

1989

Statistical Issues in Evaluation of Stake Tests ¹

Carol L. Link
Mathematical Statistician

Rodney C. De Groot
Research Plant Pathologist
U.S. Department of Agriculture, Forest Service, Forest Products Laboratory Madison, Wisconsin

This paper explores some statistical issues in the evaluation of stake tests: accounting for data variability, choice of rating scale for evaluating termite and decay damage, definition of failure, and replicability of experimental results. Data used to illustrate these issues come from 50 years of field tests conducted by the Forest Products Laboratory. These issues need to be considered as researchers are trying to improve their estimates of future benefits that could be expected from new preservatives.

Keywords: variability, rating scale, failure time.

Introduction

Field trials with preservative-treated stakes (graveyard tests) are commonly used to determine the efficacy of a given wood preservative and retention level. Wood stakes are treated to different retentions of each candidate preservative, and usually (if following ASTM (1987) or AWP (1988) standards), to different retentions of a reference preservative. Within the experimental field plot, the treated stakes are inserted vertically into the soil usually to a depth of one half their length. The amount of subsequent decay and termite attack is visually estimated at regular intervals. Therefore, field plots are often established with the intent of comparing one preservative and retention to another. One should use statistical concepts for this comparison. Because of the nature of wood preser-

vation tests, several issues must be addressed when comparing wood preservation experiments statistically. These can be summarized as data variability, choice of rating scale, definition of failure, and replicability of the field trial results.

To illustrate the concepts in this paper, we use data from the Forest Products Laboratory's (FPL) field plot in the Harrison Experimental Forest, Saucier, Mississippi. For data variability, rating scale, and definition of failure, we use groups of ten replicate southern yellow pine nominal 2 by 4 by 18 inch (5 by 10 by 46 mm) stakes that had an average time to failure because of decay and/or termites of between 14 and 15 years. The plots, preservatives, and retentions involved are presented in Table 1. For replicability of field tests, we use groups of ten replicate yellow southern pine nominal 2 by 4 by 18 inch (5 by 10 by 46 mm) stakes that are treated with coal tar creosote, zinc chloride, and pentachlorophenol.

Issues

Data Variability

Most stake test results are presented as averages, such as average time to failure and average rating at a given time. While averages are useful to examine trends in the data, they do not address data variability. For example, the average life (time to failure) of stakes listed in Table 1 is about 14 years (Gjovik and Gutzmer, 1986). However, individual lifetimes range from 2 to 26 years (Fig. 1).

Box plots (Velleman and Hoaglin, 1981) (Fig. 2) are a useful tool for summarizing data variability. A box surrounds the center 50 percent of the data from

¹This publication reports research involving pesticides. It does not contain recommendations for their use, nor does it imply that the uses discussed here have been registered. All uses of pesticides must be registered by appropriate State and/or Federal agencies before they can be recommended.

CAUTION: Pesticides can be injurious to humans, domestic animals, desirable plants, and fish or other wildlife-if they are not handled or applied properly. Use all pesticides selectively and carefully. Follow recommended practices for the disposal of surplus pesticides and pesticide containers.

²The Forest Products Laboratory is maintained in cooperation with the University of Wisconsin. This article was written and prepared by U.S. Government employees on official time, and it is therefore in the public domain and not subject to copyright.

Table 1. Description of test stakes.

Group	Plot	Preservative	Preservative retention pcf (kg/m ³)	Average stake life (year)	
				Time to 0 ^a	Time to 7 ^b
1	20	Catalytic gas-base oil (West Coast)	8.0 (128)	14.6	5.8
2	20	No. 2 fuel oil (Mid United States) with 5% pentachlorophenol	4.0 (64)	14.9	4.3
3	20	Catalytic gas-base oil (West Coast) with copper naphthenate (0.5% copper metal)	4.2 (67)	14.3	7.4
4	20	Coal-tar creosote	4.1 (66)	14.2	5.4
5	20	Coal-tar creosote, 25%, and catalytic gas-base oil (West Coast) with copper naphthenate (0.75% copper metal), 75% by volume	4.2 (67)	14.6	7.0
6	2	Chromated zinc chloride	0.49 (7.8)	14.2	4.8
7	4	Zinc chloride	0.50 (8.0)	14.2	2.4
8	4	Zinc chloride	1.02 (16.3)	14.4	3.2
9	36	Rosin oil and no. 2 fuel oil (1:7) with 2.98% pentachlorophenol	8.0 (128)	14.8	4.4
10	38	Pentachlorophenol, 5% in light aromatic solvent with vapor cleaning	4.5 (72)	14.2	1.7
11	38	Pentachlorophenol, 5% in light aromatic solvent with no cleaning	4.6 (74)	14.1	3.5
12	38	Copper naphthenate, 0.59% copper in light aromatic solvent with steaming	4.4 (70)	14.3	3.1

^aTime to failure.

^bTime when stake first reaches the ASTM rating of 7.

