

5.0 BAYESIAN REGRESSION ANALYSIS

5.1 Bayesian Regression Methodology

5.1.1 Introduction

Predictive equations are very important tools for the pavement management systems. However, databases to support the developments and updating of these models are lacking. These databases are often inadequate in sample size, noisy, or incomplete. Conventional statistical modeling tools, such as classical regression analysis, may have limited success in these applications (*Kajner et al. 1996*). A promising solution lies in the use of Bayesian regression, which explicitly allows experts to be used to supplement poor quality data (*Kwaeski and Nickeson 1997*). Bayesian regression methodology was adopted by the Canadian Strategic Highway Research Program (C-SHRP) for the Canadian Long Term Pavement Performance (C-LTPP) monitoring program. Nesbit and Sparks (*1990*) discussed the complete rationale for employing the Bayesian approach for the C-LTPP program in the report "Design of Long Term Pavement Monitoring System for the Canadian Strategic Highway Research Program."

5.1.2 An Overview of the Bayesian Regression Approach

In its simplest sense, Bayesian regression is a specialized adaption of the Bayes' Theorem involving development of multivariate regression models which explicitly consider two disparate sources of information:

1. A prior information, i.e. information that is known prior to an experiment, and
2. Experimental data, i.e. information that is derived from an experiment.

The interpretation and conclusion drawn from the experimental data can be quite different depending on what other evidence exists on the subject at hand. However, this difference in

interpretation does not simply mean biasing a result. Interpretation of results using Bayes' Theorem is a mathematically consistent way to interpret new evidence/information (*Kwaeski and Nickeson 1997*).

The Bayesian statistical method for model development, represented in Figure 5.1, is to systematically combine prior knowledge and experience with data to improve the predictive relationship. The Bayes approach calculates a meaningful and credible answer without relying solely on a small database. In doing so, the Bayes technique allows decisions to be made in the short term while improvements to the data, judgement and the model continue to be made (*Kwaeski and Nickeson 1997*).

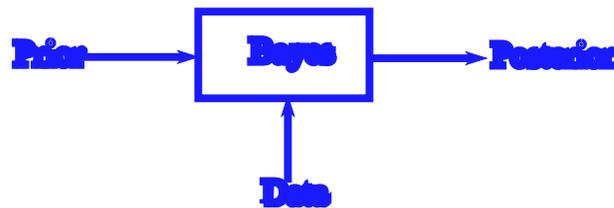


Figure 5.1 The Bayesian Statistical Approach (*Kwaeski and Nickeson 1997*)

In assembling information for Bayesian regression, data collected in the traditional manner is supplemented with prior knowledge. This approach is summarized in the Figure 5.1. The so-called 'prior' may be drawn from expert judgement, "old" data sets, or knowledge that is generally accepted in the field. Expert judgement can also be encoded by polling experts and asking them to estimate the value of the dependent variable for a combination of contributory variables. Once collected, the experts' observations are interpreted similar to the traditional data.

5.1.3 Bayesian Regression Software

Two Bayesian regression software packages, B-STAT and XLBayes, were developed by VEMAX Management, Inc., Canada, under contract to C-SHRP. B-STAT provides an EXCEL spreadsheet interface to a FORTRAN based Bayesian regression program, PC-BRAP. XLBayes, on the other hand, is a much faster Bayesian regression program based entirely in the EXCEL environment (*Kwaeski and Nickeson 1997*). The analysis features and numerical results of the two programs are identical. XLBayes was selected for this research because it is relatively straightforward and faster.

5.2 Bayesian Regression to Predict the Decrease in PSE Values

The Bayesian regression analysis using the XLBayes software requires prior data to be combined with the sample data to obtain the desired posteriors. The prior data can be drawn from the expert judgement, old data sets or knowledge that is generally accepted in the field. For this research project, the data set for a number of pavements from Districts I and IV for 1993 and 1994 were used as prior data, and the data for 1995 were used as the sample data. The same functional form and transformations of the independent variables as in the classical regression were used.

5.2.1 Developing Prior and Assembling Sample Data

The prior can be derived either subjectively using expert judgement or objectively based on existing data or models. Both approaches require that the prior information be put into either an N-prior or G-prior format. Both the N-prior or G-prior summarize a linear regression which represents the prior state of knowledge in the Bayesian regression calculation. The prior includes the coefficients of the linear regression equation along with the corresponding regression statistics such as the variance of the regression coefficients. The regression statistics indicate the certainty of the prior and are used to weigh the balance between the prior and the data in the Bayesian regression calculation. A brief overview of the information required to define the N-prior or a G-prior is provided in Table 5.1 (*Kwaeski and Nickeson 1997*). The G-prior option is typically used when the

coefficient means have been estimated directly by the experts. The G-prior derives the variance/covariance matrix for the coefficient means based on a set of independent variable data. The G-prior factor is used to increase or decrease the influence of the prior in the calculation of the posterior. The G-prior factor is denoted by g . A typical value for g is 1. This essentially gives the prior variance/covariance matrix weight equal to that of the experimental data. The greater the value of g , the more influence the prior will have on the posterior. Since the pseudo/prior data used in this research were not derived from expert opinion only, the N-prior option of Bayesian regression was used in this analysis.

Table 5.1 Required Prior Information (After Kwaeski and Nickeson 1997)

Prior Information	Required for N-prior	Required for G-prior
Means vector		
Variance/Covariance Matrix		-
G-prior data set	-	
G-prior factor	-	
Residual variance		
Degrees of freedom		

5.2.2 Results of Bayesian Regression and Selected Posterior Models

The classical regression results using pseudo data, development of the N-prior and the posterior regression coefficients for the FDBIT and PDBIT pavements have been reported in detail by Chowdhury (1998). The selected posterior models using N-prior Bayesian regression analysis are shown below.

