Rate My Stake: Interpretation of Ordinal Stake Ratings

Patricia Lebow
Grant Kirker
USDA Forest Products Laboratory
Madison, Wisconsin

ABSTRACT

Ordinal rating systems are commonly employed to evaluate biodeterioration of wood exposed outdoors over long periods of time. The purpose of these ratings is to compare the durability of test systems to nondurable wood products or known durable wood products. There are many reasons why these systems have evolved as the chosen method of evaluation, including having an overall evaluation that incorporates several characteristics thought to correlate with some "failure," balancing bias and variability because of multiple raters and simplifying calculation and interpretation. Appropriate analysis of the ordinal ratings is needed to accomplish this goal. We will briefly review the standards and associated methods and models currently in use, as well as several proposed analysis methods in relation to an example FPL data set.

Keywords: field stake test data, data quality, statistics, field performance.

INTRODUCTION

Short review of long history of field stake evaluations

Ordinal rating systems are commonly used to evaluate biodeterioration of wood exposed outdoors over long periods of time. Rating measurements taken repeatedly over time are used to measure wood degradation, as well as other phenomena, and have been used in American Wood Protection Association (AWPA) standards for decades. AWPA rating measurements, taken at particular exposure times, are basically ordered categorical measurements and have been refined over the years to portray the underlying progression of degradation as a function of decay, termites, or other agents. Typically, this progression is seen as being continuous in nature. However, it is difficult, if not impossible, to fully quantify some of the continuous measurements of degradation or even define "failure," and so ordinal rating systems have evolved as the chosen method of evaluation. Several AWPA standards refer to measurements that result in ordinal data, including evaluation standards E1, E5, E7, E14, E16, E18, E21, E23 and E24, although most refer to the standard Decay and Termite Visual Rating Scales.

In terms of preservative field stake ratings, the durability trends from this type of measurement over long periods of time indicate how particular preservatives will perform in-service. Although generally not based on statistical methods of comparison, they allow some differentiation of preservative performance over time in comparison to either an untreated control and/or some established treated control.

Cook and Morris (1995) and Morris (1998) provide an excellent history of field rating data evaluation, including efforts by Colley (1970, 1984) and Hartford (1972) (Hartford and Colley, 1984; Morris and Cook, 1995; Morris and Rae, 1995). This research has typically concentrated on characterizing the mean ordinal response over time and does not account for within-stake or stake-to-stake variation in evaluating differences between treatments. Morris (1998) also investigated the use of modes as being more indicative of individual stake behavior.

Other efforts have tried to understand and characterize deterioration behavior by using time-to-failure approaches that are common for answering reliability questions, such as how long will something last? Historically, this is how the Forest Products Laboratory (FPL) has presented data in their stake test progress report (Woodward, et al. 2011) with failure represented by some combinations of decay/termite ratings. Bartlett (1978) used a proportional hazards model to model probability of survival of field stakes in Australia. To develop the model, he chose a decay rating of "7" or below as a failure on a 9-point rating scale. He also discussed issues related to inspection times, in that stakes cannot be continuously rated but are in fact rated on some type of inspection intervals. Further, Link and De Groot (1989, 1990) emphasize the ordinality of ratings and discourage the sole use of average ratings or average lifetimes as they fail to account for stake variability. Their work discusses time-to-failure approaches with attaining a specific rating as a definition of failure, such as time-to-7 or time-to-0. Choosing a particular failure and then analyzing the sample stake data of time-to-failure allows the investigation of the sample distribution of that failure, including the ability to look at quantiles such as the first quartile (25^{th} percentile) and median (50^{th} percentile), and to compare these sample estimates for field trials of different preservatives. Also, they illustrate the usefulness of boxplots in displaying and comparing sample characteristics. Link and De Groot (1990) compare different

estimators of median lifetime and discuss construction of confidence intervals for a population median lifetime. Further work in De Groot and Evans (1998, 1999) studies how possible dose-response behavior in preservative systems is characterized by quartile and median estimates.

Survivability/reliability type methods, such as the above, have the advantage that under, the appropriate assumptions, lifetime estimates can be established with certain degrees of confidence. With the use of analysis methods that accommodate censoring; this type of analysis does not require waiting until all stakes fail.