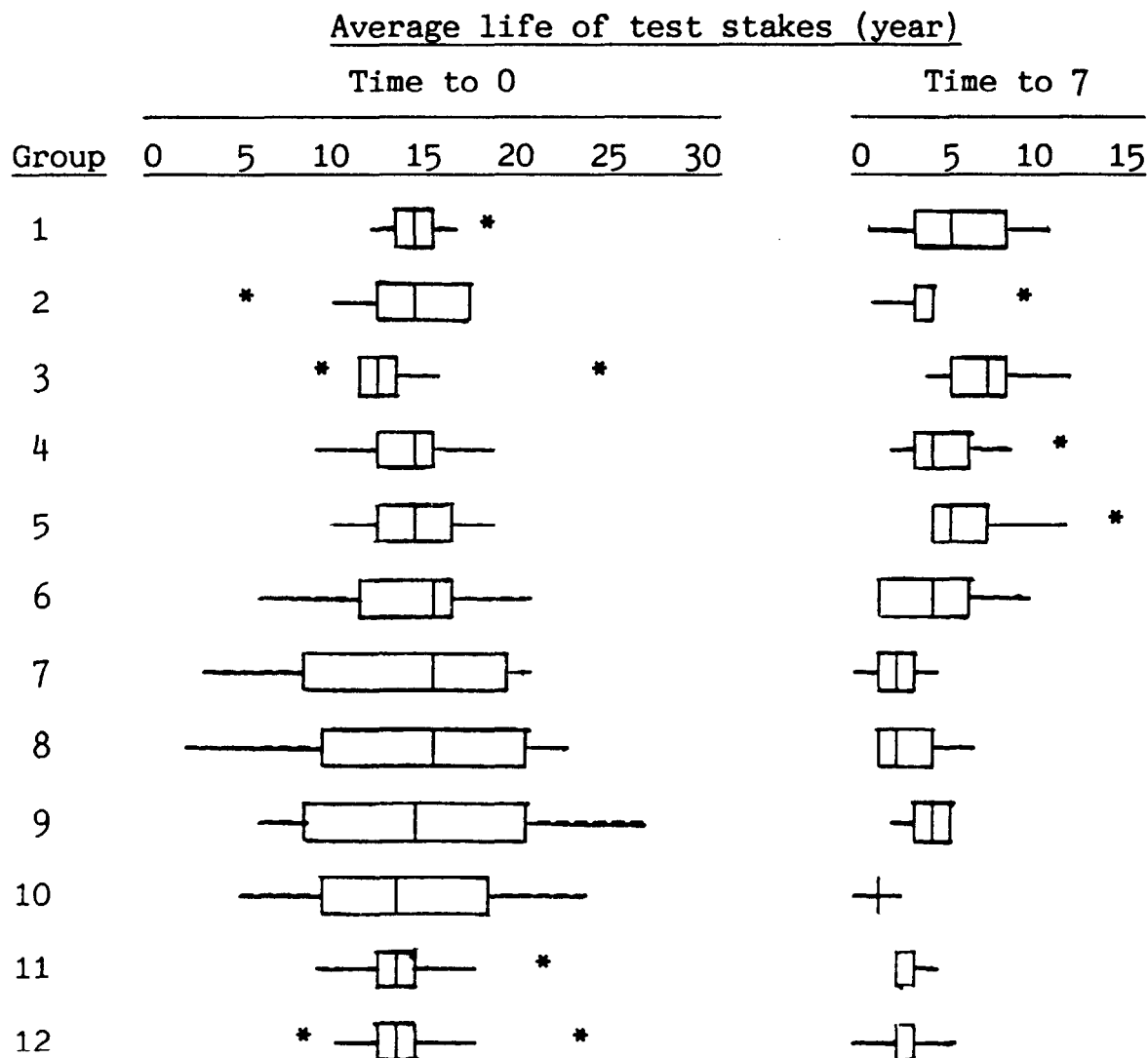


Figure 2. Box plots for groups of test stakes described in Table 1.

the first quartile to the third quartile. A vertical line in the box represents the median (second quartile). Whiskers are drawn to the minimum and maximum data points, except in the presence of outliers. Outliers (asterisk) are points that are far away from the first and third quartiles compared to the variation of the rest of the sample as measured by the interquartile range (third quartile to first quartile). Where only one vertical line occurs instead of a box, the first, second, and third quartiles are the same. Where no vertical line exists within the box, the median is the same as either the first or third quartile.

Data variability is important if early failures are a concern and if preservatives are to be compared statistically.

Rating Scale

Stakes in the field are periodically inspected and rated for damage caused by decay and termites (where applicable). These ratings may then be combined to give a single rating. The choice of rating scale and the combination of rating scales may influence the comparison of one preservative to another. Some rating scales in current use are described in Table 2.

In addition to the ratings listed in Table 2, the ASTM rating scale provides a method for obtaining a combined decay and termite rating, the stake rating: Stake rating equals minimum (decay rating, termite rating). For example, if the decay rating is 9 and termite rating is 7, the stake rating is 7.

Table 2. Rating scales for decay and termite damage to field stakes.

Type of damage	Rating scale			
	ASTM D 1758	FPL ^a	Levy & Dickinson	Borsholt
Decay				
None (sound)	10	1	0	0
Trace	9	2	1	25
Moderate	7	3	2	50
Heavy	4	4	3	75
Complete ^b	0	5	4	100
Termite				
None (sound)	10	A	--	--
Trace	9	B	--	--
Moderate	7	C	--	--
Heavy	4	D	--	--
Complete ^b	0	E	--	--

^aGjovik and Gutzmer, 1986.

^bFailure.

The scales described in Table 2 use similar but not identical terms to describe decay and/or termite damage. The ASTM rating scale uses the terms listed in Table 2. In the FPL scale, the levels of decay are described as follows: 1, no decay; 2, slightly soft or suspicious; 3, partial or limited decay; 4, severe decay; and 5, removed because of decay. The FPL rating scale terms for termite damage are A, no attack; B, nibbles or trails; C, limited attack; D, heavy attack; and E, removed because of termite attack.

During FPL inspections, stakes are often rated at intervals between rating levels. The addition of a

plus sign to a rating level indicates decay or attack in excess of that rating level; a minus sign indicates decay or attack less than that rating level.

De Groot (1983) suggested a numerical conversion of the FPL ratings to a scale similar to that used in the ASTM D 1758 (ASTM, 1987) standard as follows:

Numerical conversion:	10	9.5	9	8.5	8	7	6	5	4	2	0
Decay rating:	1	2-	2	2+	3-	3	3+	4-	4	4+	5
Termite rating:	A	B-	B	B+	C-	C	C+	D-	D	D+	E

De Groot also suggested a combined decay and termite rating, which is the product of the decay and termite numerical ratings divided by 10.

The Levy and Dickinson (1980) rating scale (Table 2) does not provide for termite attack and results are not given where termite attack occurs. Terms used to describe decay at each rating level are as follows: 0, sound; 1, slight and superficial decay; 2, evident but moderate decay; 3, severe decay; and 4, failure (almost complete loss of strength). The authors state:

As they stand the five ratings are not evenly spaced on the scale, so that it is not possible that rating 1 represents 25% decay, rating 2 = 50% decay, rating 3 = 75% decay and rating 4 = 100% decay. Only rating 0 can be said to show 0% decay but even rating 4 is not necessarily 100% decay since it may only represent a much smaller proportion of the percentage loss of strength of the stake.

The Borsholt (1979) scale describes decay rating as 0 = sound (no attack); 25, slight attack; 50, moderate attack; 75, severe attack; and 100, failure in bending apparatus. There is no provision for rating termite attack.

The Conradie and Pizzi (1984) scale rates decay and termite attack separately. Rating is expressed as a percentage from 0 (sound stake) to 100 (failed stake). The average grade is computed according to the formula

$$\text{Average grade} = D + (aR)/50$$

where D is percentage of stakes destroyed (stake considered destroyed if rating is > 50 percent), a is sum of stake ratings (for stakes with rating < 50 percent), and R = 100 - D, which is the percentage of stakes remaining. Conradie and Pizzi do not combine decay and termite ratings.

Because the FPL data do not readily translate into the Conradie and Pizzi scale, this scale is not used in the remainder of the paper.

In field sites where both decay fungi and termites are present, stakes are rated for each of these hazards separately. Some have argued that a stake should be rated for only one hazard at a time, and only one hazard should be used for comparison purposes. However, decay and termite damage are unlikely to be statistically independent phenomena, and some rating system must be devised to incorporate both decay and termite damage where both occur in a field plot. Decay fungi and termites need to be considered as competing risks for any given stake. It might be preferable to develop one rating scale for the overall condition of the stake, with an indication of the cause for deterioration.

With the exception of the average rating used by

Conradie and Pizzi, each system uses a five-point scale with similar descriptions of the five levels of decay and/or termite attack. Therefore, one could easily change from one rating scale to another. The scales used by Levy & Dickinson and Borsholt are essentially the same scale with 0 to 4 replaced by 0 to 100.