FDBIT Pavements: The selected models for FDBIT pavements are :

Distress Level 1

$$PSE = 0.123 * (AGE)^{1.5} - 9.329 * \exp[SN] + 0.106 * TH + 0.374 * PSE + 5.89 * DL1$$

(5.1)

Distress Level 2

$$PSE = 0.123 * (AGE)^{1.5} - 9.329 * \exp[SN] + 0.106 * TH + 0.374 * PSE + 6.04 * DL2 \quad (5.2)$$

For Distress Level 3

$$PSE = 0.123 * (AGE)^{1.5} - 9.329 * \exp[SN] + 0.106 * TH + 0.374 * PSE + 6.47 * DL3 \quad (5.3)$$

PDBIT Pavements: The selected models for PDBIT pavements are :

Distress Level 1

$$PSE = 0.021 * (AGE)^{1.5} - 1.873 * \exp[SN] + 0.303 * PSE + 0.392 * DL1 \quad (5.4)$$

Distress Level 2

$$PSE = 0.021 * (AGE)^{1.5} - 1.873 * \exp[SN] + 0.303 * PSE + 0.881 * DL2 \quad (5.5)$$

Distress Level 3

$$PSE = 0.021 * (AGE)^{1.5} - 1.873 * \exp[SN] + 0.303 * PSE + 1.974 * DL3 \quad (5.6)$$

where, PSE= Predicted *decrease* in PSE value,
AGE= Age of the pavement *since the last rehabilitation action* (in years),
TH = AC layer thickness (in inches),
PSE= PSE value assigned to the pavement *immediately after the last action*,
SN= Decrease in structural number, and
DL_i= Distress level due to transverse cracking (i = 1, 2, 3).

5.3 Model Evaluation

The purpose of evaluating the model results is to draw conclusions about the Bayesian posterior results. Evaluation emphasizes comparisons between the data, the prior, and the posterior. These comparisons may be used for additional iterations for analysis later on. The statistical performance of a classical regression model is typically measured by evaluating the standard error (S_e), coefficient of determination (R^2), F-statistic, and t-statistic. In Bayesian regression, only S_e and t-statistic can be evaluated. Neither R^2 nor the F-statistic can be calculated because they rely on the

experimental data which does not exist for the posterior results (*Kaweski et al 1997*).

5.3.1 Data, Prior, and Posterior PDF Plots

An important output of XLBayes is the PDF (Probability Density Function) plots for each coefficient in the model. These plots graphically compare the distribution of the same coefficient when based on the data alone, the prior alone, or the Bayesian posterior. Figures 5.2 through 5.14 show the PDF plots for all coefficients in the models developed in this study.

Under the assumptions of both classical linear regression and the Bayesian regressions, the model coefficients follow t-distribution. The width of the bell shaped curve shows the confidence in the estimating coefficients. The PDF plots of all coefficients reveal the fact that the probability distribution for the posterior estimate is 'tighter' than either the prior or the data. This is intuitively reasonable as the prior and the data reinforce each other with similar estimates of the coefficients. Bayesian regression models can always be updated by inserting more data in the model which makes the posterior more and more definitive.

5.3.2 t-Statistic

The t-test is used to determine whether a regression coefficient is significantly different from zero. The t-value for a regression coefficient is calculated by dividing the mean of the regression coefficient by its standard deviation:

$$t = b_i / s_{b_i}$$

The null hypothesis in this test is :

$$H_0 : b_n = 0$$

which is tested against the alternative hypothesis :

