National Cooperative Highway Research Program

RESEARCH RESULTS DIGEST

December 2003—Number 283

Subject Areas: IIB Pavement Design, Management, and Performance and IIIB Materials and Construction

Responsible Senior Program Officer: Edward T. Harrigan

Jackknife Testing—An Experimental Approach to Refine Model Calibration and Validation

This digest summarizes key findings from NCHRP Project 9-30, "Experimental Plan for Calibration and Validation of Hot Mix Asphalt Performance Models for Mix and Structural Design," conducted by Fugro-BRE, Inc. The digest is an abridgement of portions of the project final report by principal investigator Harold L. Von Quintus, P.E., Fugro-BRE, Inc.; Charles Schwartz, Ph.D., P.E., University of Maryland-College Park; Richard H. McCuen, Ph.D., University of Maryland-College Park; and Dragos Andrei, Ph.D., Fugro-BRE, Inc.

INTRODUCTION

This digest summarizes findings from research conducted under NCHRP Project 9-30 with the objective of developing a detailed, statistically sound, and practical experimental plan to refine the calibration and validation of the performance models incorporated in the pavement design guide (hereinafter referred to as the 2002 Design Guide) produced in NCHRP Project 1-37A with laboratory-measured hot mix asphalt (HMA) material properties.

Jackknifing is an analytical procedure for refining and confirming the calibration coefficients of mechanistic-empirical (M-E) distress prediction equations and models such as those used in the 2002 Design Guide. Jackknifing provides more reliable assessments of model prediction accuracy than the alternative use of either traditional split-sample validation or calibration goodness-of-fit statistics because jackknifing's goodness-of-fit statistics are based on predictions rather than the data used for fitting the model parameters (Miller, 1974; Mosteller and Tukey, 1977). Thus, the model validation statistics are developed independently of the data used for calibration. Multiple jackknifing is used to assess the sensitivity of the validation goodness-of-fit statistics to sample size.

To develop jackknife statistics from a sample of n sets of measured values, the data matrix is divided into two groups, one part for calibration and the other for prediction. Assume that the data matrix includes measurements of p predictor variables X_{ij}, $j = 1 \ldots p$ and a single criterion variable Y_i, with $i = 1 \ldots n$ sets of measured values. Therefore, the data matrix will have n rows and $p+1$ columns. For an n–1 jackknife validation, the procedure begins by removing one set of measurements from the data matrix and calibrating the model with the remaining n–1 sets of measurements. The kth set of measurements that was withheld is then used to predict the criterion variable Y_k, from which the error (e_1) is computed as the difference between the predicted ($\hat{Y}_k$) and measured (Y_k) values of the criterion variable. A second set of measurements is removed while replacing the first set, and the new n–1 set is used to calibrate a new model. This new calibrated model is then used with the withheld set of X values to predict Y and compute the error e_2.

The process of withholding, calibrating, and predicting is repeated until all n sets have been used for prediction. This yields n values of the error, from which the jackknifing goodness-of-fit statistics can be computed. While both the calibration statistics based on all n sets and the jackknifing statistics are computed from n measures of the error, the jackknifed errors are computed from measured X values that were not used in calibrating the model coefficients. Thus, the jackknifing goodness-of-fit statistics are considered to be independent measures of model accuracy.

TRANSPORTATION RESEARCH BOARD
OF THE NATIONAL ACADEMIES

Because sample sizes of most pavement engineering data sets are limited, one objective of model validation is to assess the sensitivity, or stability, of the accuracy of the model to sample size. To assess the stability of the jackknifed goodness-of-fit statistics, multiple jackknifing can be performed by withholding two sets of *X,* while calibrating on the remaining *n*–2 sets. Two errors are computed for each calibration based on the *n*–2 withheld sets of *X*. For small samples, the goodness-of-fit statistics for the *n*–2 jackknifing may be quite different from those for the *n*–1 jackknifing. If, however, the *n*–1 and *n*–2 jackknifing goodness-of-fit statistics are similar, the *n*–1 jackknifing statistics are not sensitive to the sample size and the statistics are stable. Stable statistics are reliable indicators of goodness-of-fit or prediction accuracy.

The primary advantage of jackknifing is that the goodness-of-fit statistics are based on predictions from data that are independent of the calibration data. Thus, they more likely indicate the accuracy of future predictions than the statistics based on calibration of all *n* data vectors. The use of multiple jackknifing to assess the stability of the prediction statistics is a second advantage of jackknifing. A third advantage is that the method is easy to apply.

Split-sample validation differs from jackknifing in that the goodness-of-fit statistics for both calibration and prediction are based on *n/*2 values (for symmetrical split sampling, the usual case) rather than *n* values. Traditional split-sample validation has the distinct disadvantage that, if *n* is small relative to the infra-space being simulated, then *n/*2 is even smaller, which produces inaccurate calibrations, inaccurate coefficients, and less reliable prediction accuracy.

A procedure proposed for the NCHRP Project 9-30 experimental plan presented in *NCHRP Research Results Digest 284* combines jackknifing with split-sample testing in what is essentially an *n/*2 jackknifing scheme termed *split-sample jackknifing*. Split-sample jackknifing provides somewhat better measures of prediction accuracy than the traditional split-sample validation. Using this procedure, the number of test sections required to accurately recalibrate the HMA distress prediction models in the 2002 Pavement Design Guide with measured materials properties was estimated from the standard error of estimate for each model published in the final report for Project 1-37A, "Development of the 2002 Guide for the Design of New and Rehabilitated Pavement Structures: Phase II."

This split-sample jackknifing procedure for calibration and validation is illustrated in the remainder of this digest using simulations of rutting performance based on measured data from in-service pavement sections at the Minnesota Road Research Project (MnROAD).

JACKKNIFE TESTING—APPLICATION OF THE EXPERIMENTAL APPROACH

Overview and Model Formulation

To illustrate the split-sample jackknifing procedure for calibration and validation, measured permanent deformation (rutting) data from five pavement sections at the MnROAD test site were used to develop a statistical simulation model. The five pavement sections—MnROAD Cells 16, 17, 18, 20, and 22—are from an in-service length of Interstate 94 in Minnesota.