Stake ratings are on an ordinal rather than numerical scale; that is, the ratings are ordered from sound to destroyed. However, the distance between any two ratings does not necessarily imply that a precise amount of damage has occurred; for example, the amount of damage that occurs from a rating of 10 to 9 on the ASTM scale is not necessarily one half the damage that occurs from 9 to 7 on that same scale. Means are not appropriate statistics for data on an ordinal scale. Therefore, means should not be used to obtain an average rating or to compare the average rating of one preservative to another.

As an example, let us consider groups 1 and 3 from Table 1. These stakes were installed in April 1948. Let us compare the average ratings from the December 1952 inspection. The FPL ratings for that inspection were as follows:

Group 1: 1B 3A 1B 1B 3C 2A 2B 3B 1B 3C
Group 3: 1A 1A 1A 3B 1A 1B 1A 1A 1B 2B

To compare the linear rating system used by Levy & Dickinson and Borsholt to that suggested by De Groot and the ASTM system, a method for combining decay and termite ratings is needed. We used a procedure that is comparable to the ASTM system; the combined linear rating is the maximum (decay rating, termite rating). Maximums are used instead of minimums because values on the ASTM rating scale decrease with damage whereas those on the linear rating scale increase with damage.

The usual procedure for comparing the means of two groups, assuming that the data are on a numerical scale, is the two-sample t-test. This is a parametric procedure since there is an underlying assumption that the data come from a normal distribution. The tested null hypothesis is that the means of the two groups are equal. A p value is used to accept or reject the null hypothesis. The p value is the probability that if the null hypothesis is true (in this case, that the sample means are equal), one would observe the given difference of means or a larger difference.

If the two-sample t-test were mistakenly used to compare the mean ratings of groups 1 and 3, the result would be the p values listed in Table 3. The variability of the p values, from using different scales, is typical of the problems associated with using a test procedure that assumes numerical data when only ordinal data are present. Depending upon the cho-

Table 3. Statistical significance of a comparison of groups 1 and 3 using different tests.

Statistical test	Rating scale	p value
Two-sample t-test	ASTM combined rating	0.0326
	De Groot's multiplicative rating	0.0260
	Linear rating ^a	0.0044
Mann-Whitney test	ASTM combined rating	0.0140
	De Groot's multiplicative rating	0.0173
	Linear rating	0.0140

^aUsed by Levy & Dickinson and Borsholt.

sen significance level, the equality of the mean December 1952 rating of groups 1 and 3 could be interpreted in different ways. For example, at a 1 percent level of significance, the null hypothesis of equal means could not be rejected if the ASTM or De Groot's multiplicative rating were used. However, at the same level of significance, the null hypothesis of equal means could be rejected if the linear rating scale used by Levy & Dickinson and Borsholt were used.

As indicated previously, stake ratings are on an ordinal rather than a numerical scale, and it is therefore inappropriate to compare group means. Instead, one should compare group medians. The correct test for equality of medians is the Mann-Whitney test, which is a nonparametric test; that is, it does not assume that the data come from any specified distribution. Table 3 shows the p values resulting from the Mann-Whitney test of the equality of group 1 and 3 medians. Note that the p values for the ASTM combined rating and the linear rating used by Levy & Dickinson and Borsholt must be the same because the order of the stake ratings is the same under both systems. The Mann-Whitney test also produces p values that are much more comparable, indicating that the choice of rating scale may not be crucial.

Because the rating scales are subjective evaluations of a stake's condition at one point in time, the

rating of the stake may or may not be monotone over time. A monotone rating indicates that the condition of the stake stays the same or deteriorates over time but never improves. However, the decay rating of a stake may indeed improve over time, if for example, termites eat all the decayed wood. If inspectors change, the rating of a stake might also improve because of different interpretations of the subjective scale. In addition, variations in the subjective rating may make the condition of a stake appear to have improved. Figures 3, 4, and 5 show the rating of three different stakes over time. The dashed line is the combined decay and termite ASTM rating and the solid line is De Groot's multiplicative combined rating.

Definition of Failure

Several definitions of the failure time of a stake have appeared in the literature. Hartford (1972) and Colley (1970) argued for time to an ASTM rating of 7. However, at this point the stake is likely to still have considerable structural integrity. An ASTM rating of 4 could be argued, if this is the point in the lifetime of the stake when the stake may be replaced. However, using either of these definitions can lead to difficulty in estimating the failure time because stake ratings are not always monotone over time. The series of FPL progress reports (such as Gjovik

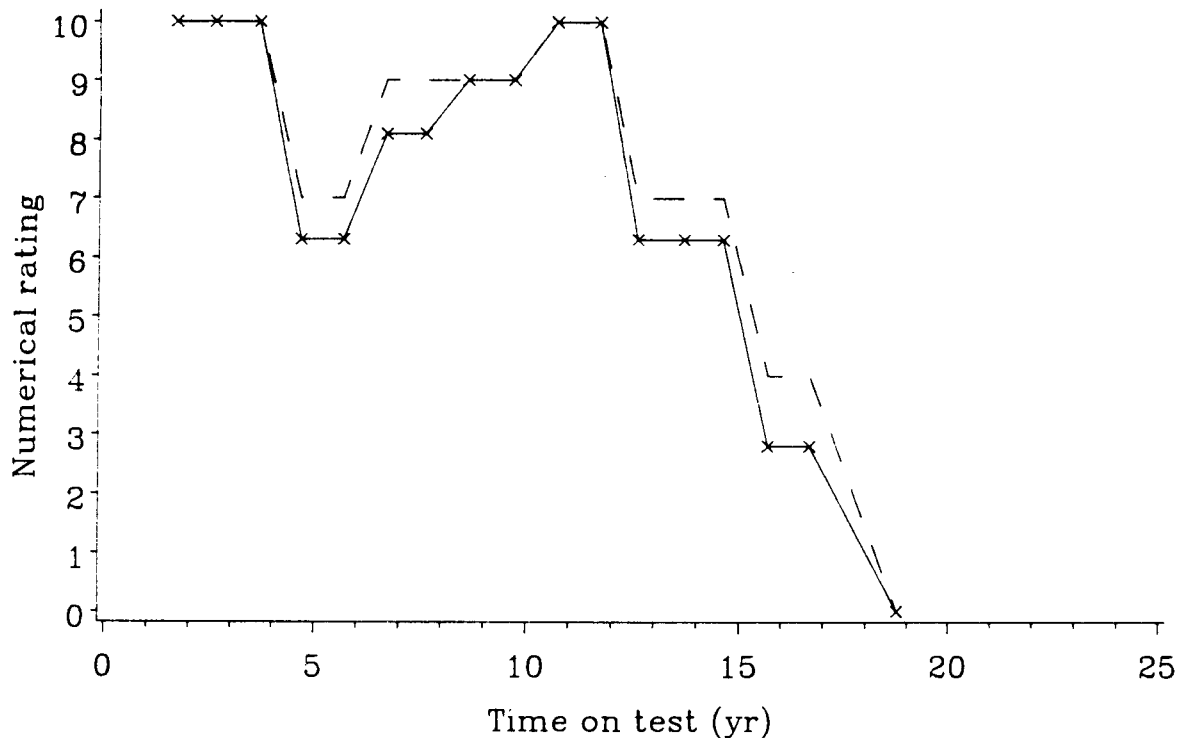


Figure 3. Rating over time of stake 374, plot 20, treated with coal-tar creosote, 25%, and catalytic gas-base oil (West Coast) with copper naphthenate (0.75% copper metal), 75% by volume to a retention of 4.1 pcf (66 kg/m³). Solid line, De Groot's multiplicative scale; dashed line combined ASTM scale. X indicates inspection dates.

