Comparison of Performance Criteria for Evaluating Stake Test Data

Stan Lebow
Grant Kirker
Patricia Lebow
USDA, Forest Service, Forest Products Laboratory
Madison, Wisconsin

ABSTRACT

Stake tests are a critical part of evaluating durability of wood in ground-contact, but there is a lack of criteria for interpreting stake test results. This paper discusses criteria that might be used to determine if short term ratings indicate satisfactory long-term performance. Ratings of 19 by 19 mm stakes from multiple plots in the Harrison Experimental Forest, Mississippi, were analyzed to determine how well performance criteria at years three, four, or five predicted durability over the longer term (nine years). Results revealed that with only three years of data, false prediction of satisfactory performance was common even when strict criteria were applied. Predictions of future performance improved with four and five years of data, although some instances of false prediction still occurred. These analyses indicate that with three to five years of data, even very slight differences between ratings of a Test treatment and a Reference preservative may indicate unsatisfactory future performance. However, variability in the ratings of the standard reference preservative can make it difficult to detect slight differences and increasing the number of replicates may be warranted. More complex statistical approaches that incorporate all years of rating data into comparisons may help to account for year to year variation in ratings. This analysis was limited to one test site and a similar type of analysis for other test sites would be beneficial.

Keywords: stake testing, ratings, performance criteria, prediction

INTRODUCTION

Predicting the long-term durability of treated wood products with field tests remains a difficult challenge. Stake tests have long been the primary method for evaluating ground-contact applications and their interpretation has been the subject of numerous papers (Hartford, 1972; Hartford and Colley, 1984; Lebow and Kirker, 2014; Link et al., 1989, 1990; Morris, 1998; Morris and Cook, 1995; Morris and Rae, 1995). However, questions remain about stake test methodology and data interpretation. These questions include minimum duration of testing, establishing criteria for acceptable performance, and the relationship between stake test data and in-service performance of actual wood products (Lebow et al., 2008).

Duration of testing can be viewed as a compromise between obtaining meaningful data and minimizing the time needed to complete testing. American Wood Protection Association (AWPA) guidelines call for a minimum of three years of stake test data at high decay hazard sites (AWPA, 2016a), although it remains unclear whether three years is always sufficient (Lebow et al., 2009). At a given year of testing the performance of the test formulations is typically judged by comparing the average ratings of the test formulations to that of a standardized reference preservative. The criteria used for judging acceptable performance relative to the reference preservative remains a work in progress. Recently, the Committee P6 Minimum Effective Threshold task group recommended that the average ratings of the test formulation be greater than or equal to (at least 100% of) the reference preservative after three and four years of exposure. Recognizing that average ratings might conceal early failures, the task group also recommended that no more than 10% of the stakes have a rating of below "9" at three years or below "7" at four years. It was subsequently suggested that the average ratings of the test preservative should not be significantly less than that of the test preservative at a 95% confidence level (AWPA, 2016b). However, statistical testing methods have not been defined and the effect of using any of these criteria needs further evaluation. Part of the difficulty in determining test duration and performance criteria is attributable to the uncertain relationship between stake test results and the durability of commodity-size products. The use of relatively small (typically 19 by 19 mm) test stakes is intended to accelerate testing but the extent of acceleration is unclear. In fact, the relationship between deterioration rate and wood dimension is highly variable (Lebow et al., 2010) and is likely a function of site conditions, preservative characteristics, and types of attacking organisms.

In this paper we analyze 19 by 19 mm stake rating data from USDA Forest Products Laboratory (FPL) stake plots located in Harrison Co., Mississippi, with the goal of improving our understanding of the role of stake test duration and performance criteria in predicting preservative performance.