The site climate is characterized as moderate to cold temperature. All sections were conventional flexible systems with approximately 8 inches of HMA over either 28 inches of good-quality base (Minnesota Department of Transportation [MnDOT] Class 3 material in Cells 16, 17, 18, and 20) or 18 inches of high-quality crushed stone base (MnDOT Class 6 material in Cell 22). The HMA at the five test sections differed in the binder grade and mixture gradation. The natural soil subgrade under all sections was a silty clay (AASHTO A-6). Traffic loading consisted of typical Interstate highway traffic mix with a volume of about 25,000 vehicles per day and approximately 12.5% trucks; these traffic conditions were identical for all five sections. The test sections were opened to traffic in 1994.

Rut depth and other performance measurements have been taken periodically since the opening of the sections to traffic. The rutting performance history used in this example spans slightly more than 3 million equivalent single-axle loads (ESALs). Eight rut depth measurements distributed over the life of the pavement were typically available for each of the test sections. Figure 1 shows the measured rut depth versus traffic histories for the five sites.

Four of the five sections (Cells 16, 17, 18, and 22) exhibited very similar rutting performance histories, with maximum rut depths of 4 to 6 mm after 3 million ESALs. The fifth section (Cell 20) exhibited a maximum rut depth of approximately 17 mm after only 2.5 million ESALs. However, this type of variability is believed to be typical among multiple sites within any given pavement condition category.

These five MnROAD test sections had previously been studied as part of the development of the simple performance test in NCHRP Project 9-19, "Superpave Support and Performance Models Management," because all conditions other than HMA mixture properties were identical. The preliminary results from this study and standard empirical model forms for pavement rutting were used to formulate a simple performance model relating the cumulative rut depth *D,* in millimeters, to the number of traffic repetitions, *N* (defined in terms of millions of 18-kip ESALs, or MESALs), and the mixture stiffness *S* in 10^6 psi:

$$D = b_0 N^{b_1} S^{b_2} \tag{1}$$

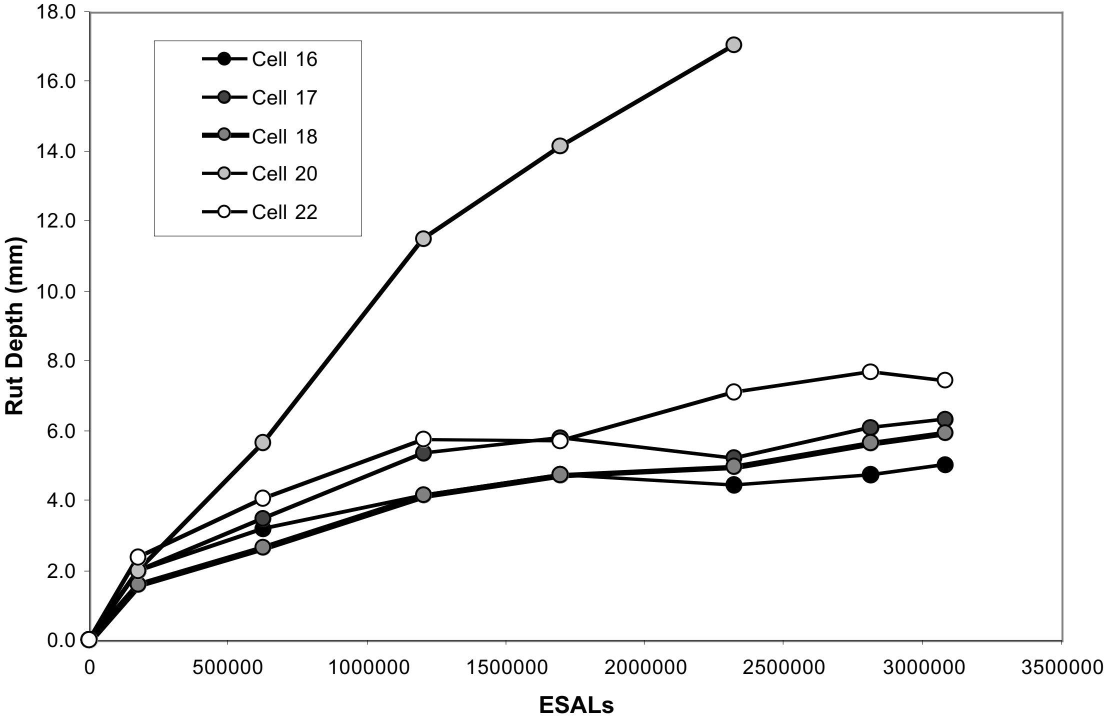


Figure 1. Measured rutting performance histories at selected MnROAD test sections.

The mixture stiffness S is quantified by the ratio $|E^*|/\sin\phi$ in accordance with the results of Task C of NCHRP Project 9-19, where $|E^*|$ is the dynamic modulus and ϕ is the phase angle for a reference temperature and loading rate. The stiffness S is intended to capture the differences in binder and gradation among the mixtures. The stiffness values for the five sites varied from 0.6 to 1.5 million psi, with a mean value of 1.15 ± 0.34 million psi.

Model Calibration

The two-predictor power model form in Equation 1 was fitted to the data using least squares regression of the logarithmic transformation:

$$\log D = \log(b_0) + b_1 \log N + b_2 \log S \qquad (2)$$

The parameters of b_0, b_1, and b_2 are the regression coefficients. These three coefficients were fitted to the entire MnROAD data set for these five pavement sections with the following result in transformed log-log space:

$$\log D = 0.695 + 0.454 \log N - 0.795 \log S \qquad (3)$$

In arithmetic space, Equation 3 can be written as

$$D = 4.96 N^{0.454} S^{-0.795} \qquad (4)$$

The calibration goodness-of-fit statistics in arithmetic space are as follows: correlation coefficient R = 0.833 (R^2 = 0.694), standard error of estimate S_e = 1.87 mm, relative standard error S_e/S_y = 0.5718, bias $\bar{e}$ = –0.255 mm, and the relative bias (bias divided by the mean D) R_b = –4.60%. The bias is minimal but not insignificant. The relative standard error ratio suggests moderate precision.

The statistics and the model of Equation 4, which are based on the MnROAD data, are used as the statistical population values for a Monte Carlo simulation exercise. The purpose of the simulation is to extend the model beyond constraints imposed by the data structure of the MnROAD data. Specifically, the number of sections, the number of observations at each section, and the maximum ESAL value at which rutting data are measured can be varied to assess their effects on the calibration and validation statistics for the model.

