993	156.883	600	17	326.410	356

Variables for data set D

Plot	An arbitrarily assigned plot number
Size	Plot size in acres
YIP	Years in the period
DBH	Initial diameter (inches)
DBHRATE	DBH growth rate (in/yr)
NTREES	Initial number of trees of this size.
GROUP	Sequence number assigned to a group of trees of same size.
BAL	Initial basal area of larger trees (including half of current group)
BA	Initial basal area (sq. ft/ac)
TPA	Initial trees per acre
NSURV	Number of surviving trees at end of growth period
BA_FINAL	Final basal area for the plot (sq. ft/ac)
TPAFINAL	Final trees per acre.

Table 5—Regression results, data set A.

Regression	b_0	b_1	b_2	–2LL
LOGISTIC	–.011508	–.462517	.908641	49.558
REG	–.002985	–.47223	.91406	49.558
RISK	.0530757	–.494347	.889641	49.562

Table 6—Predictions from three methods for the probability of an event using data set A.

X_1	X_2	N	NY1/N	P (RISK)	P (LOGISTIC)	P (REG)
1	1	10	.6	.61025	.60698	.60798
1	2	10	.8	.79216	.79302	.79461
2	1	10	.5	.48851	.49302	.49166
2	2	10	.7	.69924	.70698	.70696
Wt'd Sum		40	26	25.90	26.00	26.01

Table7—Regression results, data set B.

Regression	b_0	b_1	–2LL
LOGISTIC	–1.1115	–2.6602	64.3675
NLIN(LS)	–1.1219	–2.6270	64.3682
NLIN(wtLS)	–1.1120	–2.6586	64.3675
NLIN (LOSS)	–1.1219	–2.6270	64.3682

Table 8—Regression results, data set C.

Regression	b_0	b_1	b_2	b_3	–2LL
LOGISTIC[1]	–3.2956	–.9826	–1.5666	0.0410	8049
NLIN(LOSS)	–3.4217	–1.0068	–1.38	0.0425	8047
NLIN(LOSS) – Midpoint	–2.3943	–1.1032	–2.3609	0.0423	8068
RISK	–3.1513	–.90577	–1.6717	0.0391	8096

[1]The regression reports –2LL as 10363, based on an inflated number of observations.

variables. The RISK procedure shows the poorest -2LL value, possibly indicative of convergence problems.

Data set D is an artificial data set of four plots, each with 10 tree sizes. Four variations of the same model are examined: a Compound survival model, a Midpoint model, an Interpolated model, and a Simulation model. The differences between the models arise from different assumptions for calculating the same X variables. The Compound model uses the initial values of X's for each survival prediction, so a compound interest formula for survival is appropriate. The Midpoint model uses the average of starting and ending values for the independent variables and applies the same compound survival formula as for the previous method. The Interpolated model uses different X values for each year in the growth period; these are from linear interpolation. For both the midpoint and interpolated models, BAL/BA is not updated from the start of the period because doing so would require information on tree mortality. The Simulation model uses different X values for each year; these are iteratively recomputed based on predicted survival (for example, the computation of each year's BAL is dependent on the predicted mortality in earlier years). The NLIN (LOSS) procedure is used for fitting in all four cases. The Midpoint model, the Interpolated Model, and the Simulation model all take advantage of knowing something about DBH growth rates or end-of-period values. If the DBH growth rates are in fact predictions from a growth model, their use does not invalidate the independence of the X values. The use of the end-of-period basal area by the Midpoint and Interpolated models is a not a statistically valid procedure, but may still produce good results (Hyink and others 1985).

The Compound survival model for data set D uses the same code as given in the appendix under the heading: "C - NLIN(LOSS)". This is a reasonable approach for data where the period lengths are similar. With period lengths in the data of 8 and 12 years, this model could reasonably be applied to a 10-year period. The Midpoint model uses the same code as indicated for "C - NLIN (LOSS), Midpoint X values." The Interpolated X model uses linear interpolation code shown in the appendix. Code for the Simulation model is specific to this problem and relies upon a particular data structure (see appendix).