MATERIALS AND METHODS

The data used in this analysis are from the FPL test site within the Harrison Experimental Forest (HEF), located in southern Mississippi, approximately 15 miles north of Gulfport-Biloxi. This site is categorized as Deterioration Zone 5 according to the

AWPA and is considered a severe deterioration hazard (AWPA, 2016c). The site has active native subterranean termite populations but does not yet exhibit evidence of Formosan termite colonies. The stakes were placed in rows with 305 mm between stakes within each row and with 915 mm spacing between rows. At each inspection, the stakes were scraped lightly to remove soil and given a visual rating for decay and/or termite attack according to AWPA Standard E7 (Table 1) (AWPA, 2003). Note that the ratings for these stakes did not utilize the "9.5" rating currently included in AWPA Standard E7 (AWPA, 2016d).

Table 1. Rating scheme for evaluation of stakes.

Rating	Description of Stake Condition
10	No evidence of attack.
9	Slight attack, up to 3% of cross-sectional area.
8	Moderate attack, up to 10% of cross-sectional area.
7	Moderate to severe attack, up to 30% of cross-sectional area.
6	Severe attack, up to 50% of cross-sectional area.
4	Very severe attack, up to 75% of cross-sectional area.
0	Failure. Can be easily broken by hand or the evaluation probe can penetrate through stake.

The 19 by 19 mm stake plot data was analyzed to determine how well performance after three, four, or five years predicted performance over the longer term (i.e. nine years). The selection of nine years as the "long-term" comparison was based on data availability. It is recognized that nine years is an arbitrary time point and, as discussed above, we have a poor understanding of how nine years of 19 by 19 mm stake data compares to in-service durability of treated commodities.

Multiple plots using 19 by 19 mm stakes have been installed at HEF since the 1990's. Plots selected for analysis met the following criteria:

- Average rating of untreated controls less than "4" after two years (the criterion of a "severe" deterioration hazard recommended by Committee P6 Minimum Effective Threshold task group).
- Ratings available for one, two, three, four, five, and nine years.
- Data available for softwood species.

These criteria yielded six plots containing 102 treatment groups. Treatment groups had either 20 (five plots) or 50 (one plot) replicate stakes. The wood species was southern pine, except for one plot of Douglas-fir, which had rapid failure in untreated controls. The AWPA standardized reference preservatives equivalent to UC4A retentions were chromated copper arsenate Type C (CCA-C) in three plots and creosote, alkaline copper quat – Type B (ACQ-B), or ammoniacal copper zinc arsenate (ACZA) in the remaining three plots. Stakes were rated for both decay and termite attack, but to simplify the comparison, the lower of the decay and termite ratings for each stake was used to compute an average rating for that group.

For each plot, the average rating for each Test treatment at 3, 4, 5, and 9 years was calculated as a percent of the average rating of the standard Reference preservative (Table 2). Various criteria were then applied to the year 3, 4, and 5-year data to see how well they predicted the percentage at Year 9 (Table 3). In the first analysis, a Test treatment was judged to be "satisfactory" if its average rating was at least 90% of the Reference preservative after nine years. The analysis was then repeated with a Test treatment judged satisfactory if the average rating was at least 95% of the Reference preservative after nine years. In some cases, different criteria were applied depending on the years of test data available (Table 3). These criteria were selected with the objective of minimizing the rate of both false positive (Test treatment judged acceptable but subsequently performs poorly) and false negative (Test treatment judged unacceptable but later performs well) predictions.

Table 2. Example showing how ratings of a Test formulation were compared to those of the standard Reference preservative.

	Average Stake Ratings for Treatment Group			
	Year 3	Year 4	Year 5	Year 9
Test Treatment	9.8	9.4	9.0	8.5
Reference Preservative	10.0	9.9	9.8	9.4
Percent of Reference Preservative	**98%**[a]	**95%**[a]	**92%**[a]	**90%**[b]

[a]Criteria were applied to percentages at years 3, 4, and 5 to see how well they predicted successful performance at year 9 (long term).
[b]Test treatments were considered successful if they were at least 90% (analysis 1) or 95% (analysis 2) of the Reference treatment at year 9.

Table 3. Summary of criteria applied to Test treatments after three, four, or five years of exposure. Test formulations were judged to be satisfactory if their average rating at year nine was either 90% (analysis 1) or 95% (analysis 2) of that of the reference preservative.

