D26

DISCUSSION

Disc. 105-M69/From the November-December 2008 *ACI Materials Journal*, p. 610

Unbiased Statistical Comparison of Creep and Shrinkage Prediction Models. Paper by Zdenek P. Bažant and Guang-Hua Li

Discussion by David McDonald

FACI, PhD

The paper by Bažant and Li is very much appreciated because it outlines the difficulties that have been faced by committees reviewing shrinkage and creep models.

The paper by McDonald and Roper,[6] which this paper criticizes along with others, compares the graphical approach of plotting residuals, as shown in Fig. 1(c) and (d) of the present paper, and the coefficient of deviation statistic presented by Bažant and Panula in 1978. The paper by McDonald and Roper concluded that "the method of plotting residuals enables visual interpretation of the accuracy of prediction models of shrinkage for concrete."

The caption for Fig. 1 is "Examples of ineffectual statistical comparisons in which all the prediction models... look approximately equally good (or equally bad)." In the presented figures, it is clearly observed that data for the ACI compliance model clearly lie below the zero axis, compared with the B3 Model, which appears to be equally dispersed around the zero axis. Similarly, the ACI model for shrinkage lies above the zero axis, whereas the B3 model appears to be centrally dispersed around the zero axis. Thus, it may be rapidly assessed that the ACI model does not fit the bulk of the data well.

The graphical approach suggested by McDonald and Roper also clearly shows that as time progresses, the spread of data increases. After 1000 days, the B3 compliance model data range from approximately –120 to 100 × 10^{-6}/MPa (–0.82 to 0.69 × 10^{-6}/psi) and the shrinkage ranges from –400 to 400 microstrain. Thus, those using the models are quickly and effectively made aware that the models may be inaccurate for their particular concrete in their particular design. The bottom line in this is that none of the models are accurate and that designers should be aware of these inaccuracies.

The graphic approach of plotting residuals overcomes objections presented in the paper indicating that: a) statistical trends are not reflected; b) plots are dominated by the short term; and c) plots are dominated by long-term low-strength mixtures. These plots also clearly show the heteroscedasticity of the data and the periods of time through which data are measured. It would be interesting to observe the other models plotted in this format.

The Bažant/Li paper criticizes the ACI 209.2R-08, "Guide for Modeling and Calculating Shrinkage and Creep in Hardened Concrete," for publication of a table of statistical values, shown in the current paper as Table 2. As discussed in that document, ACI Committee 209, Creep and Shrinkage in Concrete, found that "there is no agreement as to which statistical indicator(s) should be used, which data sets should be used, or what input data should be considered." For this reason, the committee determined that the best approach was to summarize all of the findings in Table 4.3 of that document.

Table 3

B3 model	GL model
Age of concrete when drying starts	Age of concrete when drying starts
Age of concrete at loading	Age of concrete at loading
Cement type	Cement type
Concrete mean compressive strength at 28 days	Concrete mean compressive strength at 28 days
Relative humidity	Relative humidity
Volume-surface ratio	Volume-surface ratio
Cement content in concrete	—
Curing method	—
Aggregate content in concrete	—
Shape of specimen	—
Water content in concrete	—

The paper also compares the accuracy of the B3 and GL models along with others. Figure 8 shows that the standard indicator for compliance and shrinkage for the B3 models are 27.3 and 28.5%, respectively, whereas for the GL model it is 30.2 and 31.0%, respectively. The two models contain substantial differences in input data, as shown in Table 3. The B3 model requires significantly more knowledge of the concrete materials and curing, as shown in italic. The use of these additional data only improves the model standard indicator by less than 3%. It would be interesting to determine how the B3 model would perform without these variables, which are frequently not available to the designer at the time of design.

As concluded in the paper by McDonald and Roper, "The designer should not be misled in the belief that, by adding factors to the prediction, a 'better' result is achieved, particularly if no prior information on the performance of a particular concrete is available."

Disc. 105-M69/From the November-December 2008 *ACI Materials Journal*, p. 610

Unbiased Statistical Comparison of Creep and Shrinkage Prediction Models. Paper by Zdenek P. Bažant and Guang-Hua Li

Discussion by N. J. Gardner

FACI, Professor Emeritus, University of Ottawa, Ottawa, ON, Canada

This discussion has three themes: 1) to show that the ranking of the models does not change with the method of calculation of the coefficient of variation (COV); 2) to give the COVs using the information available at the design stage; and 3) to suggest COVs that should be used for structural sensitivity calculations.