Results of fitting data set D are in table 9, where two values are shown for -2LL, the "fit" value as reported by the regression program and the "true" value that would be obtained within the context of an annual simulation. The Compound model would not be applied within an annual simulation, nor can it be evaluated for the 10-year steps where it is likely to be applied. Thus, its true -2LL is unknown. The best fit is expected to be from the Simulation model. Here the mortality predictions within the regression are identical to those made within the simulator, and a

Table 9—Regression results, data set D. The NLIN (LOSS) procedure is used for fitting in all four cases. Two values of –2LL are shown: those from the fitting routine, and the "true" –2LL values that would be obtained by using the coefficient solutions in an annual growth simulation.

Model	b_0	b_1	b_2	b_3	–2LL (fit)	–2LL(true)
Compound[1]	0.1956	–1.6602	–2.1140	.058730	1606	
Midpoint[2]	–2.5369	–1.4604	–2.3461	.063074	1550	1596
Interpolated X[3]	–4.9798	–.9577	–1.4003	.04970	1538	1544
Simulation[4]	–2.9387	–1.0529	–1.7037	.0437	1538	1538

[1]Predicted annual survival rate is compounded for YIP years; X's are initial values. Suitable for 10-yr step.
[2]Annual model relying on midpoint X values. Reported –2LL =1550
[3]Annual model relying on annually interpolated X values.
[4]Annual model, with iteratively simulated annual X variables.

maximum likelihood solution is chosen; we do not address any interaction with maximum density constraints that *FVS* may employ. The Interpolated X method has a -2LL value that is almost as good; the Midpoint model is considerably poorer. A partial explanation for the interpolated X method being so good, is that a linear DBH growth model is assumed; a nonlinear model would not interpolate as well. Another result, not shown in the table, is the total number of mortality trees that would be predicted by an annual simulation. The observed mortality was 433 trees. The predicted mortality for the Midpoint, Interpolated X, and Simulation model fits were 518, 474, and 459, respectively. There is a tendency to overpredict mortality. Normally that might be indicative of model misspecification. Here there is the additional possibility that the simulation that created the data set was somehow defective.

Summary

Fitting a fixed-period length mortality model is straightforward. The model should directly predict the probability of survival (or mortality) for the entire period, without any requirement to compound the survival rate. Any of the methods presented here should be adequate for this task, with the exception that the Logit model for proportions should not be used unless the data are easily summarized into cells with a constant mortality rate. We offer this Logit model for proportions for historical completeness, not as a viable general alternative. The LOGISTIC regression procedure is the most robust. It is the easiest to apply, should not have convergence problems for typical data sets, and should normally be the preferred methodology. Each tree is one observation, and weighting is not used. Weighted least squares, NLIN(wtLS), should also be problem-free but is a bit more work. NLIN(LOSS) is the maximum likelihood option. It should provide good solutions, although convergence may slow as the number of parameters increases. RISK is a product of its time: a clever approximation to maximum likelihood when mainframe computer resources were limited. A FORTRAN compiler is required if model transformations cannot be done beforehand in an external data structure. Finally, a fixed-length survival model can be formulated as a direct prediction, as we have done here, or it could be formulated as a compound annual survival model for a fixed-length application. The appropriate methodology for fitting the latter formulation would be the same as for a variable-period length model, using initial X values.