Criteria Applied to Test Treatments	**Years in Test Analyzed**
Average rating of Test formulation is **95%** of Reference at years in test. Only **10%** of stakes may have rating of **9** or below.	3
Average rating of Test formulation is **95%** of Reference at years in test. Only **10%** of stakes may have rating of **7** or below	4 and 5
Average rating of Test formulation is **95%** of Reference at years in test. Only **30%** of stakes may have rating of **9** or below.	4 and 5
Average rating of Test formulation is **98%** of Reference at years in test. Only **20%** of stakes may have rating of **9** or below.	3, 4 and 5
Average rating of Test formulation is **99%** of Reference at years in test. Only **10%** of stakes may have rating of **9** or below.	3, 4 and 5
Average rating of Test formulation is **99.5%** of Reference at years in test. Only **10%** of stakes may have rating of **9** or below.	4
Average rating of Test formulation is **100%** of Reference at years in test. Only **10%** of stakes may have rating of **9** or below.	3
Average rating of Test formulation is **100%** of Reference at years in test. Only **10%** of stakes may have rating of **7** or below.	4 and 5
All stakes rated **10**, no comparison to Reference.	3

RESULTS AND DISCUSSION

With only three years of data, false prediction of satisfactory performance was common (Figures 1, 2). Even when the average ratings of the Test treatments were required to be 100% of the Reference treatment, false acceptance occurred for either 21% or 32% of the treatments, depending on whether the Test treatment was required to perform 90% or 95% as well as the reference preservative after nine years. False rejections occurred in 13% or 6% of the cases, again depending on the whether 90% or 95% was used for the nine-year comparison. The rate of false acceptance was lowered to 11% and 22% when the criterion of "all 10's" was applied and this was accomplished with only a slight increase in the rate of false rejections. However, in very severe test sites, such as the tropics, a criterion of "all 10's" may be overly conservative. Use of only slightly less conservative criteria after three years (Test treatment at 95%, 98% or 99% of Reference) resulted in even higher rates of false acceptance. This finding warrants caution when applying a statistical test at three years which requires only that the Test treatment not be significantly less than the Reference treatment at the 95% confidence level.

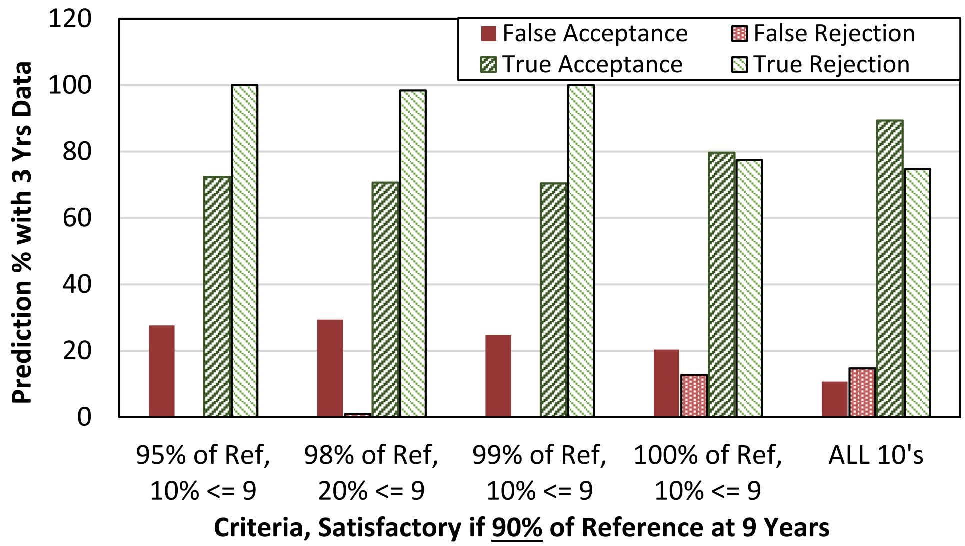

Figure 1. Percent false predictions when various criteria are applied to <u>three years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>90%</u> of the Reference preservative.