The main thesis of the paper is that least squares regression statistics is the only valid method of comparing different shrinkage and creep models. The discusser does not accept this thesis. Further, the discusser does not accept that the method described in Eq. (2) through (4) is less biased than that given in Eq. (6) and (7). Neither the population COV (Eq. (2)) nor the population correlation coefficient (Eq. (5)) or Gardner's COV (Eq. (6)) give information on how the fit systematically changes with time. There is a difference between evaluating the goodness of fit between existing models and experimental data and using least squares regression to develop a new equation.

All models are input information sensitive, for example, should the predictions be determined using mean concrete strength only or all available information, including measured concrete strength, measured modulus of elasticity, and mixture proportions.

Conventionally, the arithmetic mean is used to measure central tendency and the standard deviation is a measure of the dispersion of the individual observations from the mean of a number of relevant results. The COV is the ratio of the dispersion to the mean value.

Unfortunately, conventional statistics are not appropriate as concrete shrinkage and creep increase with time from drying or loading, respectively (heteroscedastic). The use of the generally accepted term COV is unfortunate. The term COV (RMS (Calculated – Observed)/experimental mean) used in the paper and this discussion is different from the generally accepted definition and is not appropriate for sensitivity studies.

To demonstrate the effects of different definitions on the calculated COV and the subsequent ranking of the models, the discusser attaches Tables 4, shrinkage, and 5, compliance, calculated using measured concrete strength, measured modulus of elasticity, and mixture composition. The calculations used Gardner's reduced RILEM databank for comparison with Reference 2, which only includes data sets longer than 500 days. The humidity range for compliance was 20 to 100% and below 80% relative humidity (RH) for shrinkage—swelling was not included. The comparison considered 107 data series for shrinkage and 166 data series for compliance. (This appears to be most of the long duration data indicated in Fig. 2). Individual values were removed from each data series so that successive observation ages were approximately twice the previous age. The predicted compliances (Table 2) were the calculated specific creep plus the measured immediate elastic compliance. Preferably, the interval COV should be effectively constant. Even though the values are given to three significant figures, the COVs should be rounded before ranking the methods. The COV calculated using Eq. (4) and (6) are similar, but more information is available using Eq. (6). The rankings are unchanged regardless of which indicator is used. The effect of the early age interval mean value on the interval COV is clearly seen in Table 4, where the interval COVs are larger at early ages. As stated the shrinkage calculations only considered humidities less than 80%—this may be a reason for the vast difference between the discusser's COVs and the unrealistic values given in Table 1(b). The compliance interval COVs (Table 2) are relatively constant for every time interval.

All models are input data sensitive. During design, only the mean concrete strength will be known. Tables 6 and 7 give the COVs calculated using only the mean concrete cylinder strength. If the mean value is not known, the mean concrete strength can be inferred from the specified strength. In these comparisons, B3 suffers because mixture proportions have been inferred from the cylinder strength. The easiest way to reduce the COVs is to prequalify the concrete; measure the concrete strength and modulus of elasticity, GL2000 and EC2; and use the actual mixture proportions for B3. Strangely, the shrinkage COVs for ACI 209R do not improve slightly when more input data are available.

For sensitivity calculations, the discusser recommends for design that the average interval COVs, rounded to the nearest 5%, be used for shrinkage and compliance for GL2000, B3, and EC2. The number of multiples of the COV to be used is the analyst's choice.

Table 4—Comparison of shrinkage predictions calculated using average of measured and back-calculated f_{cm} and mixture information

Time interval*	Average time, days	Average shrinkage, microstrain	GL2000 Interval COV, %	B3 Interval COV, %	EC2† Interval COV, %	ACI 209R Interval COV, %
Above 3160	5532	820	12.6	15.7	28.2	38.8
1000-3159	1588	642	17.0	19.0	27.9	36.8
316-999	570	591	20.2	18.6	31.3	32.6
100-315	186	486	20.9	19.4	32.7	33.4
33-99	69	367	26.7	25.0	41.3	41.0
10-31	21	249	28.2	22.7	40.7	55.8
3.0-9.9	5.5	111	42.2	34.2	42.8	78.3
Average interval RMS/average interval average (Eq. (6))‡			19.8%	19.6%	32.3%	39.4%
RMS (interval COV)			25.6%	22.8%	35.5%	47.7%
Population COV (without weighing)			22.6%	21.6%	35.8%	41.0%
Population COV (Eq. (4))			20.5%	20.7%	34.4%	42.1%

*Duration limits are half $\log_{10}$ intervals; that is, 0.5, 1, 1.5, 2, 2.5, etc.

†EC2 is essentially the same as CEB MC1990-99, with limits on some modifying factors.