Although the current models appear easy to understand and interpret in a perfect world, they fail to fully capture several aspects of possible durability behavior, such as identifying populations with early stages of biodeterioration activity or behavior that does not attain "failure" over an inspection period. In fact, the definition of failure leads to many issues (such as different models), and we would ideally have models that characterize time-dependent behavior while not relying on a particular definition of failure, such as the time to reach a specific rating. Because studies may often be performed at different locations, they may have different inspection intervals and neither of the preceding methods can easily accommodate this. Since rating inspections occur at intervals, "time-to" analyses can produce optimistic estimates.

Ordinal rating measurements are common in the medical community, and many statistical advances have been made in enhancing analysis of these types of data. Although many of the examples in Agresti (2010) and Molenberghs and Verbeke (2005) are medically related, both books are excellent statistical resources that discuss statistical methodology related to analyzing ordinal data, especially repeated ordinal data such as occur in stake field trials. Both discuss the advantages of modeling the probabilities of ordinal data and the various types of statistical models that are commonly used. Agresti, in particular, gives advantages and disadvantages of mean ordinal response models as opposed to models that specifically characterize probabilities of ordinal ratings or categories. Limitations of mean response models result from not having the ability to provide estimates of rating probabilities; thus, predictions are not constrained to the boundaries of the rating scale, nor do they model the functional dependencies of ratings (i.e., if you have a certain number of stakes, the sum of their frequencies is constrained). However, if an ordinal rating portrays an underlying continuous variable, and one were able to increase the number of rating categories to a moderate number of categories, then a mean response model may approximate the regression model that would result from actual measures of the underlying continuous variable (Agresti, 2010). Both authors discuss several methods for developing and analyzing ordinal probabilistic regression models for repeated ordinal data. Several studies have used these types of models for modeling ordinal deterioration/degradation, for example, in stake studies (Lebow, et al. 2007, 2009) and panel coating studies (Gobakken and Lebow, 2010; Gobakken, et al. 2010). Also, for continuous responses, it has been shown that degradation models provide more efficient estimates of reliability parameters than time-to-failure data (Hamada, 2005).

Brief review of general statistics

Hypothetically, first suppose that a random sample of a single group of treated stakes, say *n*, are exposed for a fixed number of years, at which time each is given an ordinal visual rating according to Table 1.

Table 1: Decay rating scheme from AWPA E7-09 (AWPA 2011)

Rating	Condition	Description
10	Sound	No sign or evidence of decay, wood softening or discoloration caused by microorganism attack
9.5	Trace-suspect	Some areas of discoloration and/or softening associated with superficial microorganism attack
9	Slight attack	Decay and wood softening is present. Up to 3% of the cross-sectional area is affected
8	Moderate attack	Similar to "9" but more extensive attack with 3–10% of cross-sectional area decayed
7	Moderate/Severe attack	Sample has between 10–30% of cross-sectional area decayed
6	Severe attack	Sample has between 30 – 50% of cross-sectional area decayed
4	Very severe attack	Sample has between 50 – 75% of cross-sectional area decayed
0	Failure	Sample has functionally failed. It can either be broken by hand due to decay, or the evaluation probe can penetrate through the sample

Each sample stake observation typically could be considered as coming from a discrete distribution with a particular probability associated with each rating. If we're willing to assign values (also known as scores) as in the above rating scheme such that differences between ratings are interpretable, then an expected value and variance could be determined for a single random stake observation. However, this is not ideal as it is difficult to interpret, and depending on what you are estimating, can result in estimates outside of the score rating limits.