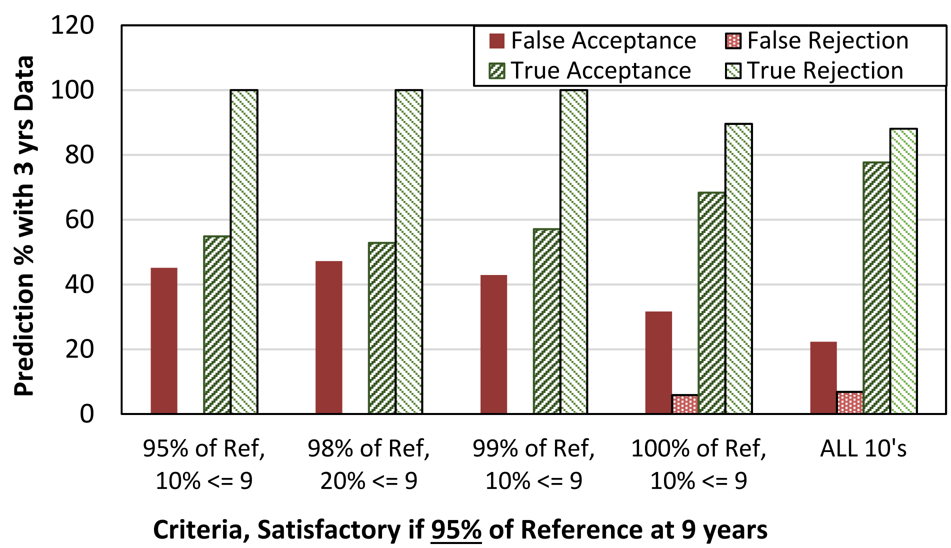

Figure 2. Percent false predictions when various criteria are applied to <u>three years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>95%</u> of the Reference preservative.

After four years of exposure, the percentage of false predictions was greatly reduced when the average ratings of Test treatments was required to be at least 100% of that of the Reference treatment (Figures 3, 4). Only 2% or 11% false acceptance occurred depending on whether the Test treatment was required to be 90% or 95% of that of the Reference treatment at year nine. In contrast, the false rejection rate was relatively high (15%) for the less stringent 90% criterion at year nine (Figure 3). As was observed with the three-year data, even slightly more lenient four-year comparison criteria (98%, 99% or 99% of Reference) caused a substantial increase in the percentage treatments with false acceptance (Figures 3, 4).

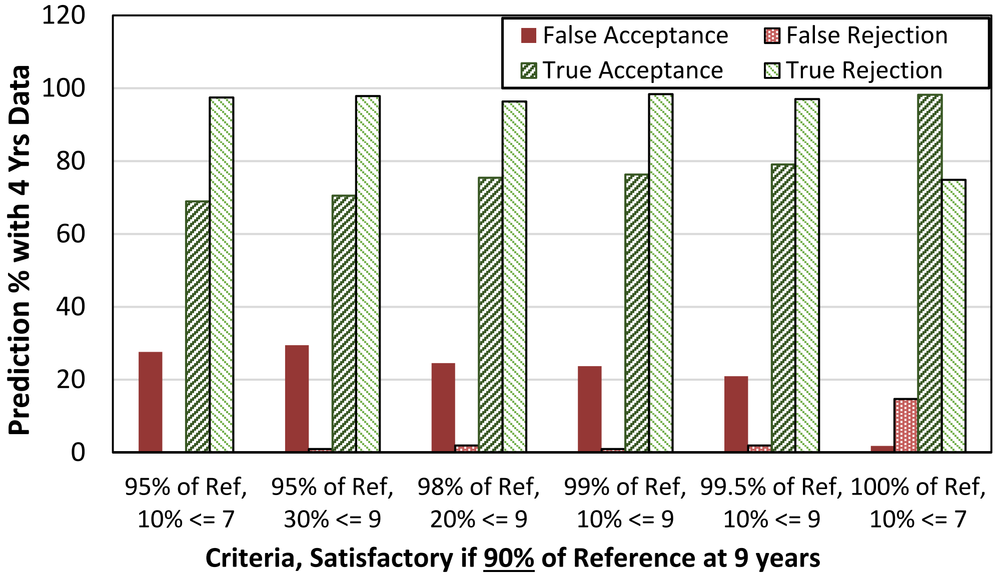

Figure 3. Percent false predictions when various criteria are applied to <u>four years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>90%</u> of the Reference preservative.