‡Calculated from information not available in table.

Table 5—Comparison of compliance predictions calculated using average of measured and back-calculated f_{cm} and mixture information

Time interval	Average time, days	Average compliance, microstrain/MPa	GL2000 Interval COV, %	B3 Interval COV, %	EC2 Interval COV, %	ACI 209R Interval COV, %
Above 3160	5661	142	26.3	29.2	34.3	35.0
1000-3159	1538	124	24.1	26.2	31.1	28.5
316-999	568	106	23.7	28.9	29.6	26.1
100-315	180	93	22.6	28.4	27.4	24.1
33-99	70	79	19.2	25.6	24.1	20.8
10-31	21	70	19.9	24.6	20.3	24.9
3.0-9.9	5.2	59	16.4	23.4	15.8	29.9
Average interval RMS/average interval average (Eq. (6))			22.6%	27.1%	27.7%	27.7%
RMS (interval COV)			20.2%	24.4%	24.7%	25.8%
Population COV (without weighing)			23.3%	28.2%	28.6%	27.6%
Population COV (Eq. (4))			24.5%	28.6%	30.8%	30.1%

Note: 1 MPa = 145 psi.

Table 6—Comparison of shrinkage predictions calculated using measured f_{cm} only

Time interval	Average time, days	Average shrinkage, microstrain	GL2000 Interval COV, %	B3 Interval COV, %	EC2 Interval COV, %	ACI 209R Interval COV, %
Above 3160	5532	820	17.2	23.4	28.2	38.6
1000-3159	1588	642	22.3	25.6	27.9	36.2
316-999	570	591	26.7	33.1	31.3	31.4
100-315	186	486	27.4	32.7	32.7	32.5
33-99	69	367	36.0	39.4	41.3	40.2
10-31	21	249	34.7	39.8	40.7	55.6
3.0-9.9	5.5	111	42.8	45.7	42.8	76.7
Average interval RMS/average interval average (Eq. (6))			25.8%	30.8%	32.3%	38.8%
RMS (interval COV)			30.7%	35.0%	35.5%	47.1%
Population COV (without weighing)			29.6%	35.2%	35.8%	40.3%
Population COV (Eq. (4))			26.9%	32.4%	34.4%	41.5%

Table 7—Comparison of compliance predictions calculated using measured f_{cm} only

Time interval	Average time, days	Average compliance, microstrain/MPa	GL2000 Interval COV, %	B3 Interval COV, %	EC2 Interval COV, %	ACI 209R Interval COV, %
Above 3160	5661	142	31.4	30.5	35.6	41.6
1000-3159	1538	124	27.8	28.0	29.5	36.0
316-999	568	106	26.5	29.3	28.8	33.9
100-315	180	93	27.0	31.1	30.3	33.6
33-99	70	79	24.5	29.2	27.7	29.8
10-31	21	70	23.3	29.0	27.1	28.6
3.0-9.9	5.2	59	22.6	27.9	27.9	30.2
Elastic	0	38.7	26.3	*	28.2	24.8
Average interval RMS/average interval average (Eq. (6))			26.9%	29.4%	30.2%	34.5%
RMS (interval COV)			24.6%	27.5%	27.8%	31.5%
Population COV (without weighing)			27.4%	30.4%	30.5%	34.9%
Population COV (Eq. (4))			28.9%	30.7%	32.2%	37.3%

*Method B3 calculates compliance directly.

AUTHORS' CLOSURE

Both discussions are appreciated for providing a welcome opportunity of clarification.

McDonald's discussion

McDonald states that his "graphic approach of plotting residuals overcomes objections presented in the paper, indicating that: a) statistical trends are not reflected; b) plots are dominated by the short term; and c) plots are dominated by long-term low-strength mixtures." This is untrue, as well his objection to labeling the statistical comparisons in his plots (Fig. 1(c) and (d)) co-opted in ACI 209.2R-08 as "ineffectual." Briefly, the reasons are as follows:

1. The median cannot distinguish the error magnitudes. For example, the median lines and the percentages of positive and negative errors would remain unchanged if the data points lying one side of the zero line were all pushed very close to that line, or very far from it.

2. McDonald's comparisons imply the optimum fit to be the median line of the database trend. The error of the median, however, is known in statistics to be much larger than the error of the mean, which is what is obtained by the least-square regression.

3. In the special case of a linear model, every meaningful statistical method must reduce to standard linear regression. Yet, McDonald's median line differs greatly from the least-square regression line.

4. The short-term data do dominate the statistical as well as the visual comparisons because they are far more numerous. Those for low-strength concretes dominate because they exhibit larger deformations. No weights were introduced by McDonald to offset this bias.