$$H_1 : b_n \neq 0$$

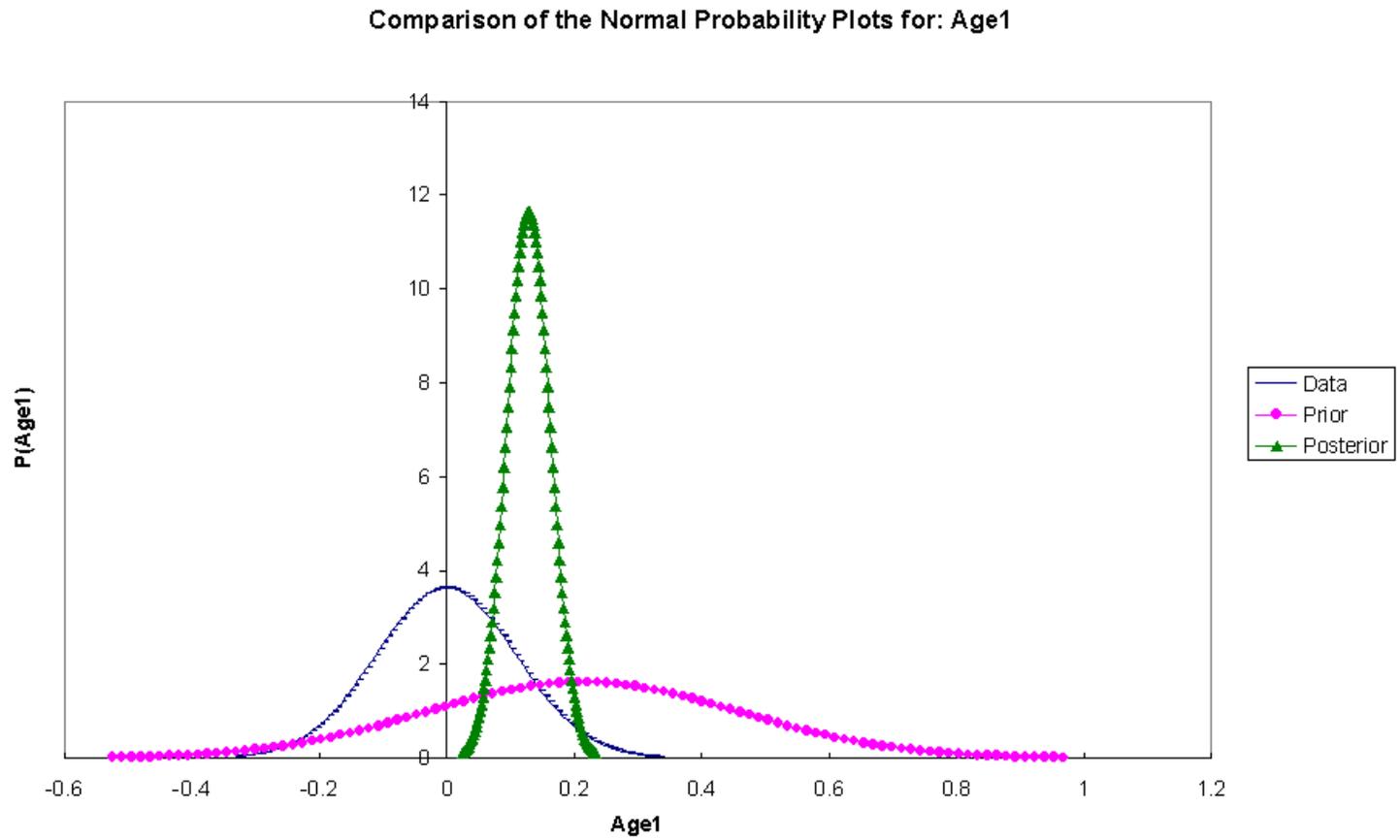


Figure 5.2 PDF Plot for Age for FDBIT Pavements

Comparison of the Normal Probability Plots for: Th

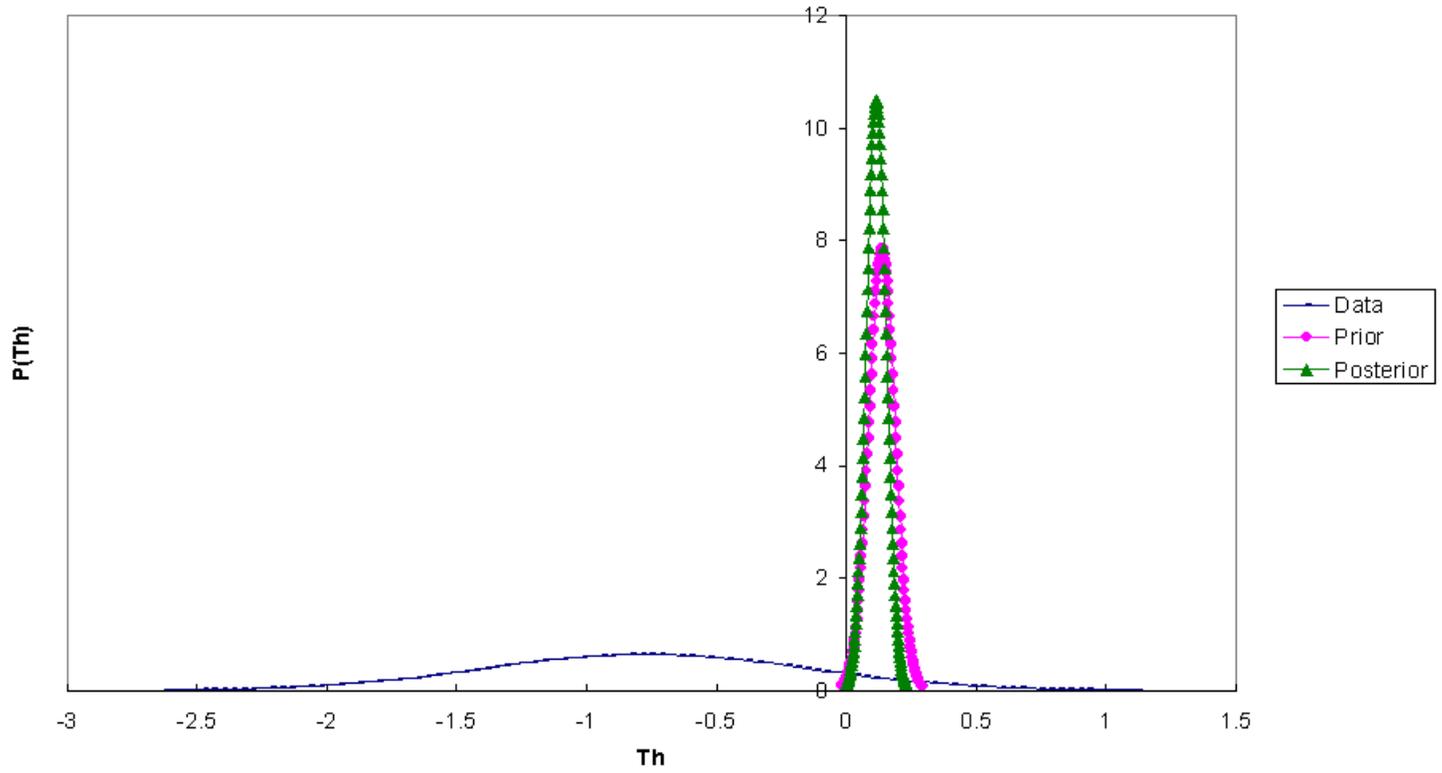


Figure 5.3 PDF Plot for Thickness for FDBIT Pavements

Comparison of the Normal Probability Plots for: Exp(dSN)