Calibration Versus Prediction Accuracy

The calibration goodness-of-fit statistics for the model of Equation 4 should not be assumed to represent prediction accuracy; true estimation of prediction accuracy requires model validation. The model of Equation 4 was therefore used as the basis for a Monte Carlo model validation simulation using *n*–1, *n*–2, and split-sample jackknifing. The validation goodness-of-fit statistics for these analyses can then be compared with those for calibration.

Rational design of pavements requires that the pavement distresses at the end of the design life just reach their limiting design values. Thus, it is the accuracy of performance predictions at the end of the pavement's design life that is of most interest in assessing the value of the model. Various factors influence the prediction accuracy of a model. Six factors considered in the present study are as follows: the underlying but unknown population correlation coefficient of the model, ρ; the desired level of confidence γ for the performance predictions; the number of traffic repetitions at the end of the pavement design life, N_d; the number of pavement sections within a region, *m*; the number of performance measurements per section, *n*; and the ESAL value at the last performance measurement, N_m.

Monte Carlo simulations (Ross, 1990; Sobol, 1994) were performed to evaluate the influence of each of these factors on the prediction accuracy of the model. The Monte Carlo model is of the following form:

$$D = b_0 N^{b_1} S^{b_2} \varepsilon \tag{5}$$

The parameters of b_0, b_1, and b_2 constants are the population values that are assumed equal to the MnROAD values determined in Equation 4, and ε is the stochastic component that is inversely related to the population correlation coefficient ρ for the model. A total of 15,000 simulations were performed for each set of study conditions. For each simulation, rut depth predictions were made at N_d values of 10 and 15 MESALs. The 15,000 predictions in each set were tabulated into histograms from which 60%, 75%, and 90% one-sided nonexceedence values were determined. A relative error D_R was then computed as the difference between the predicted rut depth $\hat{D}_\gamma$ at a specified confidence level γ and the mean population rut depth $\overline{D}$ for the same ESAL normalized by the mean population rutting depth:

$$D_R = \left(\frac{\left| \hat{D}_\gamma - \overline{D} \right|}{\overline{D}} \right) \tag{6}$$

This normalized error value should be relatively transferable to other pavement conditions. The numerator of the ratio is simply the expected error in the prediction at the specified confidence level.

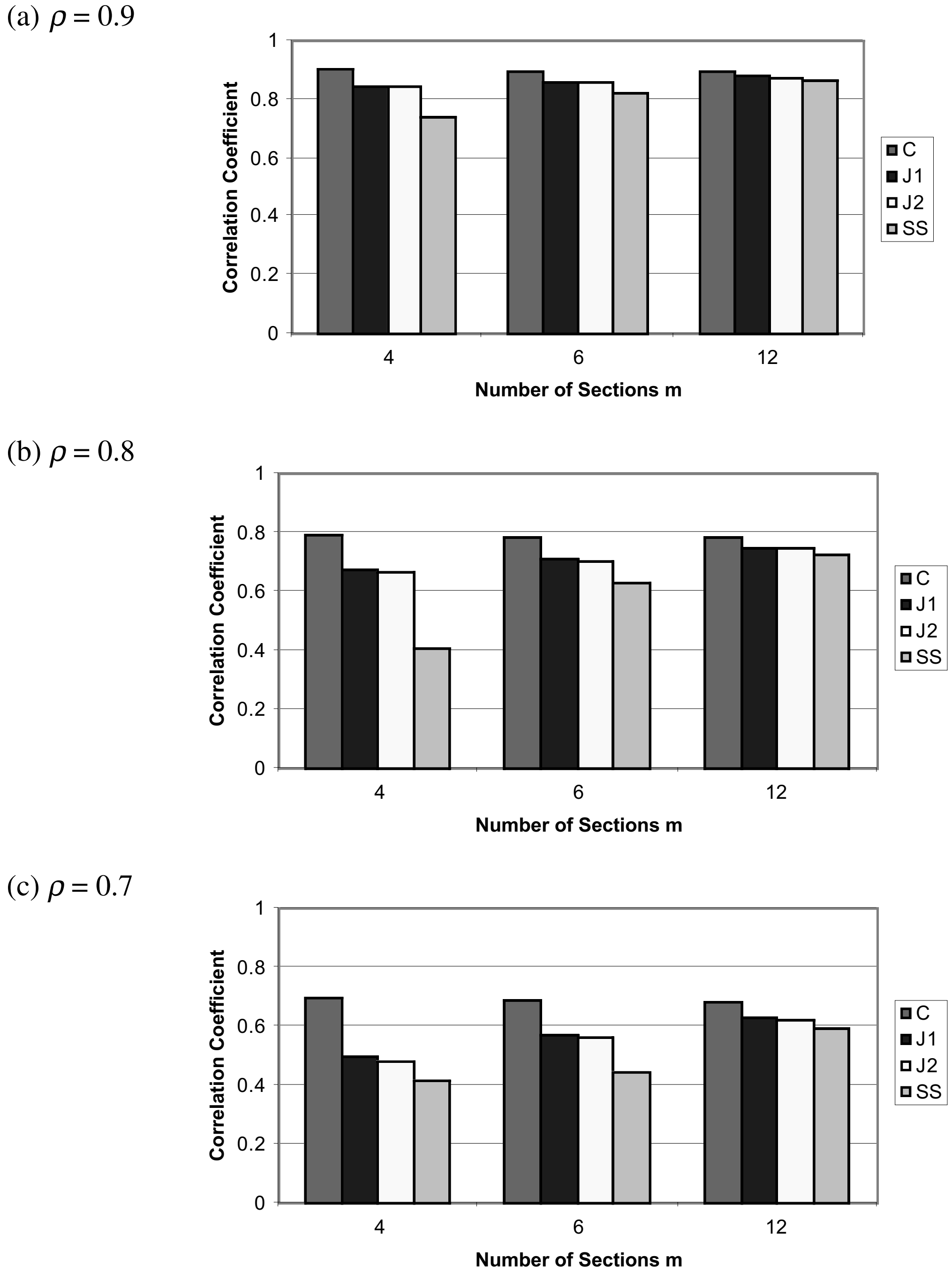
A central benefit of jackknifing is that it assesses prediction accuracy, not calibration accuracy. A principal objective of this analysis was to identify the conditions under which jackknifed prediction statistics differed significantly from calibration statistics. The differences between the calibration and validation goodness-of-fit indices should depend on several data characteristics, most notably the number of pavement sections used to develop the prediction equation.