McDonald further credits his data plots with revealing that "the spread of data increases" in time. This is, of course, true but needs no demonstration because the initial spread of shrinkage or creep data is, by definition, always zero. There is an incremental random process[62] for which an increasing spread is typical.

McDonald's statement that his "plots clearly show the heteroscedasticity of the data" is true, but only in a linear scale plot. As is well known in statistics, for least-square regression, the scale should be transformed so as to achieve near-homoscedasticity. This could have been achieved by a logarithmic transformation of the deformation scale (refer to Fig. 4(b)).

McDonald objects that the "Bažant/Li paper criticizes the ACI 209.2R-08 ... for publication of a table of statistical values, shown in the current paper as Table 2." It was mathematically demonstrated in the paper, however, that most of the data in that table were obtained by incorrect statistical methods.

McDonald's statement that "the B3 model requires significantly more knowledge of the concrete material and curing" (than other models) is baseless. Model B3 can be used even if no additional knowledge is available because the additional influencing parameters (cement, aggregate, and water contents), known to be significant, can be assigned either the default values typical of concretes in the given area or the values crudely estimated from the design strength according to the empirical equation given in Model B3.

McDonald correctly points out two types of comparison in which the coefficients of variation (COVs) of errors of Models B3 and GL are very close; however, for other types of comparison of B3 versus GL, Fig. 8 shows the percentages 23.8 versus 27.6, 33.3 versus 41.7, 29.4 versus 37.7, 34.5 versus 43.3, and in the logarithmic scale, 0.22 versus 0.40. These are not close.

The sentence cited from McDonald and Roper is inaccurate. "Better results," of course, need not be achieved, but may sometimes be achieved, which is often the case with the additional parameters of Model B3. This was demonstrated by analysis of the excessive deflections of the record-span box-girder in Palau,[63] which reached 1.61 m (5.28 ft) within 18 years. The best match of the observations was obtained with Model B3, which has a theoretical basis and has been validated and calibrated by a database using standard statistics with bias suppression. The GL model was distant second, and the ACI and CEB-fib models were far worse. These conclusions disagree with the comparisons of the existing models in ACI 209.2R-08.

Gardner's discussion

In Gardner's nonstandard regression statistics, the arithmetic mean of square roots (MSR) of the variances of errors in the individual time intervals is adopted as the measure of scatter that should be minimized. This kind of regression statistics, however, is fundamentally incorrect. The correct statistics must be based on the root of the mean of squared errors

(mean-root-square [RMS]), which follows from the requirement of maximum-likelihood fit of the data, as mathematically proven in the Appendix.

Apart from the mathematical derivation in the Appendix, the necessity of an RMS approach is most simply demonstrated by considering the linear regression, that is, the fitting of a straight line $Y_{ij} = a + bX_{ij}$ to a band of data points with a trend (b = regression line slope, a = Y-axis intercept). Setting $\partial s^2/\partial a = 0$, $\partial s^2/\partial b = 0$ (where s^2 = standard least-square optimization criterion [Eq. (10)]), one gets for a and b two linear equations

$$(\Sigma_i w_i m_i)a + (\Sigma_i w_i \Sigma_j X_{ij})b = \Sigma_i w_i \Sigma_j y_{ij} \quad (18)$$

$$(\Sigma_i w_i \Sigma_j X_{ij})a + (\Sigma_i w_i \Sigma_j X_{ij}^2)b = \Sigma_i w_i \Sigma_j y_{ij} X_{ij} \quad (19)$$

where the summations run over $I = 1, 2, \ldots n$ and $j = 1, 2, \ldots m_i$.

On the other hand, setting $\partial s_G^2/\partial a = 0$, $\partial s_G^2/\partial b = 0$ (where s_G = Gardner's comparison criterion [Eq. (9)]), one obtains

$$\sum_{i=1}^{n}\left\{\left[(m_i-1)\sum_{j=1}^{m_i}(a+bX_{ij}-y_{ij})^2\right]^{-1/2}\sum_{j=1}^{m_i}(a+bX_{ij}-y_{ij})\right\} = 0 \quad (20)$$

$$\sum_{i=1}^{n}\left\{\left[(m_i-1)\sum_{j=1}^{m_i}(a+bX_{ij}-y_{ij})^2\right]^{-1/2}\sum_{j=1}^{m_i}(a+bX_{ij}-y_{ij})X_{ij}\right\} = 0 \quad (21)$$

The fact that these equations for a and b are different and nonlinear, even for a linear model, documents that the minimization of Eq. (9), used by Gardner, is incorrect.