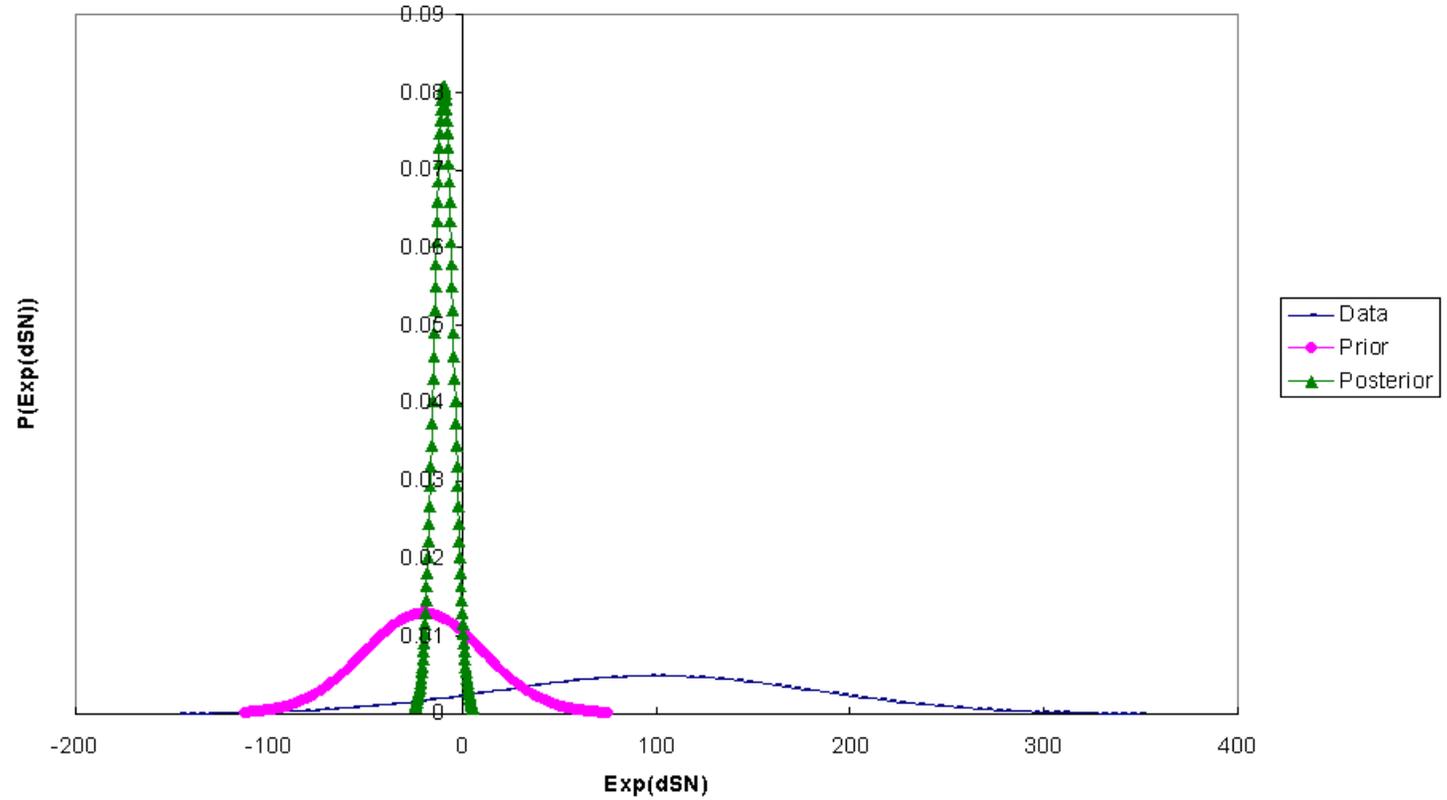


Figure 5.4 PDF Plot for Decrease in Structural Number for FDBIT Pavements

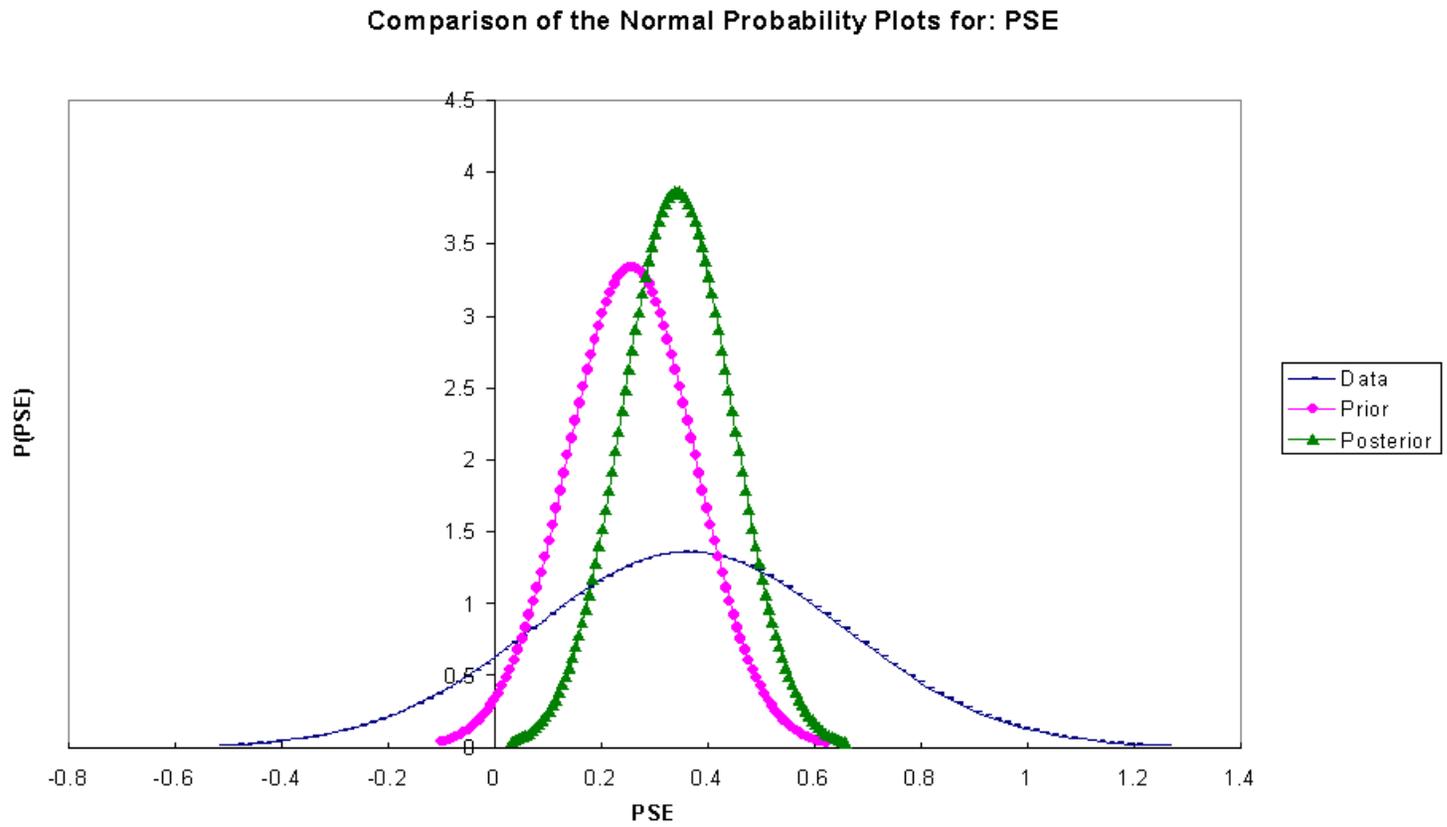