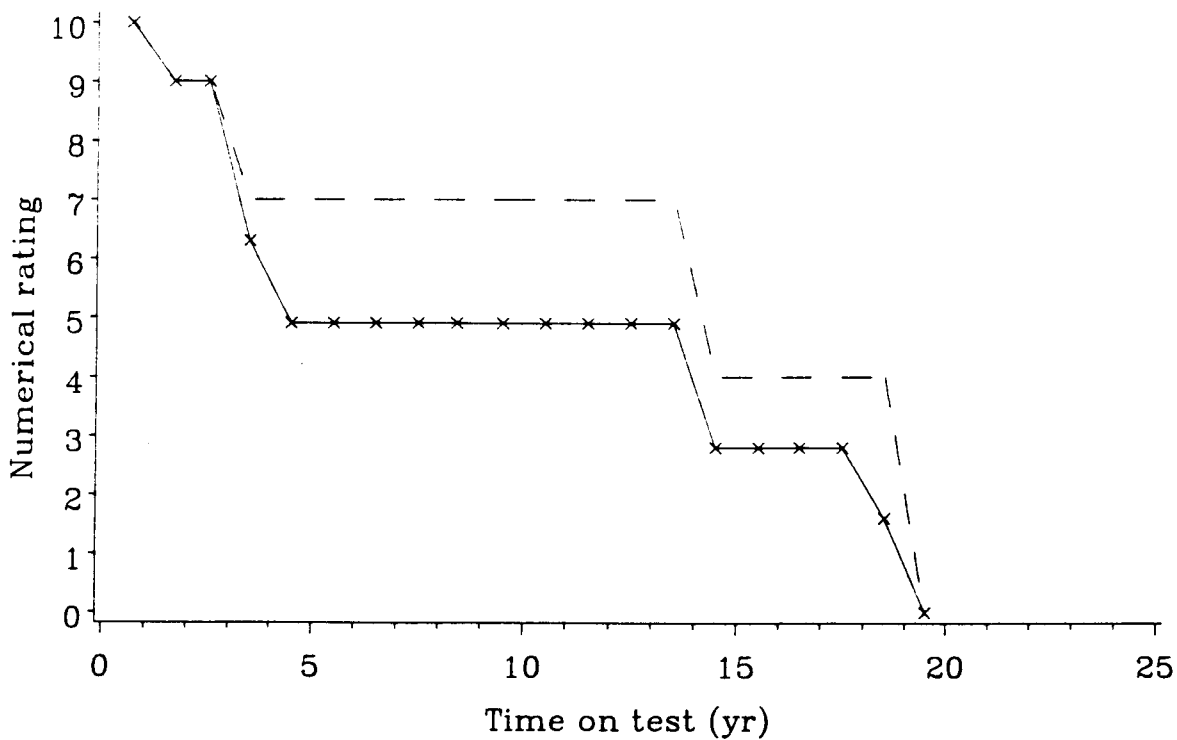


Figure 4. Rating over time of stake 81, plot 4, treated with zinc chloride to a retention of 0.5 pcf (8 kg/m³). Solid line, De Groot's multiplicative scale; dashed line, combined ASTM scale. X indicates inspection dates.

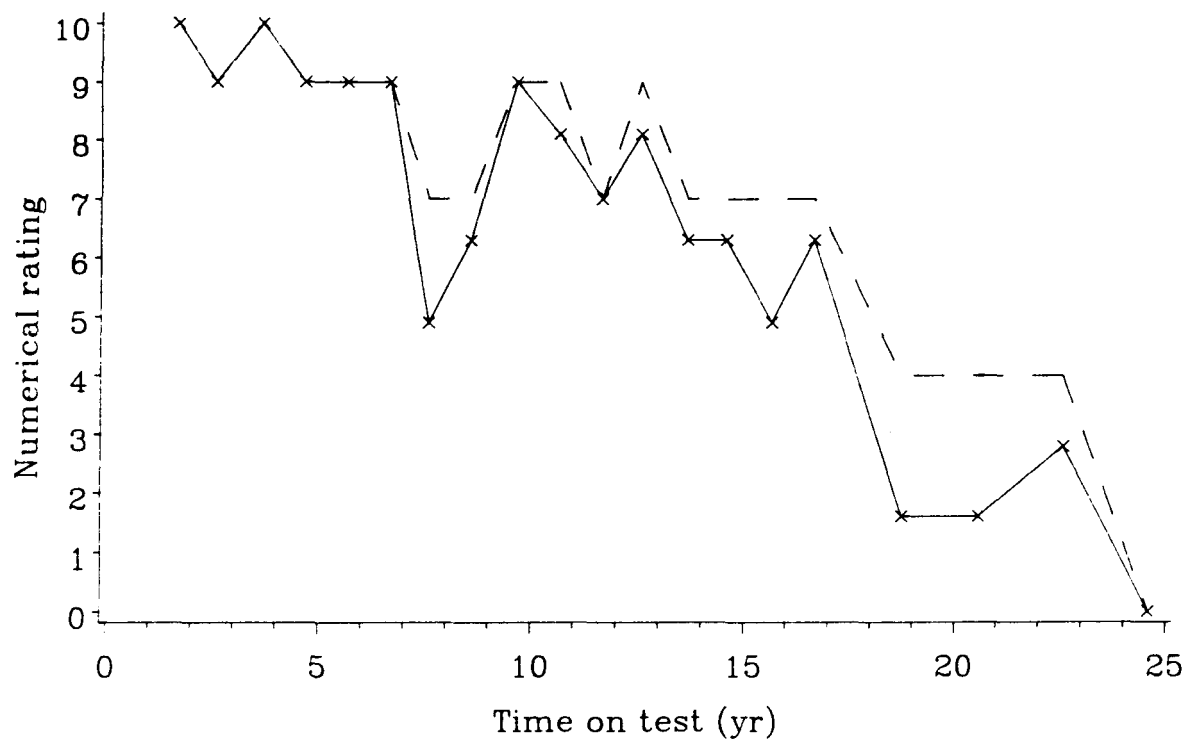


Figure 5. Rating over time of stake 326, plot 20, treated with catalytic gas-base oil (West Coast) with copper naphthenate (0.5% copper metal) to a retention of 4.2 pcf (67 kg/m³). Solid line, (De Groot's multiplicative scale; dashed line, combined ASTM scale. X indicates inspection dates.

and Gutzmer, 1986) have used time to 0 as the definition of failure time. This eliminates the potential difficulty of how to rate a stake whose condition improves over time. Comparisons of preservatives may depend upon the definition of failure time being used.