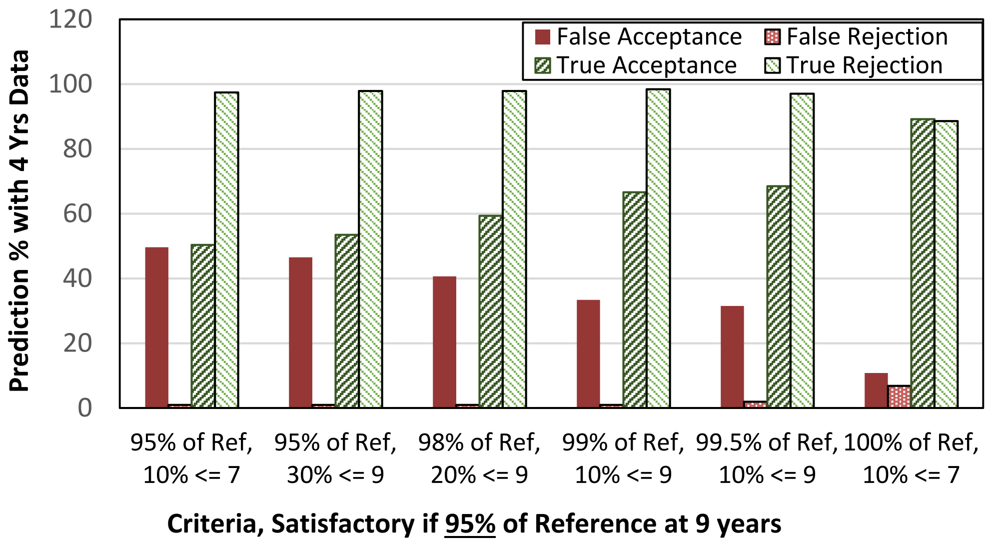

Figure 4. Percent false predictions when various criteria are applied to <u>four years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>95%</u> of the Reference preservative.

With five years of data the accuracy of prediction appeared to be improved with some criteria but not others (Figures 5, 6). When the average ratings of Test treatments were required to be at least 100% of that of the Reference treatment, the five-year data resulted in a greater percentage of false acceptance (and false rejections) than the four-year data (but fewer than the three-year data). Compared to the three and five year analyses, it is possible that the four-year data presents an overly optimistic picture of our ability to predict longer term performance (i.e. at four years). Or, that the 100% of Reference criterion with five years of data may have overestimated false predictions, since a lower percentage of false acceptance was observed with the more lenient 99% of Reference criterion. It may seem counterintuitive that the rate of false acceptance would be less with the more lenient criteria, but this can occur if more treatment groups initially meet the more lenient criteria. With the five-year data, it appears that requiring the test preservative to be within 99% that of the Reference minimizes both false acceptance and false rejection.