Figures 2 and 3 summarize values of the correlation coefficient and the standard error ratio, respectively, for selected numbers of pavement sections. The values plotted in the figures are the median statistics of all 15,000 simulations for the following cases: calibration accuracy (*C*); prediction accuracy as determined from *n*–1 jackknifing (*J1*); prediction accuracy as determined from *n*–2 jackknifing (*J2*); and prediction accuracy as determined from conventional split-sample testing (*SS*). The difference between the calibration *C* and *n*–1 jackknifing *J1* statistics reflects the difference between calibration and actual prediction accuracy (as assessed via *n*–1 jackknifing). The difference between the *n*–1 and *n*–2 jackknifing statistics indicates the effect of losing an additional piece of data from the calibration. The difference between the *n*–1 jackknifing statistics and the split-sample values indicates the effect of losing *n*/2 observations in the traditional 50-50 split-sample testing approach. The values in Figures 2 and 3 show several important general trends:

- As the median calibration correlation coefficient *C* increases, the calibration statistics become a more accurate indicator of prediction accuracy. For example, for the high assumed population correlation of 0.9 and *m* = 4, the *n*–1 jackknife prediction correlation *J1* equals 0.844, only 6.2% less than *C*. If the assumed population correlation coefficient is dropped to 0.7, the *n*–1 jackknife prediction correlation *J1* drops to 0.495, or 29.3% less than the calibration value. In other words, the prediction statistics become increasingly poorer than the calibration statistics as the sample size decreases. This fact indicates that the calibration statistics in these cases overestimate the prediction accuracy and shows the importance of the jackknife approach as a more realistic measure of true prediction accuracy for small samples.
- The *n*–1 and *n*–2 jackknifing statistics are similar for all cases studied. This result suggests that the differences between the calibration and *n*–1 prediction statistics are not due to the loss of the one observation, but rather the inability of the calibration statistics to represent prediction accuracy.
- The split-sample statistics are poor measures of prediction accuracy when the samples are small, when the population correlation is low, or both. Split-sample testing is not a reasonable substitute for jackknifing, except possibly for large samples of data.

Determinants of Prediction Accuracy

Jackknifing for model validation is of interest when modeling permanent deformations and other distresses in pavements because large databases are rarely available. Pavement distress models must usually be calibrated and validated using very limited data sets for each combination of pavement characteristics, environmental conditions, and



(a) $\rho = 0.9$

(b) $\rho = 0.8$

(c) $\rho = 0.7$

Figure 2. Comparison of correlation coefficients for calibration statistics (C) and prediction statistics from n–1 jackknifing (J1), n–2 jackknifing (J2), and split-sample jackknifing (SS) at different levels of population correlation coefficients ρ *and for different numbers of pavement sections m per pavement condition cell. (n = 4, N_d = 15, N_d = 3)*

traffic loading. Thus, it is important to know the relative influence of the key parameters on the predicted distresses relative to the prediction accuracy at the end of design life N_d. The key parameters include the number of pavement sections m, the average number of distress measurements per section n, the average duration of the distress measurement time series at each site N_m, the population correlation coefficient ρ, and the desired confidence level γ.

Based on the population model of Equation 4 and the standard error, simulated rutting performance histories were generated using Monte Carlo techniques based on the following stochastic model:

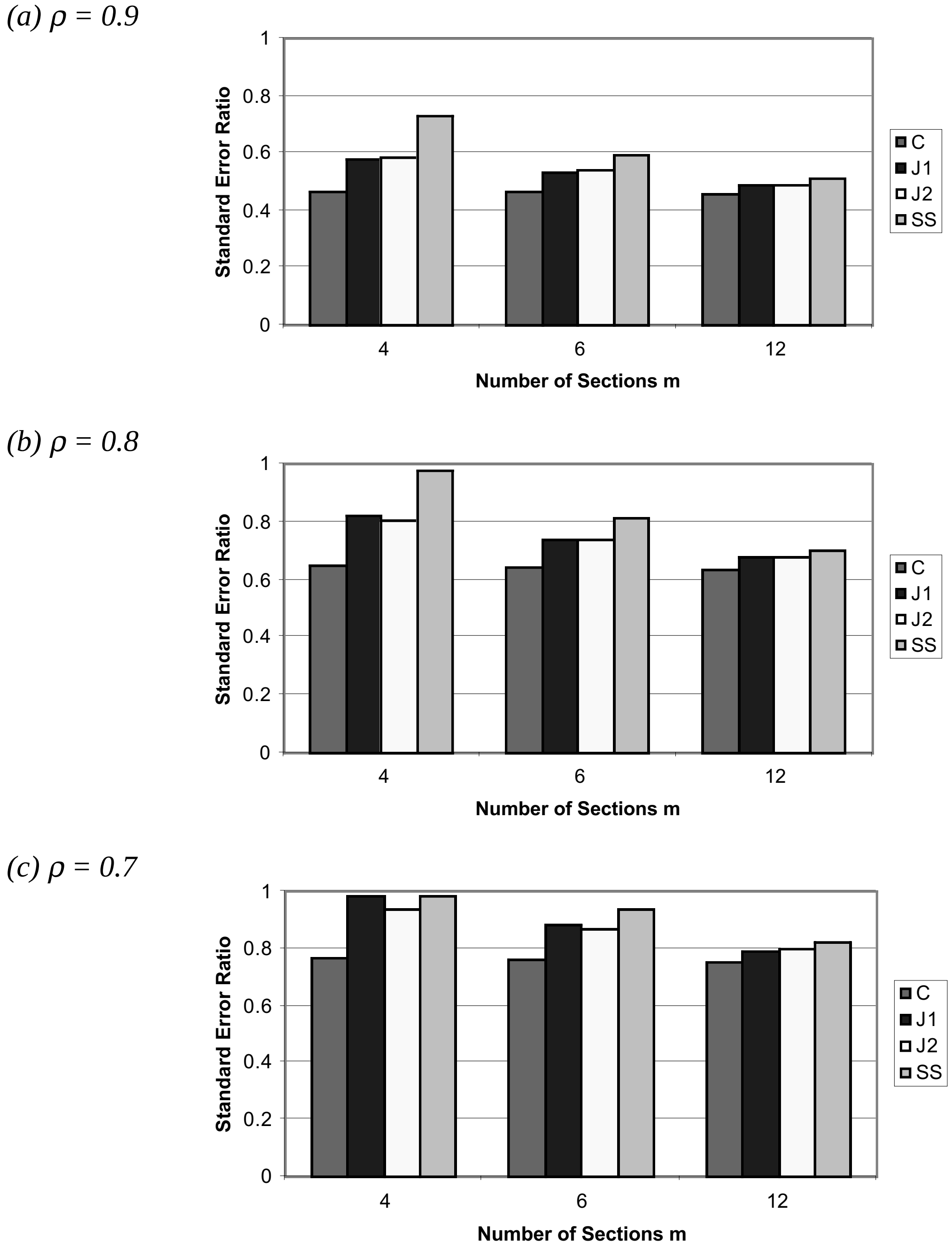


Figure 3. Comparison of standard error ratio for calibration statistics (C) and prediction statistics from n–1 jackknifing (J1), n–2 jackknifing (J2), and split-sample jackknifing (SS) at different levels of population correlation coefficients ρ *and for different numbers of pavement sections m per pavement condition cell. (n = 4, N_d = 15, N_d = 3)*