Figure 5.5 PDF Plot for PSE for FDBIT Pavements

Comparison of the Normal Probability Plots for: DL1

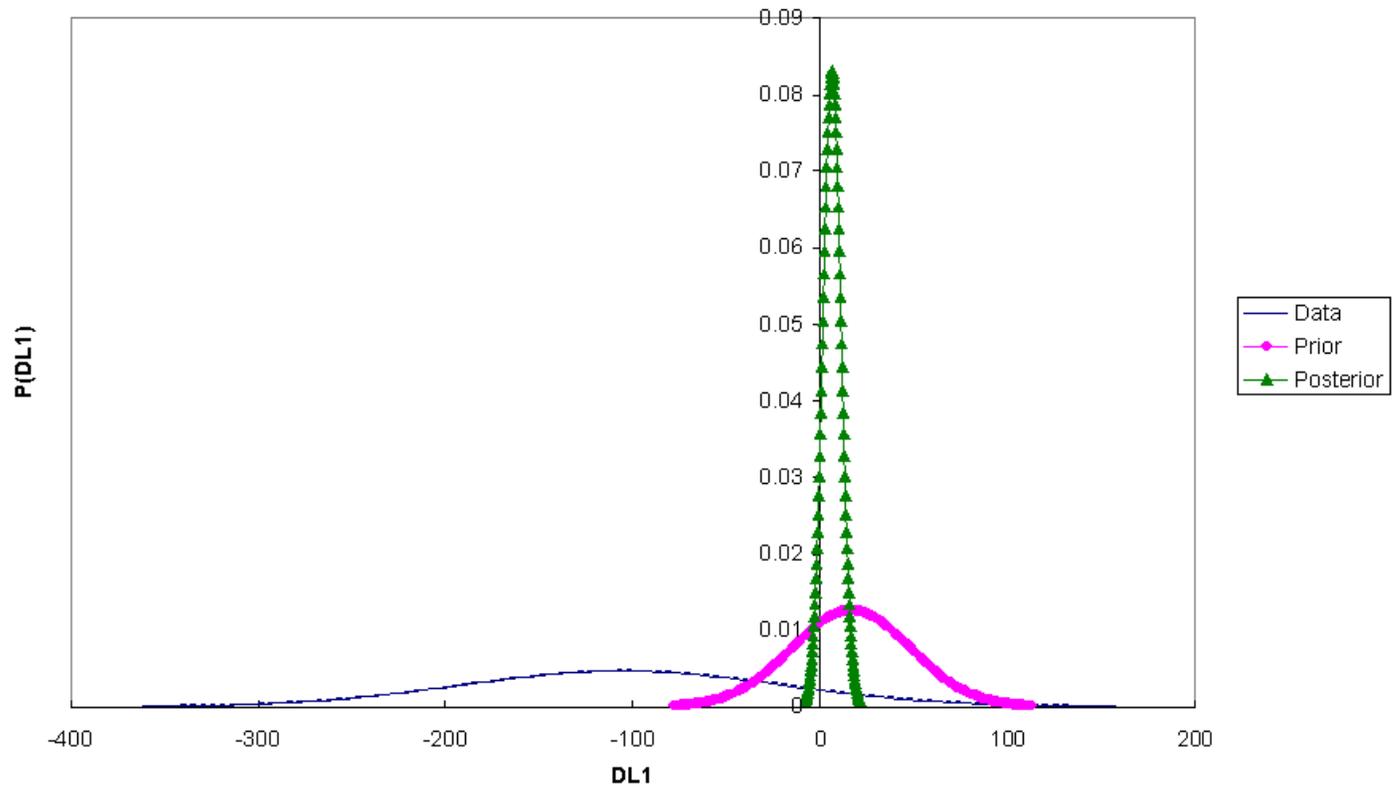


Figure 5.6 PDF Plot for