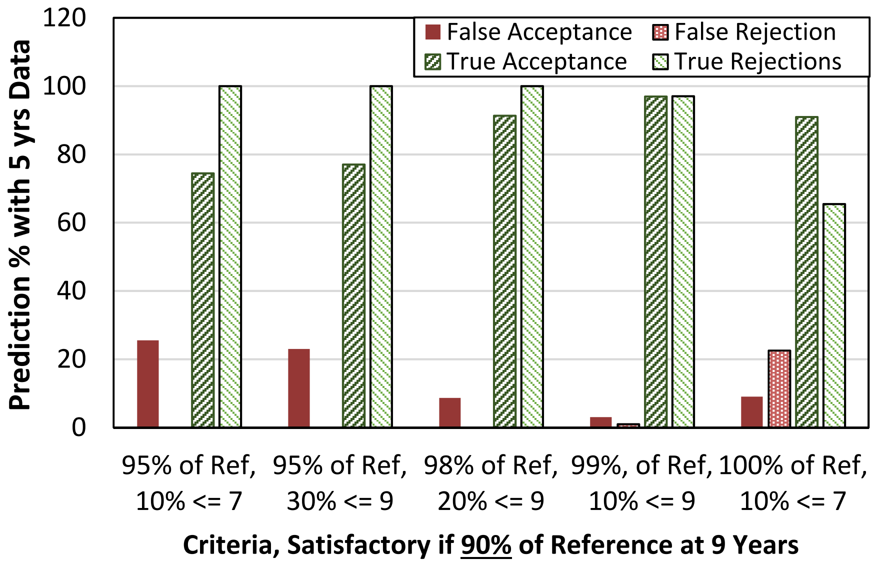

Figure 5. Percent false predictions when various criteria are applied to <u>five years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>90%</u> of the Reference preservative.

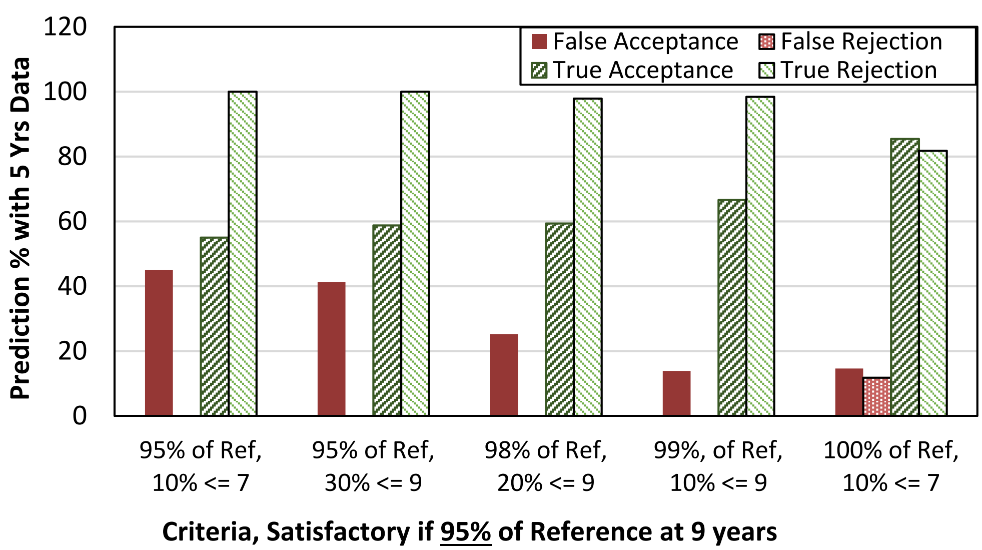

Figure 6. Percent false predictions when various criteria are applied to <u>five years</u> of 19 by 19 mm stake data. Test treatment judged to be satisfactory if average rating at year nine is at least <u>95%</u> of the Reference preservative.

In general, these analyses reveal the challenges in using short term, small stake data to predict the future performance of commodity size material over the longer term. Use of year three stake data appears particularly problematic because of the relatively high potential for accepting treatments that subsequently do not perform as expected. If three years of stake data are used, it appears prudent to use strict comparison criteria, such as requiring the average ratings of the Test preservative be at least 100% of that of the Reference preservative. Use of years four and five data is somewhat more promising, but the potential for either false acceptance or false rejection remains. Thus, although some error in predicting future performance is probably unavoidable, this data highlights the importance of continuing stake tests and periodically bringing additional data forward to AWPA committees for review.