Because stake ratings are not necessarily monotone over time, we need some specification of the time when a stake reaches an ASTM rating of 7. For the purposes of this paper, time to 7 designates the time when the stake first reaches an ASTM rating of 7, even though the rating may be higher later. This is essentially equivalent to the choice of failure time when the stake first reaches 0.

The groups of stakes listed in Table 1 were chosen because they had mean time to 0 of 14 to 15 years. It is appropriate to consider mean or median failure times because failure times are on a numerical scale, as opposed to medians for the stake ratings, which are on an ordinal scale. We have shown (Link and De Groot, submitted) that the mean and median failure times for groups of ten replicate stakes differ by at most one year 80 percent of the time and by at most two years 95 percent of the time. To compare the average failure times for the 12 groups, we could

use means or medians. Means were chosen because that is the traditional method; either means or medians produce comparable results with these data.

Using analysis of variance, none of the mean time to 0 of the groups can be considered statistically different. In fact, because of the variability of failure times for these data, the means of two groups must differ by at least 4.1 years to be statistically different at the 5 percent level. This is the difference that could be considered statistically significant if only two groups were compared. Simultaneous comparison of 12 groups would require a difference of at least 6.9 years for statistical significance at the 5 percent level because a multiple comparison procedure is needed for comparing means of many groups. We used the Tukey multiple comparison procedure. (See Dunn and Clark (1974) for more information on multiple comparison procedures.)

If time to 7 is used as the definition of failure, the 12 groups cannot be considered to have equal mean failure times. If two groups were compared, the minimum statistically significant difference would be 1.9 years. If 12 groups were compared simultaneously, the minimum statistically significant difference would be 3.2 years. Using the Tukey multiple com-

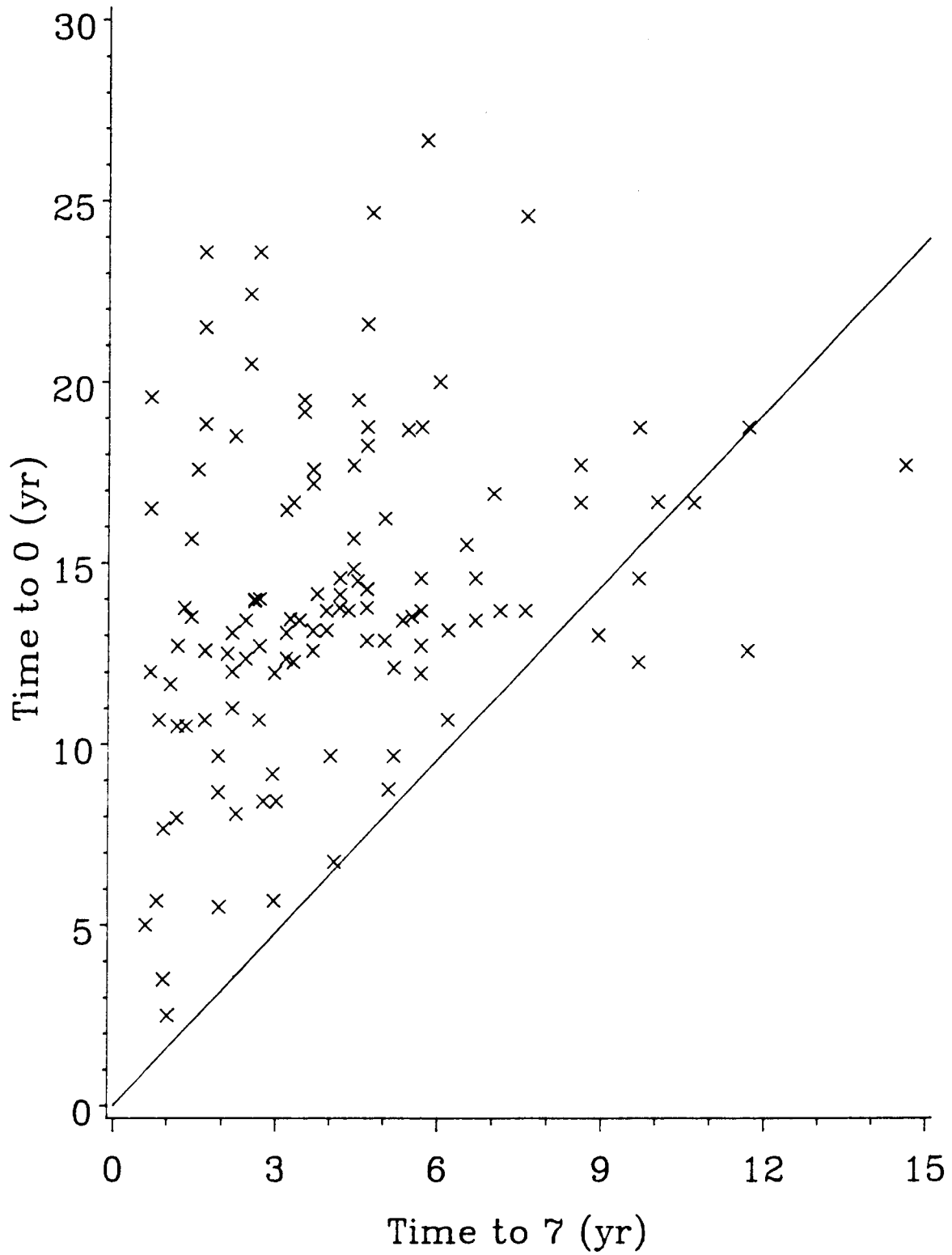


Figure 6. Time to 0 compared to time to 7 for stakes in Table 1. Line indicates where time to 7 equals 5/8 time to 0.

parison procedure for 12 groups (Table 1) results in the following comparisons (groups listed from low to high means)

10 7 12 8 11 2 9 6 4 1 5 3

Groups sharing a common underline cannot be considered statistically different at the 5 percent level.

Time to 0 and time to 7 of stakes in all 12 groups are compared in Figure 6. There appears to be little relationship between time to 0 and time to 7. Hartford (1972) considered failure time as time to 7, and calculated this value as $\frac{7}{8}$ of the reported lifetime in the FPL reports that use time to 0. This value is represented by the solid line on Figure 6. The hy-

Table 4. Average life of stakes treated with coal tar creosote. ^{a,b}

Plot	Average stake life (year)		Preservative retention pcf (kg/m ³)
	Time to 0	Time to 7	
4	17.8	6.1	4.2 (67)
5	21.3	5.3	4.6 (74)
6	24.9	6.3	3.3 (53)
20	14.2	5.4	4.1 (66)

^aFPL field trials.