For variable-length periods, the first decision is whether the model will be used in an annual simulation or in a simulation with a fixed step size of greater than 1 year. For the fixed step size, the data should be limited to growth periods of approximately that duration. The appropriate regression model is the Compound model in table 9, or the equivalent NLIN(LOSS) regression in table 8. For an annual simulation model, the best fit is assured with the Simulation model in table 9. In a real fitting exercise, the assumed DBH growth rates would be replaced by predicted growth rates (Cao 2000). As alternatives to a full Simulation model, reasonable approximations may be achieved with Interpolated X's or Midpoint X's. It is impossible to know in advance whether one of these approximate methods will be adequate. Due to the difficulty in programming the full simulation method, the use of midpoint X's may be the preferred method for model screening, and the use of interpolated X's may be adequate for final fitting. If the Midpoint or Interpolated X model is used, the Simulation model should be used to verify that the total predicted mortality is close to the observed total. This verification may be part of the validation of a completed growth simulator. Regardless of the formulation of the X's, the actual fitting method should be either NLIN (wtLS) or NLIN(LOSS).

In the foregoing, NLIN(wtLS) and NLIN(LOSS) are presented as having similar capabilities; in general, they both produce the ML solution. For the *SAS* system, NLIN(wtLS) seems to have better convergence properties than the NLIN(LOSS). For other statistical packages, this observation may not hold.

Acknowledgments

The Stand Management Cooperative, headquartered at the University of Washington, Seattle, provided the Douglas-fir data. Bill Wykoff, David Hamilton, Mark Hanus, and David Hann provided assistance with the RISK program. Additionally, Bill Wykoff patiently explained the possible ways for bringing mortality predictions into FVS. We thank Mark Hanus and Temesgen Hailemariam for helpful review comments.

References

Agresti, A.1990. Categorical data analysis. John Wiley & Sons, Inc., New York. 558 pp.

Allison, P.D.1995. Survival Analysis Using the *SAS*Æ System: A Practical Guide. Gary, NC: *SAS* Institute Inc., 292 pp.

Cao, Q. V. 2000. Prediction of annual diameter growth and survival for individual trees from periodic measurements. Forest Science 46(1): 127–131.

Chatfield, C. 1996. The analysis of time series: an introduction (5th ed.). Chapman and Hall, NY. 283 p.

Ek, A. R. and R. A. Monserud. 1974. FOREST: a computer model for simulating the growth and reproduction of mixed species forest stands. Univ. Wisconsin, College of Agriculture and Life Sciences Research Report R2635. 90 p.

Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25: 489–504.

Hamilton D.A., Jr. 1974. Event probabilities estimated by regression. USDA Forest Service Research Paper INT–152, 18p.

Hamilton, D.A. 1994. Extending the range of applicability of an individual tree mortality model. Can. J. For. Res. 20:1212–1218.

Hamilton, D.A., Jr. 1986. A logistic model of mortality in thinned and unthinned mixed conifer stands of Northern Idaho. Forest Science 32(4): 989–1000.

Hamilton, D.A. and B.M. Edwards. 1976. Modeling the probability of individual tree mortality. USDA, For. Serv. Res. Pap. Intermountain Forest and Range Exp. Station. INT–185. Ogden, UT. 22 pp.

Hann, D. W., A. S. Hester, and C. L. Olsen.1997. ORGANON User's Manual. Corvallis, OR.

Hasenauer, H., R.A. Monserud, and Gregoire, T.G. 1998. Using simultaneous regression techniques with individual tree growth models. Forest Science 44(1): 87–95.

Hosmer, D.W. and S. Lemeshow. 2000. Applied Logistic Regression (2nd Edition), New York, Wiley. 392 pp.

Hyink, D.M., W. Scott, and R. M. Leon. 1985. Some important aspects in the development of a managed stand growth model for western hemlock. pp. 9–21 *in* Forest Growth Modelling and Prediction. (ed: A.R. Ek, S.R. Shifley and T.E. Burk). USDA Forest Service General Technical Report NC–120.

Keister, T.D. 1972. Predicting individual tree mortality in simulated southern pine plantations. For. Sc. 18: 213–217.

McCullaugh, P. and J. A. Nelder, 1983. Generalized Linear Models. Chapman and Hall, New York. 261 pp.

Meeker, W. Q. 1998. Statistical methods for reliability data. John Wiley & Sons, Inc., New York. 680 pp.