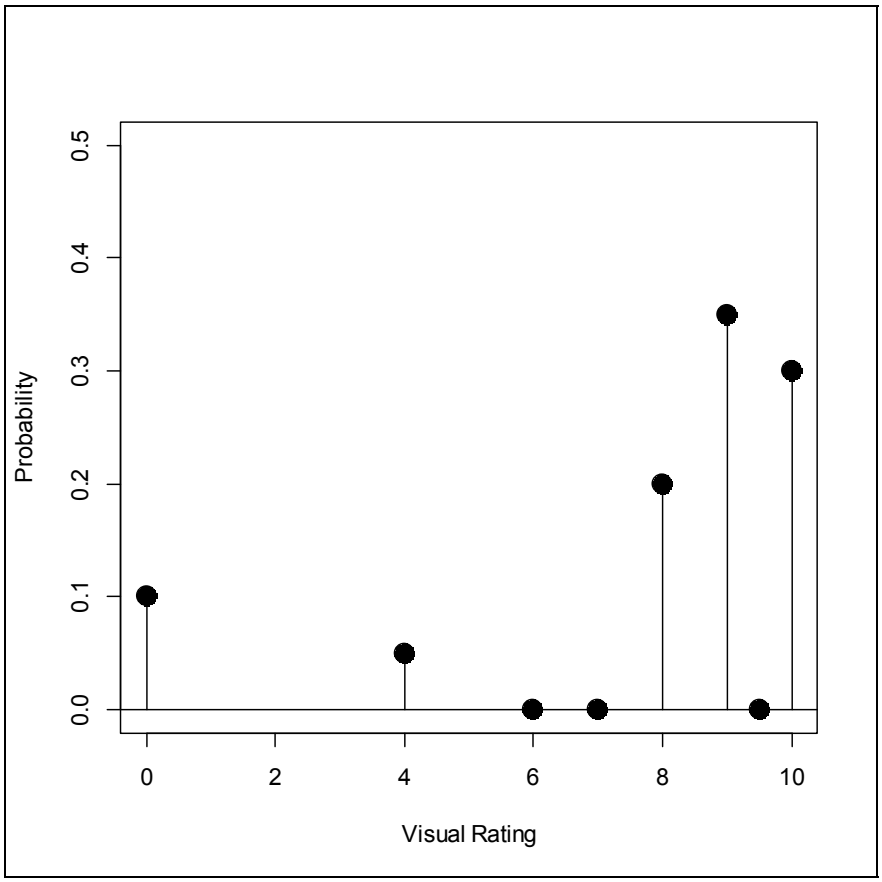

Figure 1: Hypothetical categorical distribution of visual ratings

As it is, however, the probabilities that chiefly define the distribution, we may consider the observations as coming from a categorical distribution, such as displayed in Figure 1. We are really interested in comparing those discrete probabilities for each of the rating categories. Then with multiple observations, such as occurs with our random sample, we assume that the underlying distribution is multinomial, which further describes the probabilities of joint occurrence in terms of the ratings' underlying frequencies, the possible combinations to which they could occur in a sample, and the underlying probabilities of each rating. First associate an R_k, $k = 1, 2,\ldots 8$ with each rating in Table 1, and let p_i be the probability that rating R_k will occur for a single stake ($p_1+p_2+\ldots+p_k = 1$) and assume n stakes will be tested. Then the probability that R_1 occurs x_1 times, R_2 occurs x_2 times, etc., such that $x_1+x_2+\ldots+x_k=n$, is given by

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

What statistics can we use to describe the data? For a single group of data, a tabular summary of counts (or percentages) at each rating is the most useful. This could be supplemented with the mean, median, and mode, as well as a standard deviation and some type of range. Although the mean and standard deviation are not ideal, they can give indications of central tendency and variation, especially as the ratings would approach some underlying measurement on at least an interval scale. For two independent groups of data, say, treated and untreated at one time point, again, a tabular summary of the two

rating counts for each of the two groups is best if you have the same number of observations in each group. Otherwise, you can list the observed percentages with sample sizes. Comparisons for the two groups then could be made with chi-square two-sample tests or Kolmogorov–Smirnov tests as outlined in Dickinson Gibbons (1985, esp. see chapter 6 and page 250). These methods are also extended for comparisons between more than two groups. Other nonparametric methods based on ranks that compare some central tendency (i.e., location parameter) can also be used (Mann–Whitney–Wilcoxon for two groups, Kruskal–Wallis for more than two groups), but it should be kept in mind that the ordinal rating data usually have an extensive number of equal valued responses (i.e., ties) that could compromise the accuracy of associated tests.

Statistical methods to study multiple time points

AWPA rating data are usually acquired on two (or more) groups of the same specimens repeatedly over intervals of time. Several types of models can be used to describe this type of data and to answer some of the questions, such as, after 5 years exposure, are these two (or more) treatments different? The Introduction reviewed methods that have previously been used in interpreting these types of data. We have found success in using models that model functions of the probabilities of ratings as linear functions of treatment or time, that is,

Function (rating probabilities) = linear function (factors, such as treatment type, retention, time)