Distress Level 1 for FDBIT Pavements

Comparison of the Normal Probability Plots for: DL2

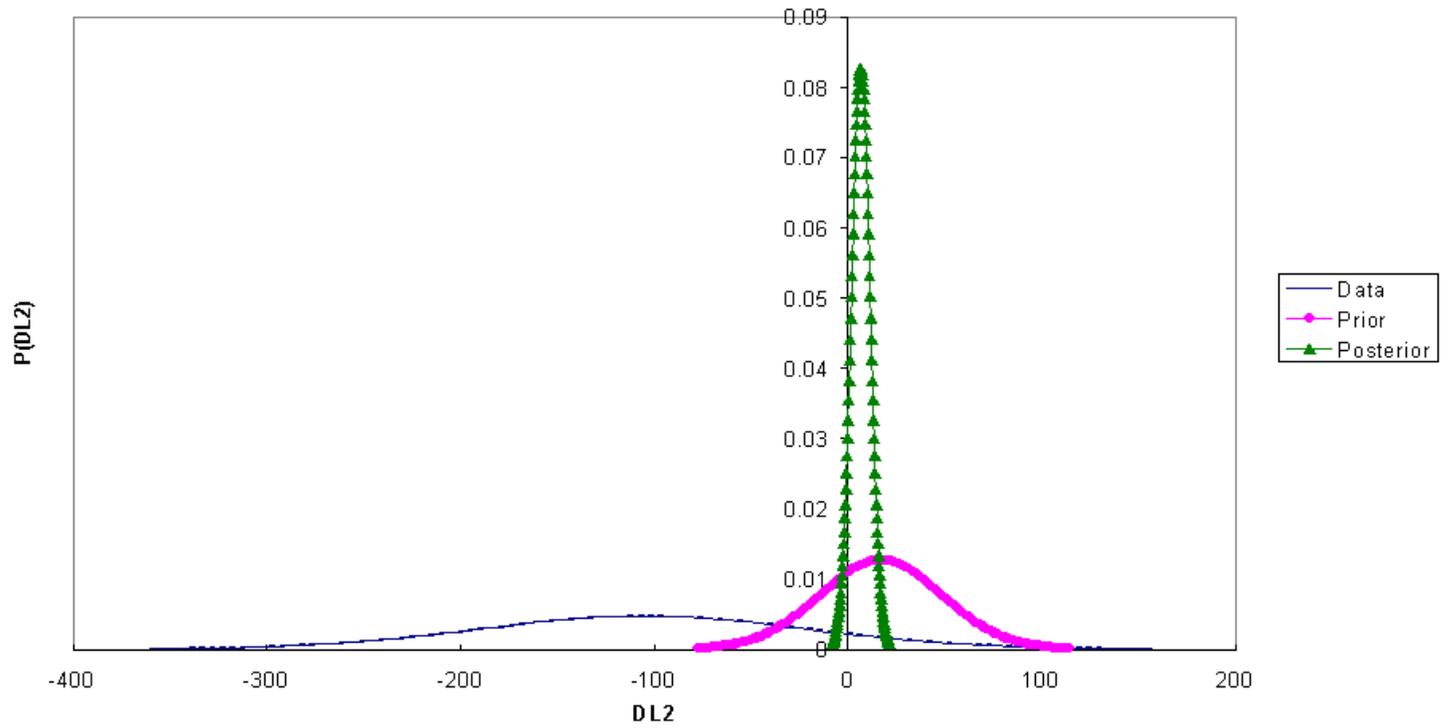


Figure 5.7 PDF Plot for Distress