^bApproximately 4 percent retention.

Table 5. Characteristics of coal-tar creosotes used in FPL plots. ^a

Creosote characteristic ^b	Plot 4	Plots 5&6 ^c	Plot 20
Specific gravity 38/15.5 °C	1.064	1.082	1.107
Benzol insoluble	0.26	0.34	0.33
Coke residue	0.85	0.92	1.46
Moisture	0.3	0.4	3.4
Distilling up to 210 °C	0.5	0.5	1.2
Distilling up to 235 °C	5.8	4.6	3.6
Distilling up to 270 °C	17.7	15.9	16.8
Distilling up to 315 °C	45.3	44.1	35.2
Distilling up to 355 °C	74.3	73.1	57.4
Distilling above 355 °C	25.3	26.4	41.9
Distillation loss	0.4	0.5	0.7

^aPlots 4,5,6, and 20. See Table 4.

^bExcept for specific gravity, all characteristics given as percentages.

^cStakes from plots 5 and 6 were treated in the same charge.

pothetical relationship that Hartford proposed does not appear to hold in the case of these data.

Replicability

There are few pure replications of preservatives and retentions in the FPL field trials. However, some data can be considered as replicates. Four plots list coal tar creosote (Table 4, Fig. 7) with retentions of 3 or 4 pcf (48 or 64 kg/m³). The properties of the creosotes are given in Table 5. Using failure time to 7, stakes in these plots cannot be considered statistically different. However, using time to 0, they are statistically different at the 5 percent level with the following multiple comparison (plot numbers):

20 4 5 6

Zinc chloride appears in four plots (Table 6, Fig. 8). Statistically, retention has little effect on average life. However, the mean failure time for the stakes in plot 32, which had retention of 0.61 pcf(9.8 kg/m³), appears to be different from the mean failure times in other groups:

Time to 7, average retention (plot number)

.3(4) .44(2) .3(2) .6(2) .61(28) .61(4) .91(2) .9(4) .61(32)

Time to 0, average retention (plot number)

.3(4) .61(4) .3(2) .44(2) .61(28) .62(2) .91(2) .9(4) .61(32)

The use of various carriers for 5 percent pentachlorophenol (at a retention of about 4 pcf (64 kg/m³)) can also be compared (Table 7, Fig. 9). Groups 11 and 12 are not included in the multiple compar-

Table 6. Average life of stakes tested with zinc chloride.

Group	Plot	Average stake life (year)		Preservative retention pcf (kg/m ³)
		Time to 0	Time to 7	
1	2	15.4	2.9	0.30 (4.8)
2	2	16.7	2.8	0.44 (7.0)
3	2	17.3	3.0	0.62 (9.9)
4	2	17.9	3.7	0.91 (14.6)
5	4	14.2	2.4	0.30 (4.8)
6	4	14.4	3.2	0.61 (9.8)
7	4	18.1	4.7	0.90 (14.4)
8	28	17.0	3.1	0.61 (9.8)
9	32	24.2	6.1	0.61 (9.8)

isons for time to 0 because not all stakes in those groups have not yet failed. The following multiple comparisons use group numbers found in Table 7. Time to 7

2 1 4 9 5 3 8 10 7 6 12 11

Time to 0

1 2 8 4 7 3 9 6 5 10

If the choice of failure is time to 0, then different carriers (other than heavy gas oil or lube oil extract—groups 11 and 12) may apparently have minimal

Average life of stakes treated with coal tar creosote (year)

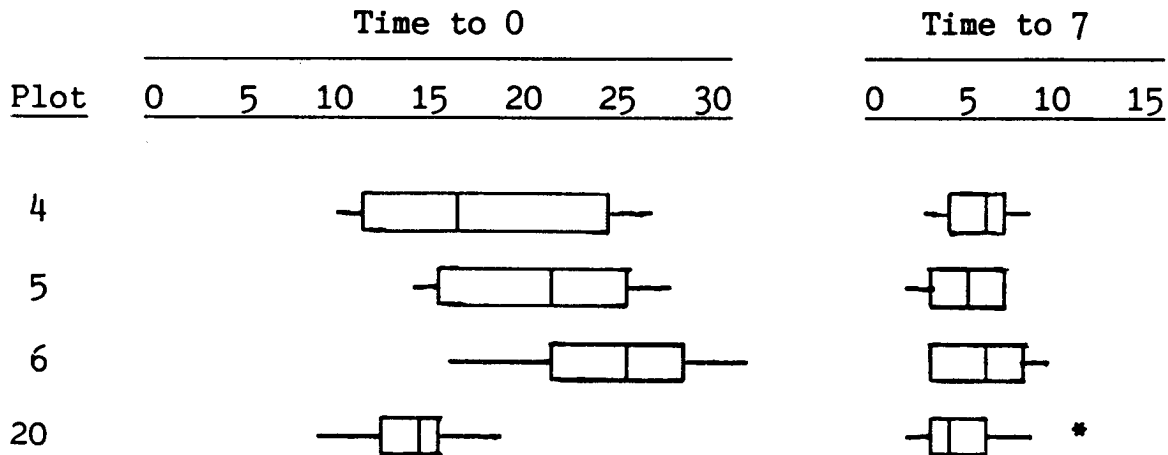


Figure 7. Box plots for stakes treated with coal tar creosote. See Table 4.