Repeated ratings over time can complicate analysis if there is an additional dependency due to the particular stake repeatedly being measured. Modeling and evaluation of these dependencies are discussed in Agresti (2002) and Molenberghs and Verbeke (2005). We further have been able to accommodate the dependencies of repeat measurements, similar to models outlined in Gobakken and Lebow (2010). There are multiple statistical methods that can be used to fit the models, and depending on the structure of the design as well as the questions you are trying to answer and the assumptions you are willing to make, one method may be more appropriate than another. In several of the data sets that have been studied so far, the generalized linear mixed models with cumulative logit link modeled as a linear function (proportional odds structure) of treatment and single or separate slope parameters for time or log-time, along with a random effect to capture the repeat measurements on each stake appear promising. To illustrate what can be done statistically with these models, we report on three treatments (one treatment at three concentrations) and an untreated control of 19 mm^2 stakes that were exposed in Mississippi for 11 years. For this set of treatments, the ratings 0, 4, 6, 7, 8, 9, and 10 were used.

In this analysis, the following cumulative logit model that models the probability of higher stake ratings versus lower stake ratings as a function of the fixed effects, including preservative treatment and exposure time, and a random effect due to repeated measurements was assumed:

$$\text{logit}[P(Y_{ijk} \geq R \mid X_i, t_{ij}, Z_k)] = \beta_R + \beta_{1i} X_i + \beta_{2i} X_i \ln(t_{ij}) + b_k Z_k$$

where
$\text{logit}(p) = \log(p/(1-p))$
Y_{ijk} = stake rating ijk
R = 10, 9, 8, 7, 6, 4 (visual rating, the last rating 0 is implied from the model)
X_i = treatment group i (A, B, C, U)
t_{ij} = exposure time j (1, 2, 3, 4, 5, 7, 9, 11 years)
b_k = random effect for stake k
Z_k = stake k from treatment group i
β = model parameters.

This generalized linear mixed model assumes the random effects for stake k, b_k, are independently, normally distributed with mean zero and variance τ^2. Models were fit using SAS® V9.2 (Cary, NC) with the GLIMMIX procedure. Hypotheses of interest were tested with linear contrasts between treatment groups after fixed exposure times, and p-values were adjusted using the simulation multiple comparison procedure. That is, performance comparisons of the treatments i and j were made using comparisons of log odds ratios at time t:

$$\begin{aligned}\text{logit}[P(Y \geq R \mid X_i)] - \text{logit}[P(Y \geq R \mid X_j)] &= \log\left[\frac{P(Y \geq R \mid X_i)/P(Y < R \mid X_i)}{P(Y \geq R \mid X_j)/P(Y < R \mid X_j)}\right] \\ &= \beta_{1i} X_i + \beta_{2i} X_i \ln(t) - \beta_{1j} X_j - \beta_{2j} X_j \ln(t)\end{aligned}$$

and were constructed using various levels of data to determine how the models performed in differentiating the treatments over time with the availability of different amounts of data. From the models, we can estimate probabilities of each rating at each time point, and then compare the model's average evolution over time versus what is the typically calculated mean rating response over time (Figure 2).

An odds ratio close to one indicates no detectable difference in ratings between groups (i.e., either no difference or insufficient evidence to detect a difference). Because cumulative probabilities of higher visual ratings were modeled, an odds ratio greater than one indicates the first part of the hypothesis has higher probability of higher ordered ratings than the second part of the hypothesis. Correspondingly, an odds ratio less than unity indicates that the second part of the hypothesis has a higher probability of higher ordered ratings than the first part. There are multiple model fitting methods available for generalized linear mixed models, even within SAS, including those that fit the marginal model, such as general estimating equations (with proc GENMOD) and data approximation via iterations of linear mixed methods (with proc GLIMMIX, see Molenberghs and Verbeke (2005), or The GLIMMIX Procedure (SAS 2008) for details).

From these models, we obtain estimates of probabilities of the occurrence of different ratings over time. We can then simulate a random effect that will shift the probabilities for an individual stake, and then using the rating values, calculate what the expected average rating for that stake would look like. With this collection of simulated stakes of expected ratings, we then can calculate their average to estimate what the average evolution of may look like. Figure 2 shows 20 simulated expected stake rating responses for each of the four treatments from the models fit with 11 years of data; it also shows the average evolutions for each of the treatments (based on 1,000 simulations each) as well as the typical observed average rating.