Level 2 for FDBIT Pavements

Comparison of the Normal Probability Plots for: DL3

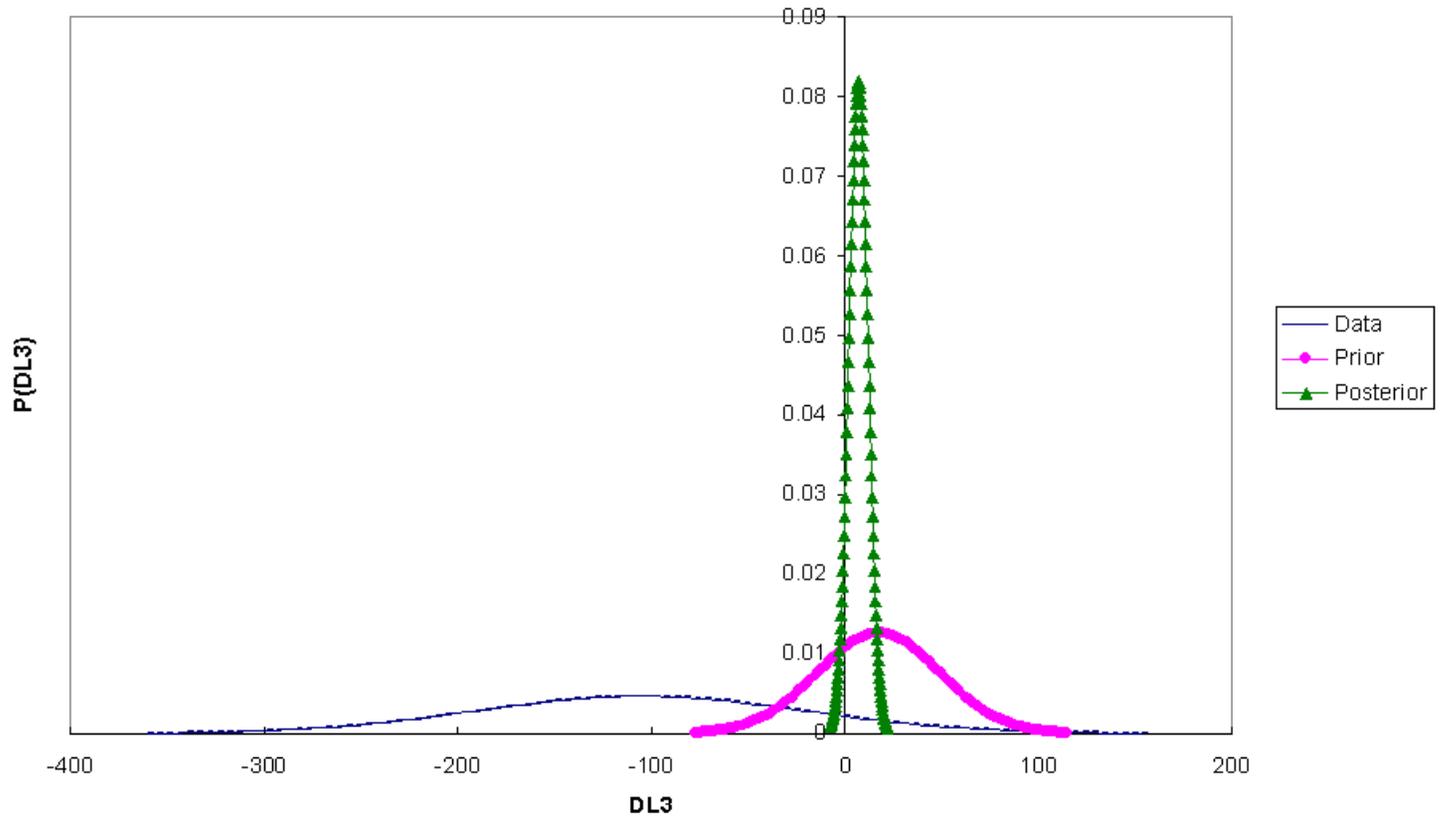
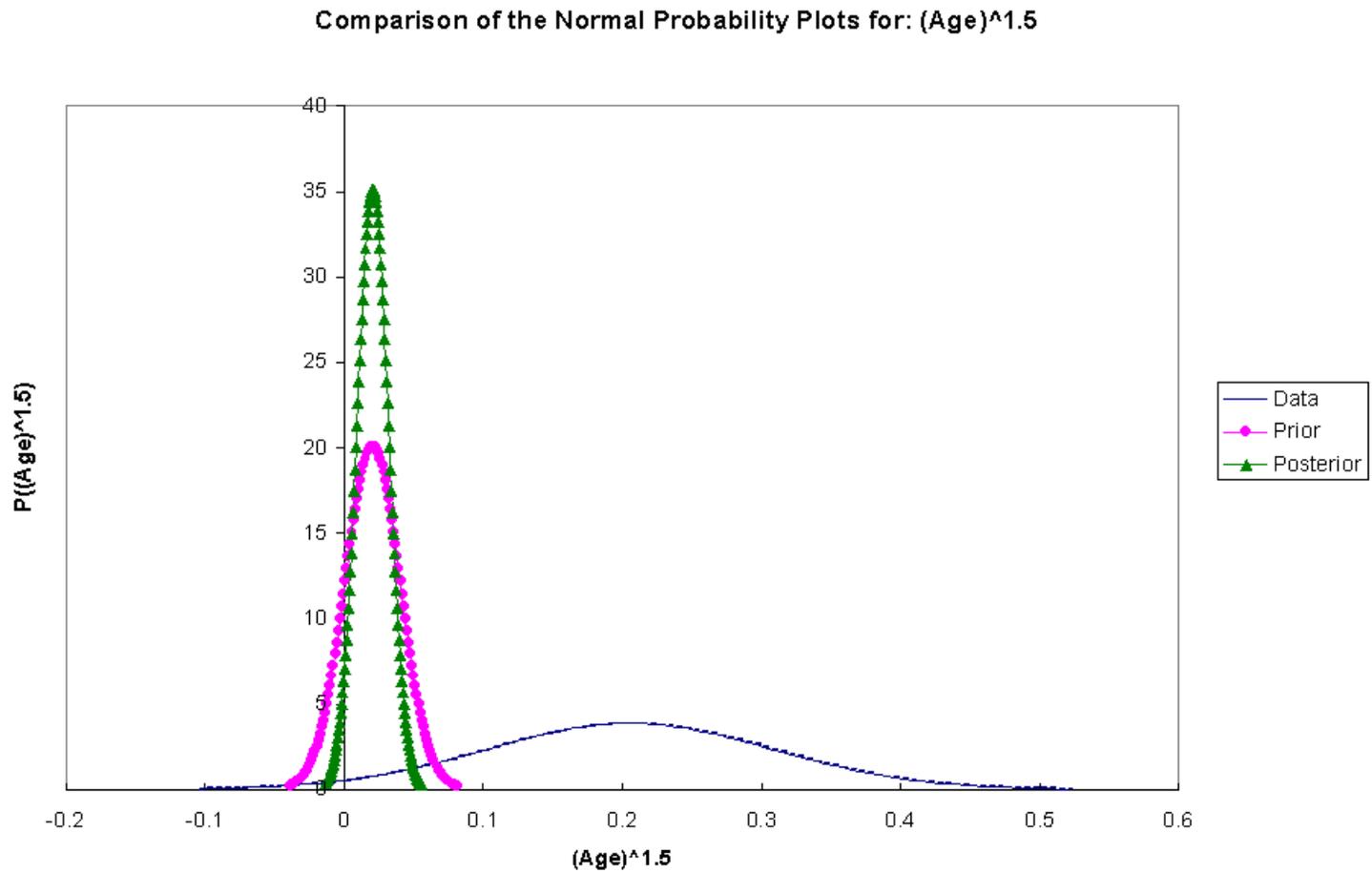