Mendenhall, W. and R. L. Scheaffer. 1973. Mathematical statistics and applications. Duxbury Press, North Scituate, MA. 649 pp.

Mitchell, K. J. 1975. Dynamics and simulated yield of Douglas-fir. For. Sci. Monogr. 17.

Monserud, R.A. 1976. Simulation of forest mortality. Forest Sci. 22: 438–444.

Monserud, R.A. and Sterba, H. 1999. Modelling individual tree mortality for Austrian forest species. Forest Ecology and Management 113: 109–123.

Stage, A.R. 1973. Prognosis model for stand development. USDA, For. Serv. Res. Pap. Intermountain Forest and Range Exp. Station. INT-137. Ogden, UT. 32 pp.

Stage, A.R. and Wykoff, W.R. 1998. Adapting distance-independent forest growth models to represent spatial variability: Effects of sampling design on model coefficients. Forest Science 44 (2):224–238.

Teck, R., M. Moeur, and J. Adams (eds.). 1997. Proceedings: Forest vegetation simulator conference. 1997 Feb. 3-7, Fort Collins, CO. Gen. Tech. Rep. INT-GTR-373. Ogden, UT: U.S.D.A. For. Serv., Int. Res. Stn.

Walker, S.H. and D.B. Duncan, 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika 54: 167-179.

Wykoff, W.R., N.L. Crookston and A.R. Stage 1982. User's guide to the stand prognosis model. Gen. Tech. Rpt. INT–133.Ogden, Utah: U.S. Dept. of Agriculture, Forest Service, Intermountain For & Rge. Expt. Sta. 112 p.

Appendix

Code for selected regression examples, by data set (A, B, C, D).

A - LOGISTIC

Proc LOGISTIC for data set A (40 observations)

```
PROC LOGISTIC ORDER=INTERNAL DESCENDING;
  MODEL Y = X1 X2 ;
```

A - REG

Logit Model for proportions, using classed formulation of data set A

```
Data Step
  P = NY1 / N ;
  LOGITY = LOG(P/(1-P)) ;

PROC REG;
MODEL LOGITY = X1 X2;
WEIGHT N ;
```

A - RISK

Control lines for running data set A (40 records) with the RISK program.

```
TEST PROBLEM 1
          1          1      3      40        8
                  9999      3       5.0
    4
00000
(4F2.0)
 PROB XONE XTWO
000010000200003
```

B- LOGISTIC

```
PROC LOGISTIC ORDER=INTERNAL DESCENDING;
  MODEL Y= X1;
  OUTPUT OUT=fileout P = PRED;

PROC TABULATE;
  CLASS X1;
  VAR Y PRED;
  TABLE X1 ALL, SUM*( Y PRED);
```

B - LOGISTIC (for classed data set of 11 records)

```
PROC LOGISTIC;
  MODEL NY1 / N = X1;
```

B - NLIN (LS)

```
PROC NLIN ;
  PARMS B0 = -.5 B1 = -2 ;
  LOGIT = B0 + B1 * X1;
  Yhat = EXP(LOGIT) / (1 + exp(LOGIT)) ;
  MODEL Y = YHAT;
```

B- NLIN (wtLS)

```
PROC NLIN ;
  PARMS B0 = -.5 B1 = -2 ;
  LOGIT = B0 + B1 * X1;
  Yhat = EXP(LOGIT) / (1 + exp(LOGIT)) ;
  MODEL Y = YHAT;
  OUTPUT OUT=File2 P=Pred1 ;

PROC NLIN data=File2;
  PARMS B0 = -.5 B1 = -2 ;
  LOGIT = B0 + B1 * X1;
  Yhat = EXP(LOGIT) / (1 + exp(LOGIT)) ;
  VAR = Pred1 * (1-PRED1);
  _WEIGHT_ = 1/VAR ;
  MODEL Y = YHAT;
  OUTPUT OUT=File3 P=Pred2 ;
```