It is also important to note that the findings of this analysis are a function of the test site, methods used, and rating system. The relationship between ratings of Test treatments and Reference preservatives is likely to vary between test sites as a result of differences in soil properties, climate, and types of organisms present. In addition, this analysis did not include the more recent “9.5” rating, which may help to improve predictions by creating more differentiation in the higher ratings. With relatively few years of data, even moderately effective treatments may not have stake ratings below “9” and the differentiation between treatments is largely based on ratings in the 9-10 range. The number of replicate stakes installed for each treatment group may also affect variability in the average ratings and thus the comparison between Test and Reference preservatives. When data packets are brought to AWPA for review 10 replicates is common, but the plots in this analysis all had at least 20 replicates per treatment group. The use of more replicates improves prediction by lessening the effect of one or two low or high ratings and even small variations in the average rating of the Reference preservative can have a relatively large effect on the comparison to the Test treatment. Because multiple Test treatments are typically compared to a single Reference, it may be worthwhile to increase the number of replicates used at least for the Reference treatment. Consistency of ratings from year to year also may be a factor. FPL uses multiple raters and stakes may be rated by different people from one year to the next. Stake ratings are inherently subjective and it is likely that some raters tend to give higher or lower ratings than others. Thus, it is not highly unusual to have a stake which is rated a “9” one year be rated a “10” the following year. Although use of multiple raters helps to minimize overall bias, it can contribute to variability in ratings between years.

The use of statistical methods, rather than averages, to compare Test and Reference treatments warrants further discussion. To this point no statistical test has been specified for this purpose and thus a statistical method was not included in this analysis. Selection of an appropriate critical test may be challenging, as ratings are ordinal data and not normally distributed. There are rank-sum statistical tests that do not assume a normal distribution, but these tests can be less powerful when there are a substantial number of equal value responses between the groups evaluated (Lebow and Kirker, 2014). Equal value responses would be typical in the early years of stake tests because most stakes in both Test and Reference groups are likely to be rated either “10”, “9.5”, or “9”. Rank sum tests also have some characteristics that could be concerning in evaluating stake ratings. For example, a test group with two stakes rated “0” and eight stakes rated “10” would have the same rank-sum as a group with two stakes rated “9.5” and eight stakes rated “10”.

Because the detection of slight differences in ratings is needed to identify weaker treatments with few years of test data, it may be necessary to employ more complex statistical models that consider ratings over time rather than comparing individual years. Lebow and Kirker (2014) had some success modeling the probabilities of ratings as a function of treatment and time. More specifically, they discuss a generalized linear mixed model with cumulative logit link modeled as linear function of treatment and slope parameters for time along with a random effect to account for the use of repeat measurements. The example below 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 measurement:

$$\text{logit}[P(Y_{ijkl} \leq m \mid X_i, C_j, t_{ijk}, Z_l)] = \beta_m + \beta_{1i} X_i + \beta_{2i} X_i \ln(t_{ijk}) + b_l Z_l$$

Where:
Y_{ijkl}=stake rating *ijkl*
m=10,9,8,7,6,4 (visual rating)
X_i=treatment group *i* (A,B,C,U)
t_{ijk}=exposure time *k* (1,2,3,4,5,7,9,11 years)
b_l=random effect for stake *l*
Z_l=stake *l* from treatment group *ij*
β=model parameters

Mean comparisons of treatments can be made using comparisons of log odds ratios and constructed using various years of data. However, even with these more sophisticated models it may be difficult to detect differences between Test and Reference treatment groups with only three years of data (Lebow and Kirker, 2014).

CONCLUSIONS

This analysis indicates that predicting future durability is difficult with only three years of 19 by 19 mm stake data. Even when strict performance criteria were applied, 20 – 30% of treatments that were judged as acceptable performed poorly over the longer term. It is important to note that the data analyzed in this paper did not include the "9.5" rating which may help to provide distinction between Test treatments and Reference preservatives in the early years of evaluation. However, caution is still warranted because even 11 – 22% of Test treatments, rated as all "10s" after three years, performed poorly at nine years. It appears that even very slight indications of vulnerability early in testing are strong indicators of poor future performance and the challenge lies in separating these slight differences from the variability that occurs in stake rating. Predictions might be improved through application of more complex statistical analysis tools and use of more replicate stakes for each treatment group. Increasing replication for the Reference preservative may be especially beneficial because it serves as the basis for comparison for multiple Test treatments.