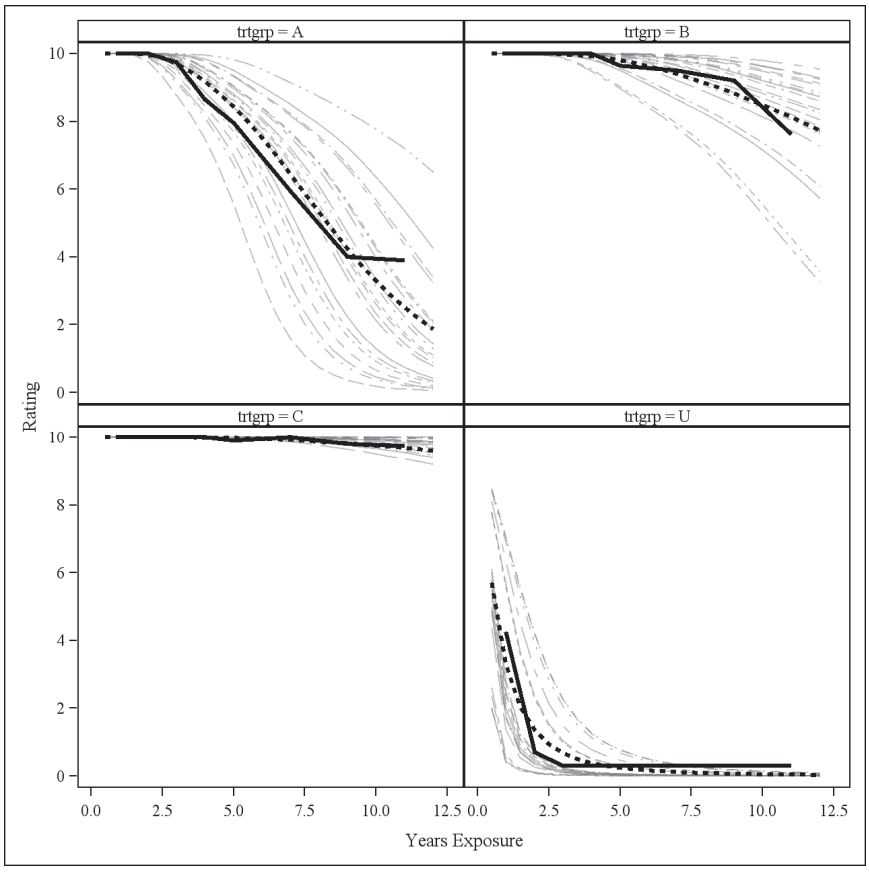

Figure 2: Observed mean rating (solid heavy line) and average evolution (dashed heavy line) over time for three treatments (A, B, C) and untreated control (U) from model fit with 11 years of data. Graph supplemented with expected responses of 20 simulated "stakes" for each treatment (gray lines)

Based on the model's fit with 11 years of data, we can compare the treatments at multiple time points to determine where differences occur statistically (see Table 2). The untreated is different at all tested time points, whereas A and B become statistically different at 3 years, A and C at 3 years, and B and C at 7 years.

Table 2: Treatment comparisons of functions of the probabilities for the treatments were performed and adjusted based on models fit with 11 years of data. Values below 0.05 indicate a significant difference

Treatment comparison	Adjusted *P*-value of treatment difference at specified years					
	1	3	5	7	9	11
A vs B	0.3968	0.0034	<0.0001	<0.0001	<0.0001	<0.0001
A vs C	0.5558	0.0156	<0.0001	<0.0001	<0.0001	<0.0001
A vs U	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
B vs C	0.9214	0.4601	0.0984	0.0017	<0.0001	0.0002
B vs U	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
C vs U	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

To illustrate how these models perform under the availability of differing periods of data, we fit the same models to 3 years, 5 years, and 7 years of rating data. Figure 3 shows the predictive ability of these models to predict average evolutions to years beyond that used to fit them. As expected, as more long-exposure data become available, the predictive performance improves. Because we are modeling probabilities, their predictions will stay within the range of rating values.