And optionally, one more iteration

B - NLIN(LOSS)

```
PROC NLIN ;
  PARMS B0 = -.5 B1 = -2 ;
  LOGIT = B0 + B1 * X1;
  Yhat = EXP(LOGIT) / (1 + exp(LOGIT)) ;
  IF Y = 1 then N2LLIKE = -2*LOG(Yhat);
           else N2LLIKE = -2*LOG(1-Yhat);
_LOSS_ = N2LLIKE;
MODEL Y = YHAT;
OUTPUT OUT=Filename P = PredY;
```

C - LOGISTIC

```
PROC LOGISTIC ORDER=INTERNAL ;
  MODEL MORT/YIP = DBH BAL_BA BA ;
note:  YIP = years in period.
       MORT=1 if tree dies, else 0.
       BAL_BA = BAL / BA
```

C - NLIN(LOSS).

```
PROC NLIN METHOD=DUD MAXITER=130;
  PARMS
    B0 = -3 B1 = -1 B2 = -1 B3 =0.05 ;

  LOGIT = B0 + B1 *DBH + B2 * BAL_BA + B3*BA;
  P_annual = exp(LOGIT) / (1 + exp(LOGIT) );
  P_surv = ( 1 - P_annual)**YIP;
  P_mort = 1 - P_surv ;
  IF MORT = 1 then N2LLIKE = -2*log(P_mort);
                  else N2LLIKE =-2*log(1 - P_mort) ;
  _LOSS_ = N2LLIKE;
  MODEL MORT = P_mort;
  OUTPUT OUT=OUTF P=Pmort;
```

The above does not converge satisfactorily. It was followed up by several regressions with B1 fixed at various values, including B1 = -1.38, and improved initial guesses for the other parameters.

```
note:  YIP = years in period.
       MORT=1 if tree dies, else 0.
       BAL_BA = BAL / BA
```

C - NLIN (LOSS), Midpoint X values.

Same as the above code, but preceded by a data step redefining the independent variables:

```
BAL_BA unchanged
F = (YIP/2 -0.5)/YIP ;
BA = BA + F*(BAfinal - BA) ;
DBH = DBH + F * (DBHgrown - DBH)
```

where BA_{final} is the basal area at the end of the period, and DBH_{grown} is from a by-plot regression of ending DBH (for the surviving trees) versus the initial DBH.

C - RISK

```
RISK with SMC data for DF
        1.        1.      4.     64121.     15.       0.      0.
        0.   999999.      5.        20.
  8
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
(8F9.5)
  A0BAL/B    BA   DBH   MORT   1     1     step
  1      2     3     4     5     6     7     8
```

Note: This is the control file that was actually used. The resultant coefficients are in a different order than is shown in our results (Table 8).

D - Interpolated X.

```
Data Step
  BAL_BA=BAL/BA;

  DO I=1 to NSURV;
  MORT=0;
  OUTPUT;
  END;

  NDEAD = NTREES-NSURV;
  if NDEAD> 0 then DO I=1 to NDEAD;
                        MORT=1;
                        OUTPUT;
                        END;

PROC NLIN METHOD=DUD ;
  PARMS B0 =-3 B1=-1 B2 =-1 B3=.04;

PER_SURV = 1.0 ;
Label PER_SURV='Pr{Surv for Period}';

  DO I = 1 to YIP;
    BAi = BA + (I-1)/YIP * (BA_FINAL-BA);
    DBHi = DBH + (I-1) * DBHRATE ;
    F = B0 + B2* BAL_BA + B3*BAi + B1*DBHi;
    P_mort = exp(F) / (1 + exp(F) );
    P_surv = ( 1 - P_mort);
    PER_SURV = PER_SURV * P_SURV;
  END;

PER_MORT = 1.0-PER_SURV;
IF MORT = 1 then N2LLIKE = -2*log(PER_mort);
                else N2LLIKE =-2*log(PER_SURV) ;

_LOSS_ = N2LLIKE;
MODEL MORT = PER_mort;
OUTPUT OUT=OUTF P=Pmort;
```