REFERENCES

1. AWPA. 2003. E7-01. Standard Method of Evaluating Wood Preservatives by Field Tests with Stakes. Book of Standards. American Wood Preservers' Association, Selma, AL. 522p.
2. AWPA. 2016a. Guidance Document A: Data Requirements for Listing Preservatives in AWPA Standards. Book of Standards. American Wood Protection Association, Birmingham AL. 628p.
3. AWPA. 2016b. Minutes, Committee P6, Methods for the Evaluation of Wood Protectants. Fall Technical Meeting, September 12 -15, San Antonio, Texas. http://www.awpa.com/members/committees/P6/P6.aspx
4. AWPA. 2016c. Commodity Specification D: Poles. Book of Standards. American Wood Protection Association, Birmingham AL. 628p.
5. AWPA. 2016d. E7-15: Standard Field Test for Evaluation of Wood Preservatives to be used in Ground Contact (UC4A, UC4B, UC4C); Stake Test. Book of Standards. American Wood Protection Association, Birmingham AL. 628p.
6. Hartford, W.H. 1972. The interpretation of test plot data. Proceedings of the 68th Annual Meeting of the American Wood Preservers' Association. Washington, DC. 68:67-82.
7. Hartford, W.H.; Colley, R.H. 1984. The rationale of preservative evaluation by field testing and mathematical modeling. Proceedings of the 80th Annual Meeting of the American Wood Preservers' Association. Stevensville, MD. 80:257-269.
8. Lebow, P., Kirker, G. 2014. Rate My Stake: Interpretation of Ordinal Stake Ratings. Proceedings of the 108th Annual Meeting of the American Wood Protection Association. Birmingham, AL. 108:156-164.
9. Lebow, S., Hatfield, C. and J. Evans. 2010. A comparison of the durability of two stake sizes in southern Mississippi. Proceedings of the 105th Annual Meeting of the American Wood Protection Association. Birmingham, AL. 105:66-72.
10. Lebow, S., Woodward, B., and Lebow, P. 2009. Observations on the predictive value of short-term stake tests. Proceedings of the 104th Annual Meeting of the American Wood Protection Association. Birmingham, AL. 104:85-88.
11. Lebow, S.T., Woodward, B.M., Lebow, P.K. and C. Clausen. 2008. The need for performance criteria in evaluating the durability of wood products. Review of the Faculty of Forestry, University of Istanbul. 58(2):29-42.
12. Link, C. L.; DeGroot, R.C. 1990. Predicting Effectiveness of wood preservatives from small sample field trials. Wood and Fiber Science, 22(1):92-108.
13. Link, C. L.; DeGroot, R.C. 1989. Statistical Issues in Evaluation of Stake Tests. Proceedings of the 85th Annual Meeting of the American Wood Preservers' Association. Birmingham, AL. 85:170-185
14. Morris, P.I. 1998. Beyond the log probability model. Proceedings of the 94th Annual Meeting of the American Wood Preservers' Association. Granbury, TX. 94:267-273.
15. Morris, P.I.; Cook, J.A. 1995. Predicting preservative performance by field testing. Wood preservation in the '90s and beyond. Proc. No. 7308. Madison, WI: Forest Products Society. pp.116-121.
16. 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.

PROCEEDINGS

One Hundred Thirteenth Annual Meeting

of the

AMERICAN
WOOD PROTECTION
ASSOCIATION

Encore at Wynn Las Vegas
Las Vegas, Nevada
April 9-11, 2017

VOLUME 113

AMERICAN WOOD PROTECTION ASSOCIATION
P O BOX 361784 • BIRMINGHAM, ALABAMA 35236-1784 • USA

ISSN 0066-1198

AMERICAN WOOD PROTECTION ASSOCIATION
P.O. BOX 361784
BIRMINGHAM, ALABAMA 35236-1784
USA

Colin McCown, *Editor*
Todd Gothard and Beth Staula, *Assistant Editors*