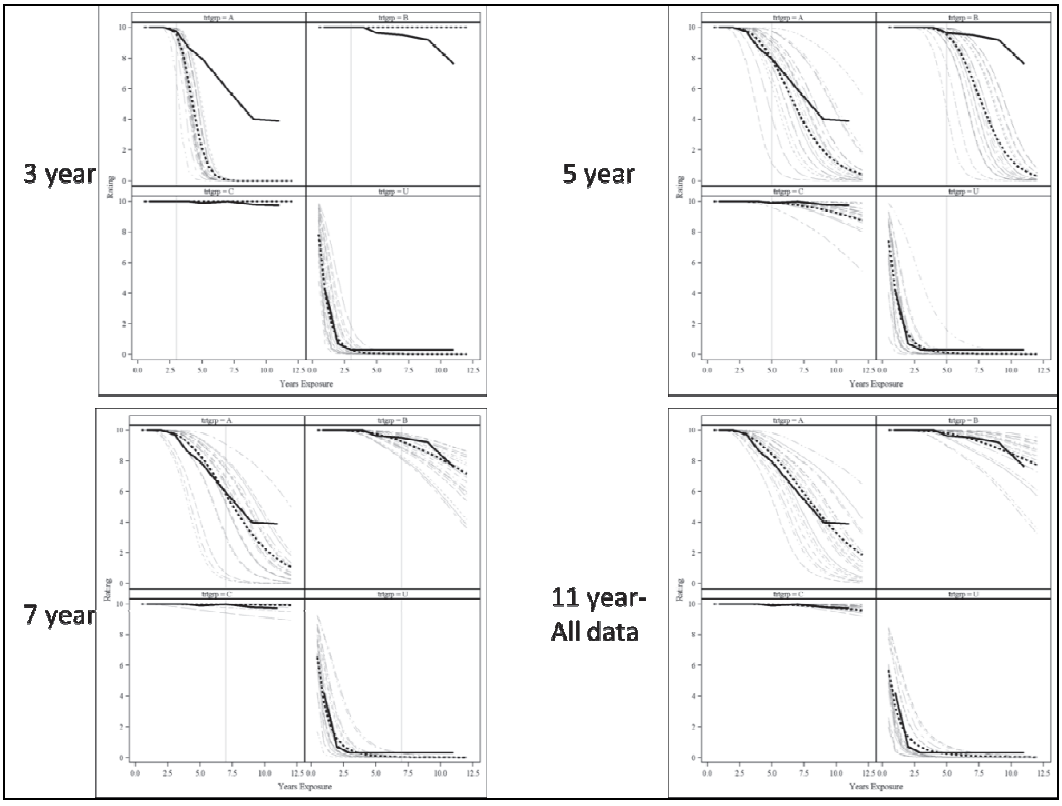

Figure 3: Based on the availability of 3, 5, 7 or 11 years of data, each of sub-figures gives the respective model's predictive performance of trend. The average evolution (dashed heavy line) over time for three treatments (A, B, and C) and untreated control (U) from models overlayed with the actual observed mean ratings (solid heavy line). Graph supplemented with expected responses of 20 simulated "stakes" for each treatment (gray lines)

To compare these models' ability to detect treatment differences after various exposure periods, we re-ran treatment comparisons for each of the models (and additionally a model based on 4 years of data). See Table 3. This shows that with 3 or 4 years of data, it is difficult to detect treatment differences, but at 5 years reasonable differences are detected. With 7 years, we can predict differences at 11 years.

Table 3: Models were fit with increasing levels of data (3, 4, 5, 7 years) and mean comparisons of functions of the probabilities for the treatments were performed and adjusted (multiple comparisons adjusted within years of data model is based on). Values below 0.05 indicate a significant difference

Years of data model is based on (Years)	Treatment comparison	Adjusted *P*-value of treatment difference at specified years					
		1	3	5	7	9	11
3	A vs B	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	A vs C	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	A vs U	0.8139	0.0007	0.9864	0.9999	1.0000	0.9999
	B vs C	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	B vs U	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	C vs U	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	A vs B	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	A vs C	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	A vs U	0.0005	<0.0001	0.0005	0.0160	0.1261	0.3248
	B vs C	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	B vs U	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	C vs U	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
5	A vs B	0.6327	0.2923	0.0217	0.8544	0.9955	0.9955
	A vs C	0.9955	0.4267	0.0027	0.1808	0.3579	0.4654
	A vs U	<0.0001	<0.0001	<0.0001	0.0005	0.0211	0.1260
	B vs C	0.8389	0.9955	0.2613	0.4369	0.5273	0.5786
	B vs U	0.0259	<0.0001	<0.0001	0.0125	0.2923	0.6327
	C vs U	0.2613	<0.0001	<0.0001	0.0009	0.0336	0.1333
7	A vs B	0.9074	0.0840	<0.0001	<0.0001	0.0011	0.0049
	A vs C	0.9074	0.0840	<0.0001	<0.0001	<0.0001	0.0004
	A vs U	<0.0001	<0.0001	<0.0001	<0.0001	0.0001	0.0049
	B vs C	0.8434	0.8023	0.0700	0.0281	0.0495	0.0700
	B vs U	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
	C vs U	0.0094	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