D - simulation model

```
*          Collapse to one record per plot;
DATA A;
  SET [original data]; by PLOT;
  ARRAY     VDBH{10}      DBH1-DBH10 ;
  ARRAY     VRATE{10}     RATE1-RATE10;
  ARRAY     VN{10}        N1-N10;
  ARRAY     VTPA{10}      TPA1-TPA10;
  ARRAY     VNSURV{10}    NSURV1-NSURV10;
  ARRAY     VBAL{10}      BAL1-BAL10;

   VDBH{GROUP}      = DBH;
   VRATE{GROUP}     = DBHRATE;
   VN{GROUP}        = NTREES;
   VTPA{GROUP}      = NTREES/SIZE ;
   VNSURV{GROUP}    = NSURV;
   VBAL{GROUP}      = BAL ;

  DEATHS = (TPA-TPAFINAL)*SIZE;

  if LAST.PLOT then output;

  RETAIN DBH1-DBH10
         RATE1-RATE10
         N1-N10
         TPA1-TPA10
         NSURV1-NSURV10
         BAL1-BAL10 ;

KEEP PLOT SIZE YIP TPA BA DEATHS
         DBH1-DBH10
         RATE1-RATE10
         N1-N10
         TPA1-TPA10
         NSURV1-NSURV10
         BAL1-BAL10;

PROC NLIN METHOD=DUD ;
  PARMS B0 =-3 B1=-1 B2 =-1 B3=.04;

ARRAY   VDBH{10}      DBH1-DBH10;
ARRAY   VRATE{10}     RATE1-RATE10;
ARRAY   VN{10}        N1-N10;
ARRAY   VTPA{10}      TPA1-TPA10;
ARRAY   VNSURV{10}    NSURV1-NSURV10;
ARRAY   VBAL{10}      BAL1-BAL10;
ARRAY   GDBH{10}      GDBH1-GDBH10;
ARRAY   GTPA{10}      GTPA1-GTPA10;
ARRAY   GBAL{10}      GBAL1-GBAL10;

  *       START-UP before growth;
  DO I=1 to 10;
          GDBH{I}=VDBH{I};
          GTPA{I}=VTPA{I};
          GBAL{I}=VBAL{I};
  END;
  GBA = BA;

DO   IYEAR = 1 to YIP;

BASUM=0;
DO I = 1 to 10;
  F = B0 + B2* GBAL{I}/GBA + B3*GBA + B1*GDBH{I};
  P_mort = exp(F) / (1 + exp(F) );
  P_surv = ( 1 - P_mort);
  GTPA{I} = GTPA{I} * P_SURV;
  GDBH{I} = GDBH{I} + VRATE{I};
  BAgroup = GTPA{I} * 0.005454154 *GDBH{I}*GDBH{I};
  GBAL{I} = BASUM + BAGROUP/2;
  BAsum = BAsum + BAgroup ;
END;
  GBA= BAsum;
END;

*          Evaluate all the losses;
  N2LLIKE=0;
  PRDeaths = 0;
  DO I = 1 to 10;
  PR_SURV = GTPA{I}/VTPA{I} ;
  PR_MORT = 1-PR_SURV;
  PrDEATHS = PRDEATHS + VN{I} * PR_MORT;

  CNT_DEAD = VN{I} - VNSURV{I};
  CNT_LIVE = VNSURV{I};

  THISLOSS = CNT_DEAD * ( - 2*log(PR_MORT))
                +CNT_LIVE * (- 2* log(PR_SURV)) ;
  N2LLIKE = N2LLIKE + THISLOSS;
  END;
  _LOSS_ = N2LLIKE;
  MODEL DEATHS = PRDEATHS;
  ID GTPA1-GTPA10;
  OUTPUT OUT=OUTF P=Pred;
PROC PRINT;
```