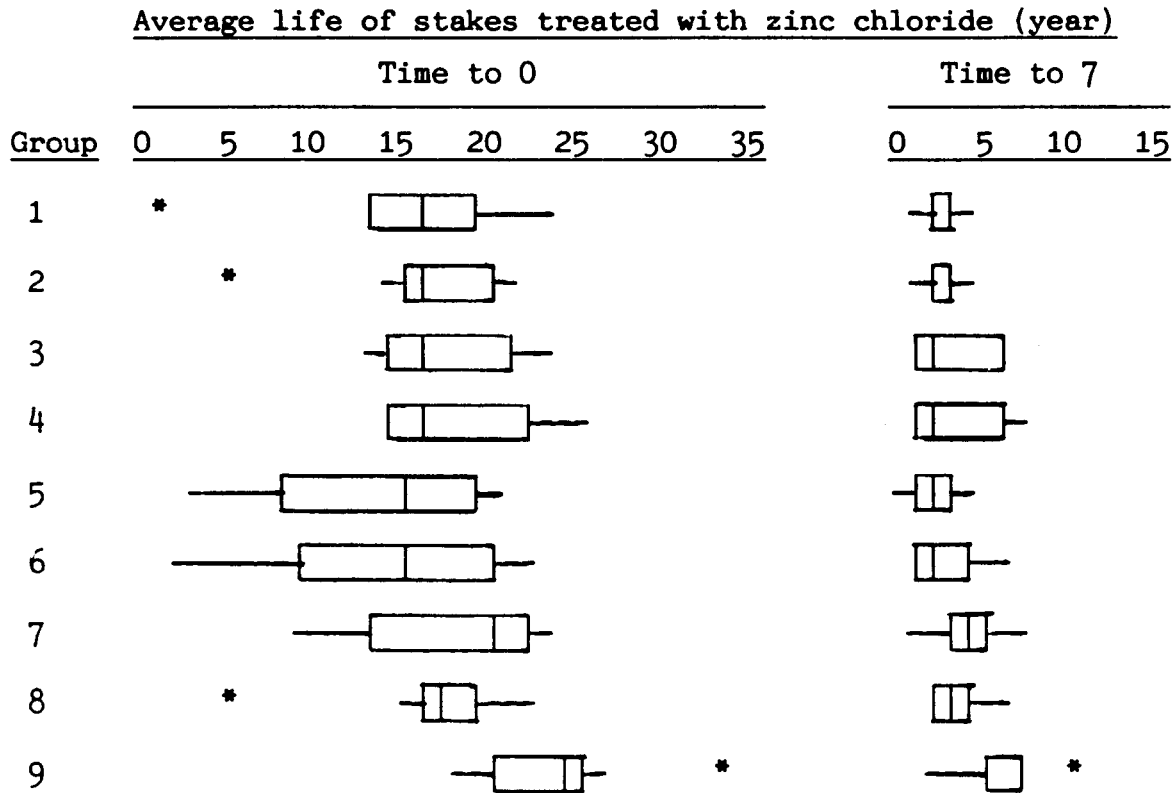


Figure 8. Box plots for stakes treated with zinc chloride. See Table 6.

influence on average stake lifetime. However, if time to 7 is the failure criterion, then more carriers give statistically different average failure times.

Differences between replicates from one plot to another may be a function of several things: actual treatment procedure and preservative retention in a stake, varying presence of decay fungi and termites between plots, the particular stakes used, the potential for inconsistency in subjective ratings, and other extraneous variables.

One method for examining plot-to-plot variation over time is to look at the failure times of untreated control stakes. The average failure times (Link and De Groot, submitted) of untreated control stakes do not vary much over time. However, a cautionary note should be added on the use of controls to compare relative decay and/or termite hazards within any given plot. The untreated controls generally fail within a few years, while the plot may be active for decades. Therefore, the untreated control stakes give information about only relative decay and/or termite hazards for a small proportion of the plot's lifetime.

In this section, we have commented on replicability only within a given site. There is likely to be

considerable site-to-site variability, depending upon the varying decay and/or termite hazards. This variability only increases the difficulty of comparing preservatives because replicability within a given site is not always present.

Conclusions and Recommendations

1. Field trials of preservative-treated stakes are used for various objectives. Because of the large inherent variability in stakes themselves, variable preservative retention for a given stake, and variability of conditions within a plot or site as well as over time, researchers cannot precisely evaluate any given preservative-retention combination with ten replicate stakes. Before a remedy can be proposed for this lack of precision, one needs to reexamine the goals of field testing.

a. If the goal is to screen large numbers of preservatives and retentions, the present setup may be appropriate and will yield an indication of performance by an average time to failure, with confidence intervals expected to be at least 4 years.

Table 7. Average life of stakes treated with pentachlorophenol (5 percent) with various carriers.^a

Group	Carrier	Preservative retention pcf (kg/m ³)	Average stake life (year)	
			Time to 0	Time to 7
1	Fortified petroleum oils and mixtures commercial aromatic solvent (mid U.S.)	4.2 (67)	10.9	2.9
2	Stoddard solvent (mid U.S.)	4.0 (64)	13.7	2.5
3	No. 2 fuel oil (mid U.S.)	4.0 (64)	14.9	4.3
4	Heavy thermal side cut (mid U.S.)	4.0 (64)	14.0	3.0
5	No. 2 diesel oil (West Coast)	4.1 (66)	17.0	4.3
6	Catalytic gas-base oil (West Coast)	4.2 (67)	16.3	7.8
7	No. 300 fuel oil (West Coast)	4.0 (64)	14.6	7.8
8	No. 400 fuel oil (West Coast)	4.2 (67)	13.9	6.8
9	Light gas oil (mid U.S.)	4.0 (64)	15.6	3.9
10	Denver no. 3 blend (50-50 topped crude residual and recycled overhead gas oil)	4.0 (64)	19.5	7.4
11	Heavy gas oil (mid U.S.)	4.1 (66)	^b	14.6
12	Lube oil extract (Texas)	4.2 (67)	^b	11.3

^aFPL field trials, plot 20.

^bNot all 10 stakes have failed

b. If the goal is to describe the failure distribution for one particular preservative and retention including lower percentiles, then larger sample sizes must be used. For example, the fifth percentile failure time of a group often replicate stakes cannot be estimated. Larger sample sizes will also give shorter confidence intervals for the mean or median failure time. The sample size needed depends on the exact goals desired.

c. If the goal is to compare several preservatives and/or retention levels, treated stakes for each preservative and/or retention level should be present in the same plot. This is not always the case in FPL field plots. (As we have described, replicability from one plot to another may be poor.) Experimental designs described for field plots in 'standard' AWWA or ASTM procedures identify the need for including in each plot stakes