We are working with further data sets to determine the applicability of the modeling approach to other treatment types. One advantage of this approach is the modeled relationship between stake behavior and collective behavior that can be built into simulation models.

FUTURE MODELING AND CONCLUSION

Ideally, we would like to establish models for standard treatments and use simulation to determine best methods for determining equivalency or noninferiority to those treatments. This will also allow the comparison of the ordinal degradation model approach against time-to-failure approaches. Currently we have developed a simulation model based on this particular set of data that will help in the design of future simulation models for other data sets. Figure 4 illustrates how the observed data for 20 stakes in treatment group A compares to simulated ratings over time for 20 stakes from treatment group A using parameter estimates from our model.

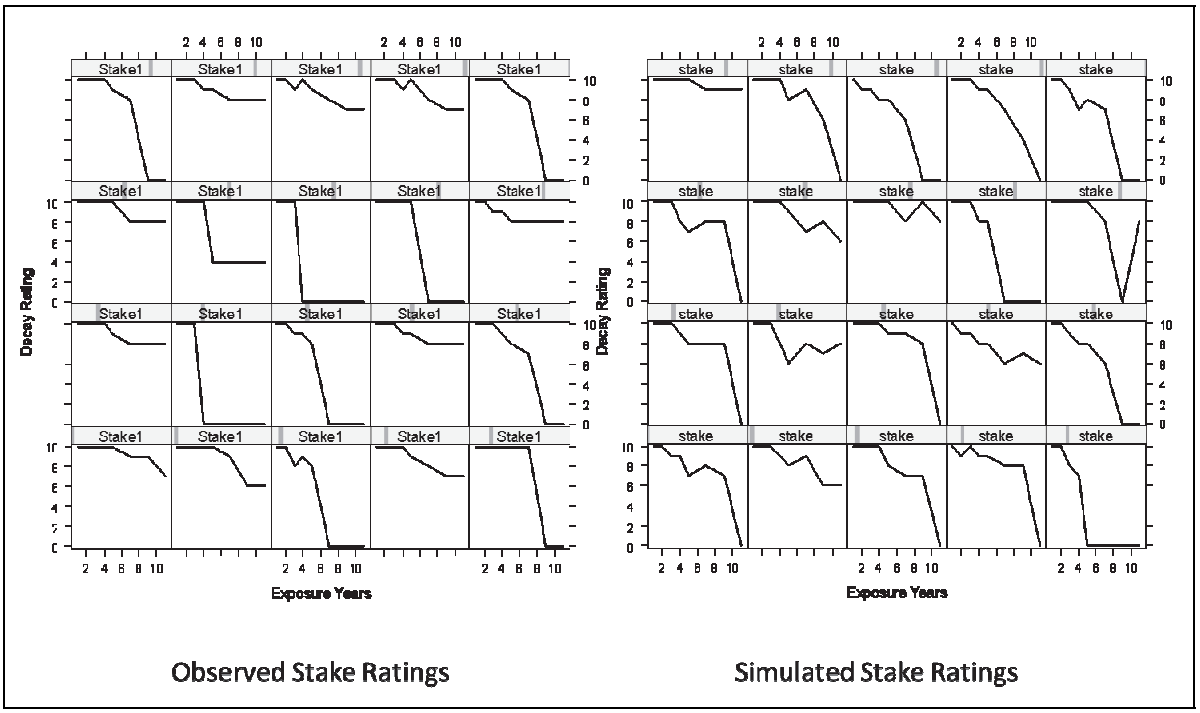

Figure 4: Observed stake ratings for 20 stakes from treatment group A and a set of simulated "raw" ratings for 20 stakes