$$\log D = \log(\beta_0) + \beta_1 \log N + \beta_2 \log S + z \log \sigma_e \qquad (7)$$

in which β_i ($i = 0, 1, 2$) is the assumed population parameters inferred from the MnROAD data, z is a standard normal deviate, and σ_e is the population value of the standard error of estimate. The stiffness S was held constant over time for each pavement section but varied randomly using a normal distribution among the sections. The range for each of the other parameters was selected to span typical ranges that might be reasonable and expected in actual pavement data sets.

The number of available sections per calibration/validation matrix cell (m) ranged from 4 to 24. The average

TABLE 1 Coefficients (c_0, c_1, c_2, c_3) for distress ratio model of Equation 8 as a function of ultimate ESAL, confidence coefficient (γ), and population correlation coefficient (ρ)

MESAL	γ	ρ	c_0	c_1	c_2	c_3
15	60	0.70	0.2351	–0.2945	–0.1883	–0.1468
		0.80	0.1704	–0.2148	–0.1378	–0.1176
		0.90	0.0990	–0.0888	–0.0555	–0.0527
	75	0.70	0.6602	–0.2797	–0.1753	–0.1388
		0.80	0.4789	–0.1972	–0.1255	–0.1084
		0.90	0.2918	–0.0820	–0.05395	–0.04443
	90	0.70	1.3981	–0.2705	–0.1670	–0.1267
		0.80	1.0752	–0.1801	–0.1198	–0.0951
		0.90	0.7486	–0.07616	–0.05137	–0.03558
10	60	0.70	0.1736	–0.2144	–0.1430	–0.1335
		0.80	0.1337	–0.1568	–0.0989	–0.1033
		0.90	0.08789	–0.05953	–0.03713	–0.04607
	75	0.70	0.4830	–0.1962	–0.1287	–0.1215
		0.80	0.3778	–0.1352	–0.0898	–0.09455
		0.90	0.26264	–0.05454	–0.03929	–0.03707
	90	0.70	1.0010	–0.1825	–0.1151	–0.1001
		0.80	0.8404	–0.1151	–0.0920	–0.0782
		0.90	0.6608	–0.04928	–0.04220	–0.03095

number of distress measurements per site (n) ranged from 4 to 8. The desired confidence levels were set at 60%, 75%, and 90%. The population correlation coefficient was varied from 0.7 to 0.9, which reflects explained variances from 50% to 80%.[1] The duration of the distress measurement time series (N_m) was varied between 2 and 6 MESALs. For each combination of parameters, 15,000 samples were generated, and predictions were made for each section at design lives (N_d) of 10 and 15 MESALs. The distributions of the 15,000 predictions were compiled and the upper bounds computed for the 60%, 75%, and 90% confidence levels. The bounds were normalized using the mean design life predictions from the assumed population model for each of the two design lives in terms of the relative distress error ratio (D_R) given in Equation 6.

The Monte Carlo simulation yielded 15,000 distress error ratios for each combination of the six parameters (n, m, N_m, ρ, γ, and N_d). The distress error ratios were then regressed on the parameters of n, m, and N_m using a multiple variable power mathematical model form:

$$D_R = c_0 n^{c_1} m^{c_2} N_m^{c_3} \tag{8}$$

in which c_i (i = 0, 1, 2, 3) are the regression coefficients evaluated using least squares on the logarithms of the four variables. Equation 8 was estimated separately for each value of population correlation coefficient ρ, desired confidence level for prediction accuracy γ, and pavement design life N_d. The fitted coefficients are given in Table 1 for the various values of ρ, γ, and N_d. The correlation coefficients for the fitted models ranged from 0.854 to 0.948 and the standard error ratios ranged from 0.335 to 0.547, both of which suggest good agreement between the predicted and simulated distress error ratios.

The coefficients for n, m, and N_m are all negative, which indicates that the distress error ratio decreases as each of the parameters increases. For most of the equations, the standardized partial regression coefficients for the log-linear models suggest that the number of sections (m) was slightly more important than either the average number of distress measurements per section (n) or the duration of the distress measurement time series (N_m).

Effect of Confidence Level

The desired level of confidence γ for the prediction accuracy is the most critical factor. As shown in Figure 4, the distress error ratio is about 0.1 at a 60% confidence level, 0.25 at a 75% confidence level, and 0.6 at a 90% confidence level. A larger D_R value—i.e., a wider interval—reflects the greater confidence in the accuracy of the predic-

[1] The actual population correlation coefficient for the MnROAD data set as inferred from regression equation 4 was 0.833. Regressions of this model form to other selected field pavement data (not reported here) from the FHWA Accelerated Load Facility (ALF) (Lanes 5, 7, 8, 9, 10, 11, and 12 in the ALF binder study) and WesTrack (Sections 2, 4, 7, 15, 23, and 24) found correlation coefficient values of 0.65 and 0.95, which are on the same order as the value from the MnROAD data set and which suggest reasonable ranges for the population correlation coefficient.