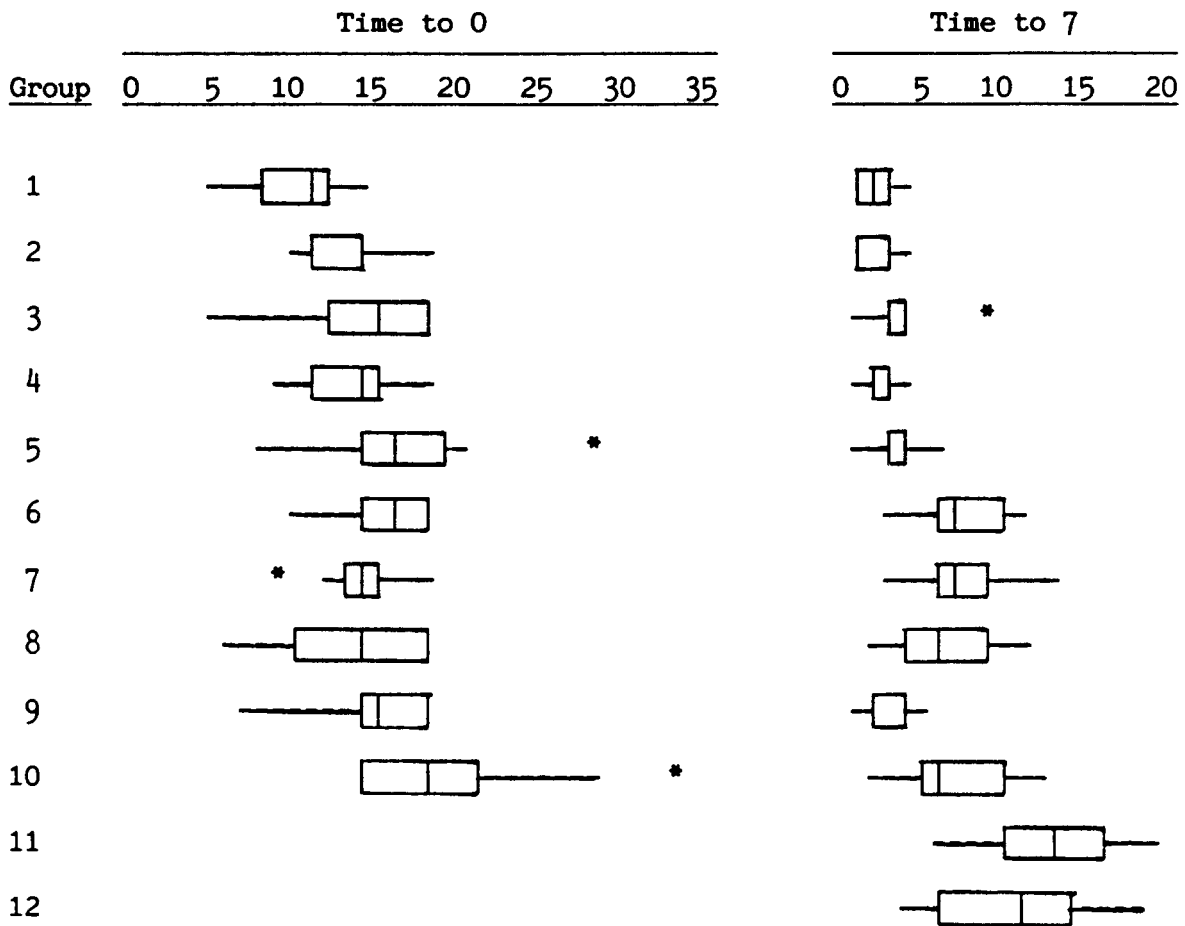
Average stake life of stakes treated with pentachlorophenol (year)

Figure 9. Box plots for stakes treated with pentachlorophenol with different carriers. See Table 7.

treated with reference preservatives. Comparative data of control stakes treated with chromated copper arsenate (CCA) and untreated stakes are one of the minimum report requirements of AWP standard M7 (AWPA, 1988).

2. Given the difference of preservative comparisons based on different definitions of failure time, we recommend that appropriate standards-formulating committees reconsider the formal definition of failure for stakes in field plots. The question is, at what point do we want to compare different preservatives: time until complete failure or an intermediate point in time? Comparisons of various preservatives are unlikely to be the same if different failure criteria are chosen because the preservatives have different degradation patterns,

3. If ratings of preservative-treated stakes are to be compared at a given time, a nonparametric anal-

ysis must be used. The comparison of means is inappropriate because ratings are on ordinal scales. The nonparametric test is insensitive to the choice of units on the scale, as long as the units retain the same ordering.

4. Because current rating scales are subjective and ratings may not be consistent over time, the addition of a nondestructive measurement of the condition of the stake might be useful. This could be provided by a stress wave or mechanical measurement of the modulus of elasticity (MOE). Hand-portable devices are available for mechanical measurement of MOE.

5. Reports on field tests of preservative-treated stakes should include some measure of variability. Box plots are one possible choice. The variability of field-tested preservative-treated stakes should not be hidden by the use of an average value only.

6. Replicability of the results in different plots appears to be questionable, but not enough data are

available for examining this problem. We suggest that this question be considered further. Data should be compared from many locations to better understand plot-to-plot variability. It is likely that a preservative should be independently tested in several plots over time to adequately judge its performance.

Acknowledgment

This research has been partially supported by USDA Forest Service competitive grant 87-FSTY-9-0254.

Literature Cited

- ASTM. 1987. Standard method of evaluating wood preservatives by field tests with stakes. American Society for Testing and Materials. Philadelphia, PA.
- AWPA. 1988. Standard method of evaluating wood preservatives by field tests with stakes. American Wood-Preservers' Association Standard M7-83. Stevensville, MD.
- Borsholt, Erik. 1979. NWPC field test no. 1 with pressure preservatives results during 10 years' testing. Nordic Wood Preservation Council, NWPC-INF. No. 9, ISSN 0358-707 X.
- Colley, Reginald H. 1970. The practical meaning of preservative evaluation tests. Proceedings of the American Wood-Preservers Association. 16-37.
- Conradie, W.E. and Pizzi, A. 1984. Biological effectiveness of ground-contact wood preservatives as determined by field exposure stake tests. Council for Scientific and Industrial Research Special Report HOUT, Pretoria, South Africa, ISBN 0 7988 2987 7.
- De Groot, Rodney C. 1983. FS-FPL-3212, Amendment to study no. 3-80-15. Non-parametric tests of field and laboratory stake tests.
- Dunn, Olive Jean and Clark, Virginia A. 1974. Applied statistics: analysis of variance and regression. John Wiley and Sons, New York. 80-85.
- Gjovik, L.R. and Gutzmer, D.I. 1986. Comparison of wood preservatives in stake tests (1985 progress report). USDA Forest Service Research Note FPL-02, May 1986.
- Hartford, Winslow H. 1972. The interpretation of test plot data. Proceedings of the American Wood-Preservers' Association. 67-82.
- Levy, J.F. and Dickinson, D.J. 1980. Preliminary results from the field experiment to determine the performance of preservative treated hardwoods with particular reference to soft rot. The International Research Group on Wood Preservation, Working Group III, Sub-group 1, Document No: IRG/WP/3164.
- Link, Carol L. and De Groot, Rodney C. Predicting effectiveness of a wood preservative from field trials that use ten replicate stakes, submitted to Wood and Fiber Science.
- Velleman, Paul F. and Hoaglin, David C. 1981. Applications, basics, and computing of exploratory data analysis. Duxbury Press, Belmont, CA.
- SESSION CHAIRMAN GJOVIK: I guess I just have one comment and I've shared this with you before. We have to be careful that we don't improve this thing so we can't use it. It is empirical data and the number you put on that stake depends on how bad a night you had before a lot of times. And you saw those still flipping on these curves. I can't imagine how it was done for a year and a half, but it could have been. But you have to keep this thing in perspective. I would caution all you people who start to play games with these numbers, they are empirical numbers. Thank you.
- This concludes the Technical Session and I'll turn the program back over to the President.
- PRESIDENT MARTINELL: Thank you, Lee. As Lee says this winds up the session for this morning. Remember this afternoon is free. The banquet will be held in this room tonight beginning at seven o'clock. So we'll now adjourn the session.

In: Proceedings of the 85th annual meeting of the American Wood-Preservers' Association; 1989 April 23-26; San Francisco, CA. Stevensville, MD: American Wood-Preservers' Association; 1989: 179-185. Vol. 85.