REFERENCES

1. Agresti, A. 2010. Analysis of ordinal categorical data. Hoboken, NJ: John Wiley & Sons, Inc. 396 pp.
2. AWPA. 2011. Book of standards. Birmingham, AL: American Wood Protection Association. 582 pp.
3. Bartlett, N.R. 1978. A survival model for a wood preservative trial. Biometrics. 34(4):673-679.
4. Colley, R.H. 1970. The practical meaning of preservative evaluation tests. In: Proceedings, American Wood Preservers' Association. Washington, DC. 66:16-35.
5. Colley, R.H. 1984. What is the log-probability index? In: Proceedings, American Wood Preservers' Association. Stevensville, MD. 80:247-256.
6. Cook, J.A.; Morris, P.I. 1995. Modelling data from stake tests of waterborne wood preservatives. Forest Prod. J. 45(11/12):61-65.
7. De Groot, R.C.; Evans, J.W. 1998. Patterns of long-term performance—how well are they predicted from accelerated tests and should evaluations consider parameters other than averages? IRG/WP/98-20130. Stockholm, Sweden. International Research Group on Wood Preservation. 12 pp.
8. De Groot, R.C.; Evans, J. 1999. Does more preservative mean a better product? Forest Prod. J. 49(9):58-68.
9. Dickinson Gibbons, J. 1985. Nonparametric methods for quantitative analysis. Second Edition. Columbus, OH: American Sciences Press, Inc. 481 pp.
10. Gobakken, L.R.; Høibø, O.A.; Solheim, H. 2010. Factors influencing surface mould growth on wooden claddings exposed outdoors. Wood Mater. Sci. Eng. 5(1):1-12.
11. Gobakken, L.; Lebow, P.K. 2010. Modelling mould growth on coated modified and unmodified wood substrates exposed outdoors. Wood Sci. Technol. 44(2):315-333.
12. Hamada, M. 2005. Using degradation data to assess reliability. Qual. Eng. 17:615-620.
13. Hartford, W.H. 1972. The interpretation of test plot data. In: Proceedings, American Wood Preservers' Association. Washington, DC. 68:67-82.
14. Hartford, W.H.; Colley, R.H. 1984. The rationale of preservative evaluation by field testing and mathematical modeling. In: Proceedings, American Wood Preservers' Association. Stevensville, MD. 80:257-269.
15. Lebow, S.; Woodward, B.; Lebow, P. 2007. Ground contact durability of wood treated with borax-copper preservative. In: Proceedings, American Wood Protection Association. Birmingham, AL. 103:82-87.

16. Lebow, S.T.; Lebow, P.K.; Woodward, B.M.; Halverson, S.A.; Abbott, W.; West, M.M. 2009. Efficacy of a borax-copper preservative in exposed applications. Res. Pap. FPL-RP-655. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 11 pp.
17. Link, C.L.; De Groot, R.C. 1989. Statistical issues in evaluation of stake tests. In: Proceedings, American Wood Preservers' Association. Stevensville, MD. 85:170-185.
18. Link, C.L.; De Groot, R.C. 1990. Predicting effectiveness of wood preservatives from small sample field trials. Wood Fiber Sci. 22(1):92-108.
19. Molenberghs, G.; Verbeke, G. 2005. Models for discrete longitudinal data. New York: Springer-Verlag. 683 pp.
20. Morris, P.I. 1998. Beyond the log probability model. In: Proceedings, American Wood Preservers' Association. Granbury, TX. 94:267-273.
21. Morris, P.I.; Cook, J.A. 1995. Predicting preservative performance by field testing. In: Wood preservation in the '90s and beyond. Proc. No. 7308. Madison, WI: Forest Products Society. pp.116-121.
22. Morris, P.I.; Rae, S. 1995. What information can we glean from field testing? IRG/WP 95-20057. Stockholm, Sweden. International Research Group on Wood Preservation. 16 pp.
23. SAS Institute, Inc. 2008. SAS® user's guide, Version 9.2. Cary, NC: SAS Institute, Inc.
24. Woodward, B. M., Hatfield, C.A., Lebow, S.T. 2011. Comparison of Wood Preservatives in Stake Tests: 2011 Progress Report. Research Note FPL-RN-02. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 120 p.

Lebow, P; Kirker, G. 2014. Rate my stake: Interpretation of ordinal stake ratings. Proceedings of the One Hundred Eighth Annual Meeting of the American Wood Protection Association, Nashville, TN, April 29-May 2, 2012. Birmingham, AL. 108: pp 156-164.