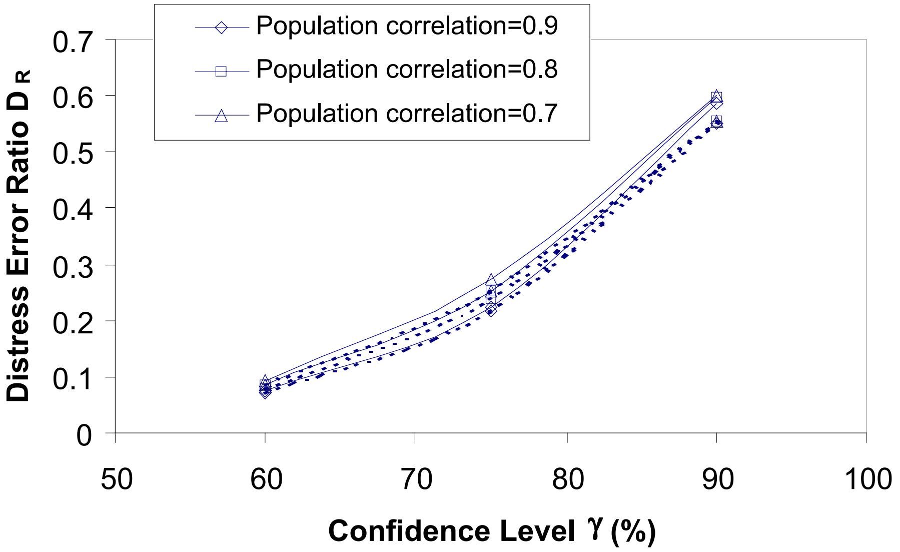


Figure 4. Effect of confidence level γ on the distress error ratio D_R. (m = 5, n = 5, N_m = 3 MESALS; solid lines are N_d = 15 MESALs, dashed lines are N_d = 10 MESALs)

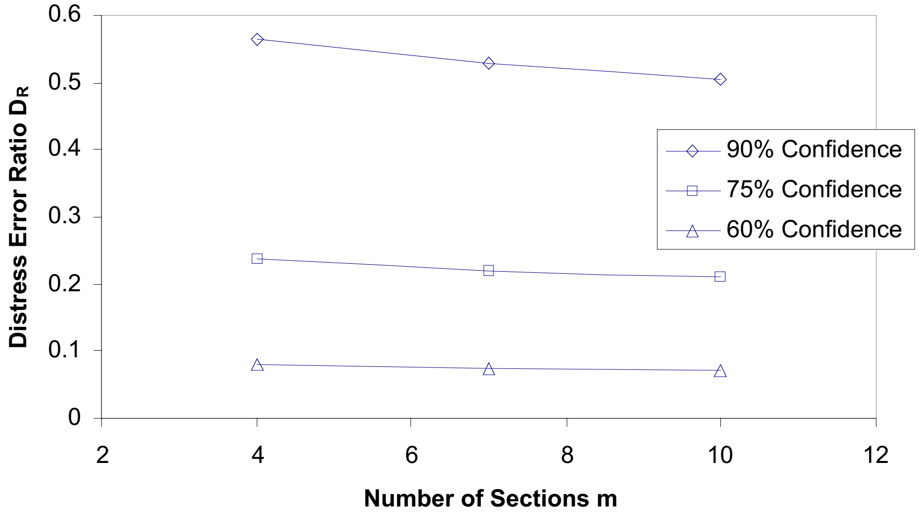


Figure 5. Effect of number of pavement sections m on distress error ratio D_R as a function of desired confidence level γ. (ρ = 0.8, n = 8, N_m = 3 MESALs, N_d = 15 MESALs)

tion. In other words, one can only be 60% confident that the predicted value is within 10% of the "true" value but 90% confident that it is within 60% of the actual. The selection of the desired level of confidence is a policy decision, as it reflects on the cost and safety of projects.

The influence of the other five parameters (Figures 5 through 9) is small relative to the variation associated with the confidence level. However, these other variations are not insignificant. The effects of different parameters may be interactive; therefore, the effect of any one parameter may depend on the magnitudes of the other parameters.

Effect of Number of Sections

The number of pavement sections (*m*) per cell needed to achieve a reasonable level of accuracy is an important decision parameter. Figure 5 shows the variation in the distress error ratio, as a function of *m* for selected values of the other

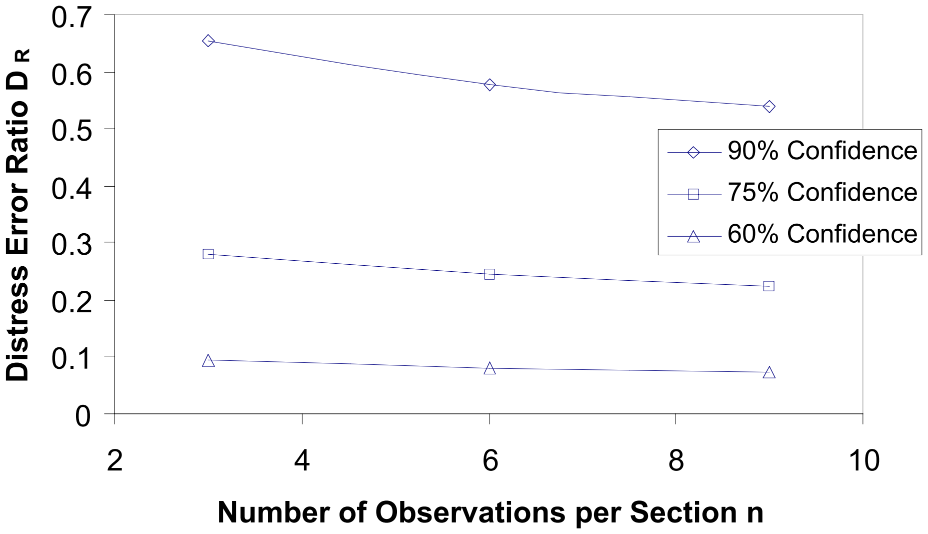


Figure 6. Effect of number of observations per section n on distress error ratio D_R as a function of desired confidence level γ. *(*ρ *= 0.8, m = 5, N_m = 3 MESALs, N_d = 15 MESALs)*