It has been thought that, because the creep and shrinkage data fitting is a nonlinear problem, the minimization of s_G (Eq. (9)) does not have to include linear regression as a special case. But this idea is again incorrect. Imagine the model $Y = a + bX + cX^2$, in which the nonlinearity is caused by a single parameter c, and consider a series of data-fitting problems with a smaller and smaller c, until c vanishes. At some value of c, one would have to switch from Gardner's regression to the standard least-square regression. The switch would lead to a sudden discontinuous jump in the optimum parameter values, which is unjustifiable. Moreover, at which value to switch? At $c = 0.001$? Or $c = 0.5$? Or $c = 0.999$? Obviously, none makes sense. So, model comparisons based on s_G, and generally on any statistics that cannot yield the straight-line fit of data, are meaningless.

Note also that Gardner cites no treatise on regression statistics that would justify his method (Eq. (6) and (7) in the paper). References 25, 26, 43-45, and 48-57 represent the basic sources on the fitting of experimental databases. They all use the least-square regression, and none use Gardner's approach.

Calling Eq. (2) "the population coefficient of variation" suggests that Gardner might regard the problem as population statistics. This is not the case. In spite of the subdivision of time into intervals, we face a problem of statistical regression, and Eq. (2) represents the COV of regression errors (that is, the standard error of regression divided by the mean).

Gardner compensates for unequal numbers of data points in various equal-size intervals of the logarithmic time scale by assigning to the mean and variance in each interval the same weight. This is reasonable but is equivalent to assigning to each point in interval i the weight $w_i = 1/m_i$, as done in the paper.

The subdivision into time intervals does not convert our regression problem to population statistics. Especially, it does not mean that the minimized expression would be the mean of square roots (MSR).

In Tables 5 and 6, Gardner shows that, in some cases, the differences between his statistics and the statistics of least-square regression are insignificant. That is true. But there are cases where the differences are significant. Refer to the ranking of the models in Table 2 of the paper, taken from ACI 209.2R-08. That ranking is very different from the least-square ranking in the paper (Fig. 5), and happens to be more favorable to Model GL.

The subject of sensitivity analysis brought up by Gardner is important but again it makes no sense to evaluate sensitivity to various parameters from changes of nonstandard COVs that are not based on least squares.

Gardner is certainly right in pointing out that there is a difference between the goodness-of-fit criteria and the least-square regression. In creep and shrinkage, however, we are still struggling with the central range statistics, which yield the optimum mean curves of compliance and shrinkage and perhaps their COV, while the goodness-of-fit criteria are important for the distribution tails, which are important for structural safety where failure probability should not exceed 10^{-6}. In durability design, however, we should be satisfied if only one in approximately 20 structures, not one in a million, requires repair after 20 years. Knowing the COV is sufficient for that purpose.

Hence, the assumption of Gaussian (or normal) errors, underlying the least-square regression, is justified. It can be introduced either for the actual scale of deformations, or for the logarithmic scale (the latter being equivalent to lognormal errors in the actual scale). For suppressing heteroscedasticity, the latter is better, but from other viewpoints not, and the choice is debatable. Both were considered in the paper, and the results were similar.

Finally, it should be emphasized that, because of the severe scarcity of laboratory data for long times and large specimens, a more productive way to compare concrete creep prediction models is to check their predictions of the observed multi-decade deflections of large structures. The recent long-time deflection analyses of the bridge in Palau and other bridges document much more clearly than any database statistics that the comparisons in ACI 209.2R-08 are misleading. If accepted, premature serviceability losses would likely result, typically within approximately 20 years, whereas AASHTO recommends the design lifetime to exceed 50 years and the transportation departments nowadays usually require at least 100 years (as, for example, for the rebuilt I-35 bridge in Minneapolis, MN).

Errata: After the second equality sign in Eq. (4), the factor $\overline{w}/n$ needs to be inserted to achieve proper normalization of statistical weights.

REFERENCES

62. Cinlar, E.; Bažant, Z. P.; and Osman, E., "Stochastic Process for Extrapolating Concrete Creep," *Journal of the Engineering Mechanics Divisions*, ASCE, V. 103, 1977, pp. 1069-1088.

63. Bažant, Z. P.; Li, G.-H.; Yu, Q.; Klein, G.; and Krístek, V., "Explanations of Excessive Long-Time Deflections of Collapsed Record-Span Box Girder Bridge in Palau," *Preliminary Structural Engineering Report* 09-09/A222e, Northwestern University, Evanston, IL, 2009.