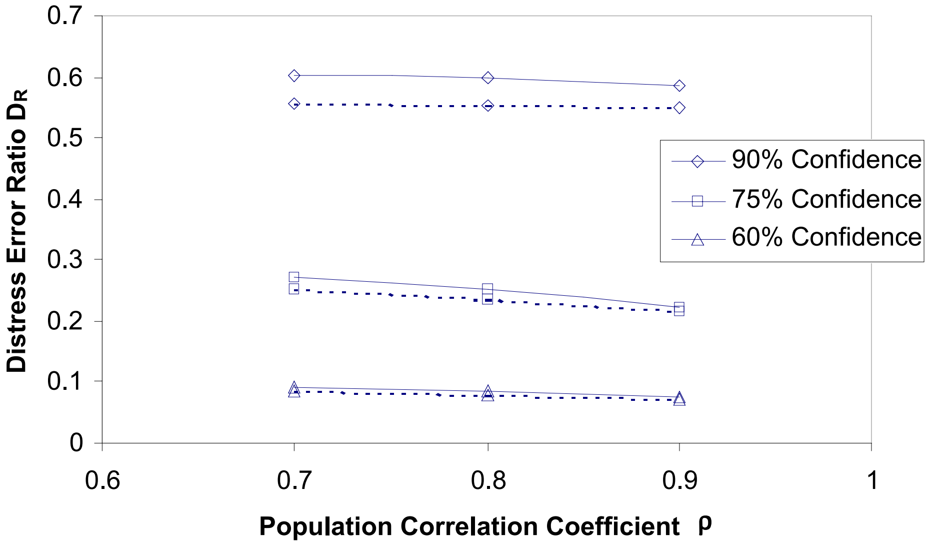


Figure 7. Effect of population correlation coefficient ρ *on distress error ratio D_R as a function of desired confidence level* γ. *(m = 5, n = 5, N_m = 3 MESALs; solid lines are N_d = 15 MESALs, dashed lines are N_d = 10 MESALs)*

parameters. The distress error ratio decreases by between 5% and 18% as *m* increases from 4 to 10 sites. For example, for a 75% level of confidence, a population correlation coefficient of 0.8, a time series duration N_m of 3, and an average number of measurements per section *n* of 8, the distress error ratios are 0.237, 0.221, and 0.211 for values of *m* of 4, 7, and 10, respectively; this trend remains relatively consistent for other combinations of γ, ρ, *n*, N_m, and N_d. The distress error ratio decreases with an increasing number of pavement sections because the uncertainty decreases as the amount of data available increases.

Effect of Number of Distress Measurements

The average number of distress measurements taken at each pavement section (*n*) is also a reflection of the information content of the data available for calibration. In the simulations, the number of distress measurements per section

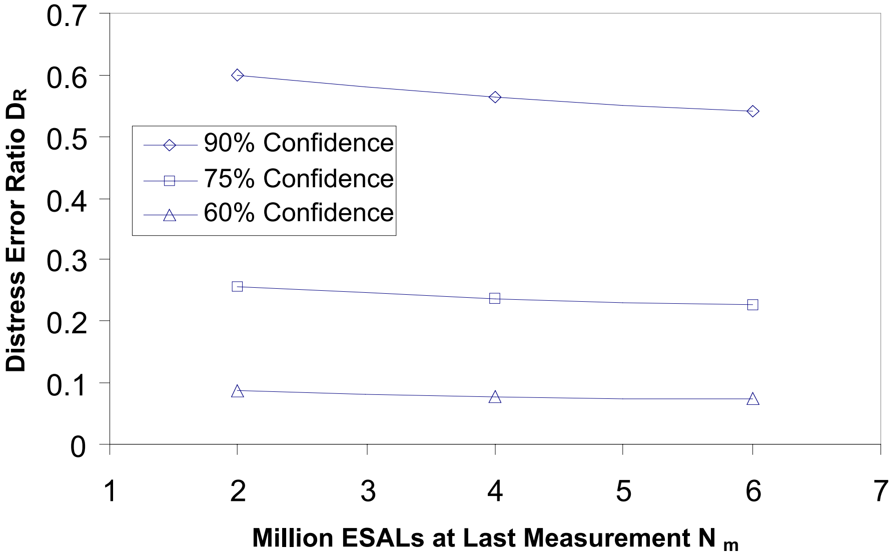


Figure 8. Effect of measurement time series duration N_m on distress error ratio D_R as a function of desired confidence level γ. ($\rho = 0.8$, $m = 5$, $n = 6$, $N_d = 15$ MESALs)

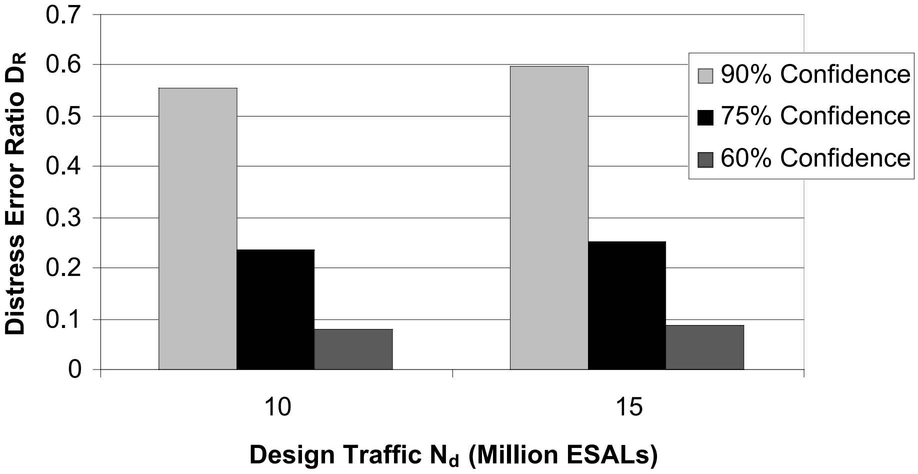


Figure 9. Effect of pavement design life N_d on distress error ratio D_R as a function of desired confidence level γ. ($\rho = 0.8$, $m = 5$, $n = 3$, $N_m = 3$ MESALs)

was varied from 4 to 8, which reflects the range typically available in pavement distress databases (e.g., the long-term pavement performance [LTPP] database). As shown in Figure 6, the distress error ratio decreases by between 8% and 34% as *n* increases from 3 to 9. The actual effect depends to some extent on the values of the other parameters, but the average number of distress measurements for each pavement section appears to be slightly more important than the number of sections for the ranges of values considered in this study.

This observation is confirmed by examining the relative sensitivity of D_R with respect to *m* and *n* (McCuen, 2003). Taking the partial derivatives of Equation 8 for D_R with respect to *m* and *n* yields:

$$\frac{\partial D_R}{\partial m} = c_0 c_2 n^{c_1} m^{c_2 - 1} N_m^{c_3} \tag{9}$$

$$\frac{\partial D_R}{\partial n} = c_0 c_1 n^{c_1 - 1} m^{c_2} N_m^{c_3} \tag{10}$$

For ρ = 0.8, γ = 0.75, and N_d = 15 MESALs, Table 1 gives values c_0 = 0.4789, c_1 = –0.1972, c_2 = –0.1255, and c_3 = –0.1084. Substituting these values into Equations 9 and 10 with N_m = 3 MESALs gives:

$$\frac{\partial D_R}{\partial m} = -0.0677 n^{-0.1972} m^{-1.1255} \tag{11}$$

$$\frac{\partial D_R}{\partial n} = -0.1064 n^{-1.1972} m^{-0.1255} \tag{12}$$

If one further assumes that *m* and *n* are comparable in magnitude (as they often will be in real pavement databases) and thus $m \cong n = k$, then $\frac{\partial D_R}{\partial m} \cong -0.07k^{-1.3}$, which is less (in an absolute value sense) than $\frac{\partial D_R}{\partial n} \cong -0.11k^{-1.3}$. In other words, the distress error ratio is *slightly* more sensitive to the average number of distress measurements per pavement section than to the total number of sections. The overall sensitivity of D_R to *n* is greatest for the poorer correlation coefficients, with variations of D_R over 30% for ρ = 0.7 and only about 9% for ρ = 0.9.

Effect of Population Correlation Coefficient

The population correlation coefficient (ρ) reflects the accuracy of the underlying model. In practice, this would be estimated using sample values obtained from typical databases. For example, the MnROAD database resulted in a correlation coefficient of 0.83. The correlation coefficient is a measure of the error in the predictions, with the error decreasing with increasing ρ.

Figure 7 shows the variation in the distress error ratio for values ρ of 0.7, 0.8, and 0.9. Values are given for design traffic levels N_d of 10 and 15 MESALs for the three target confidence levels. In general, the variation in the distress error ratio with population correlation coefficient is less than 10%, which suggests that for correlation coefficients greater than 0.7 the influence is minimal.

Effect of Distress Measurement Time Series

The duration over which distress measurements are made (N_m) also influences the magnitude of the distress error ratio. Databases that include long distress measurement histories including values that were measured near the pavement design traffic level (N_d) should provide more accurate rutting depth estimates than relatively short records of values taken soon after the pavement was opened to traffic. Therefore, simulations were performed for distress measurement time series durations (N_m) of 2, 4, and 6 MESALs; in each case, it was assumed that the average number of measurements per section (*n*) was distributed uniformly throughout the time series. Note that the largest ESAL value in the MnROAD data set was 3.1 million.

As shown in Figure 8, the distress error ratios decreased by between 8% and 16% as N_m increased from 2 to 6. This is slightly less than the influence of either *m* or *n* on the prediction accuracy. The greatest influence occurs at the lower population correlation coefficients, which is consistent with the lower overall accuracy of the model under these conditions.

Effect of Pavement Design Life

The distress error ratio should decrease as the pavement design life (N_d) increases. This reflects the inevitably lower prediction accuracy associated with extrapolating further beyond the range of the measured distress data. Figure 9 shows the variation in D_R for N_d values of 10 and 15 MESALs at the different target confidence levels. The maximum change in D_R at a given confidence level is about 10% for the parameter ranges examined in this study. The pavement design life thus appears to have less effect on the accuracy of prediction than the other decision parameters have.

Estimating Sample Sizes and Sites

The regression equations for predicting the distress error ratio (Equation 8 and Table 1) can be used to assess the amount of data (number of sections *m* and average number of distress measurements per site *n*) needed to provide any level of prediction accuracy at a specified desired confidence level. As an example, suppose that the duration of the distress measurement time series N_m was limited to 3 by the existing available data, the population correlation coefficient was estimated at 0.8, the design maximum rut depth was 25±6 mm (i.e., $D_R \cong 0.25$) at a pavement design life N_d of 15 MESALs, and policy required a confidence level of 75%. Table 2 summarizes some of the various combinations of *m* and *n* that could achieve the target distress error ratio value of 0.25. These values are not unreasonable for real pavement performance model calibration and validation scenarios.

Summary of the Example

A Monte Carlo simulation model was developed to investigate systematically the effects of the various parameters

TABLE 2 Combinations of *m* and *n* to achieve R_D of 0.25 (ρ = 0.8, γ = 75%, N_m = 3, N_d = 15)

m	n	R_D
3	8	0.246
4	7	0.243
5	6	0.244
6	5	0.247
8	4	0.249

influencing calibration and prediction accuracy for pavement performance models. The parameters studied in the simulations were the population correlation coefficient (ρ); the number of available pavement sections per calibration/validation matrix cell (*m*); the average number of distress measurements per section (*n*); the desired confidence level for prediction accuracy (γ); the duration of the distress measurement time series (N_m); and the pavement design life (N_d). The conclusions from this study were that γ was by far the most important parameter. This was followed (in rough order of decreasing importance) by *n*, *m*, N_m, N_d, and ρ.

Relative sensitivity (McCuen, 2003) is a reliable measure of the relative importance of a predictor variable. For the multiple-power model, a first-order estimate of the relative sensitivity is the magnitude of the exponents of the predictor variables. Therefore, the c_i (i = 1, 2, 3) values of Table 1 can be used directly as measures of the relative importance of the variables. In every case, the coefficient c_i for the average number of distress measurements per section *n* is the largest in magnitude, which indicates that the number of measurements per section is the most important factor. The values of c_2 and c_3 are usually similar in magnitude, although c_2 is more often greater than c_3. Therefore, the number of pavement sections *m* is slightly more important than the total duration of the distress measurement time series N_m, although both are slightly less important than *n*.

The results from this study are obviously limited by the specific conditions analyzed in the simulations and, more importantly, by the assumption that the form and statistics of the rutting simulation model in Equation 1 are fair representations of most pavement conditions. This equation was calibrated using measured performance data from selected sections at the MnROAD experimental facility. This is clearly an open question. However, the performance model calibration completed as part of NCHRP Project 1-37A provides excellent additional data on representative levels of accuracy of the performance models and on the range and variability of the inputs to these models.

REFERENCES

McCuen, R. H., *Modeling Hydrologic Change*, CRC Press, Boca Raton, Florida, 2003.

Miller, R. G., Jr., *The Jackknife—A Review*, Biometrika, Vol. 61, pp. 1–15, 1974.

Mosteller, F., and Tukey, J. W., *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley Publishing Co., Reading, Massachusetts, 1977.

Ross, S. M., *A Course in Simulation*, Macmillan Publishing Co., New York, 1990.

Sobol, I. M., *A Primer for the Monte Carlo Method*, CRC Press, Boca Raton, Florida, 1994.