Figure 5.8 PDF Plot for Distress Level 3 for FDBIT Pavements



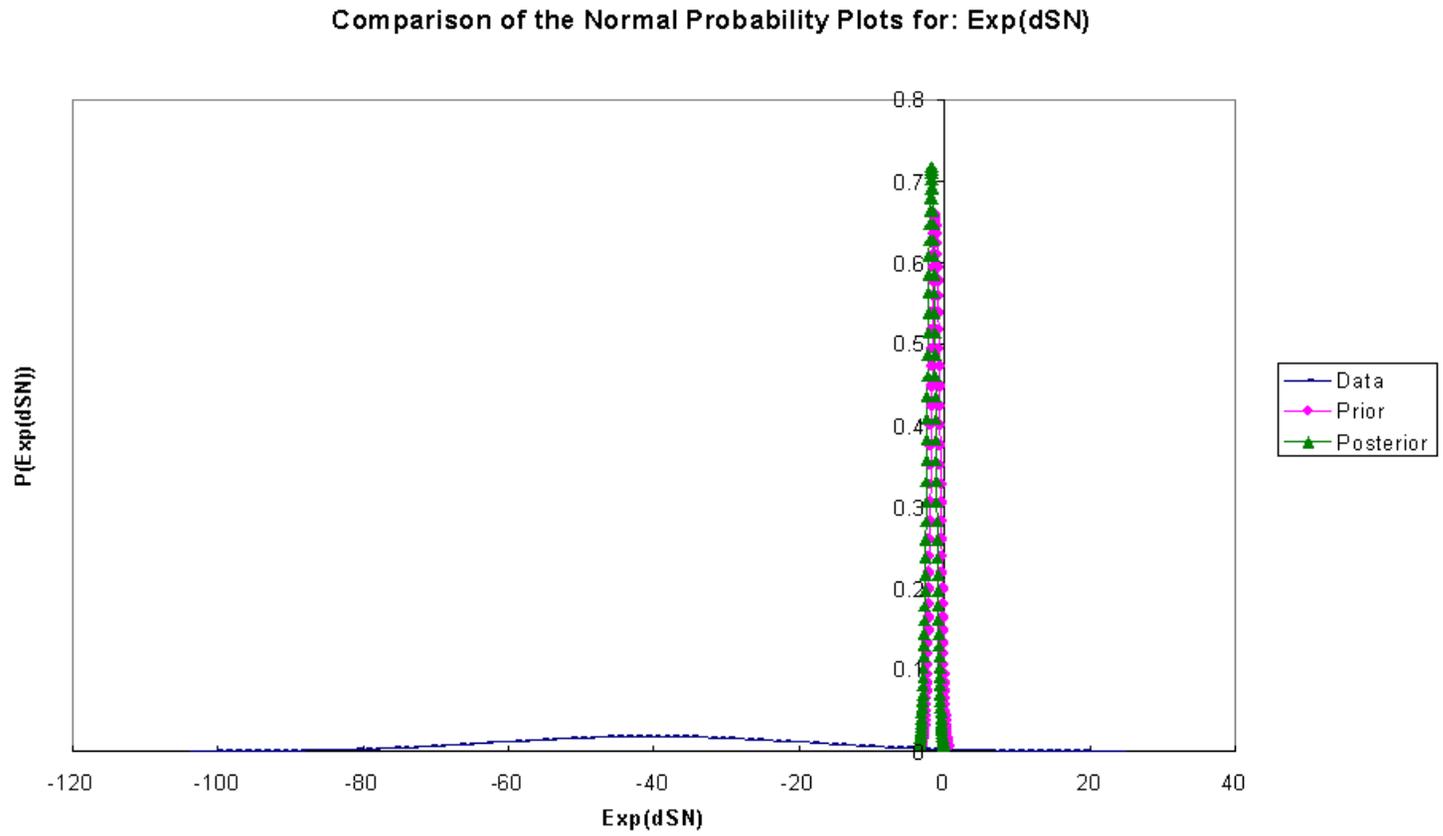


Figure 5.10 PDF Plot for Decrease in Structural Number for PDBIT Pavements

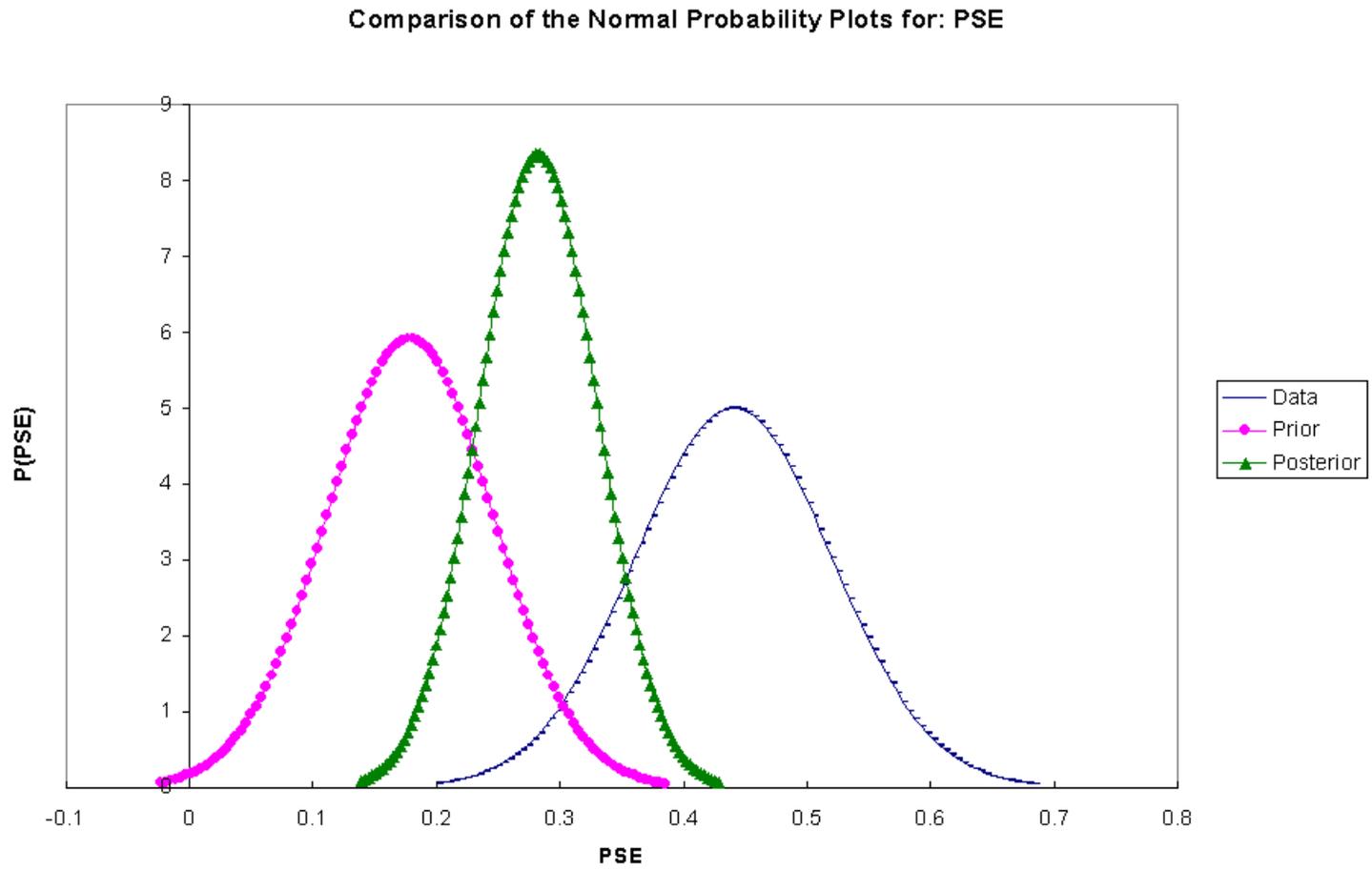


Figure 5.11 PDF Plot for PSE for PDBIT Pavements

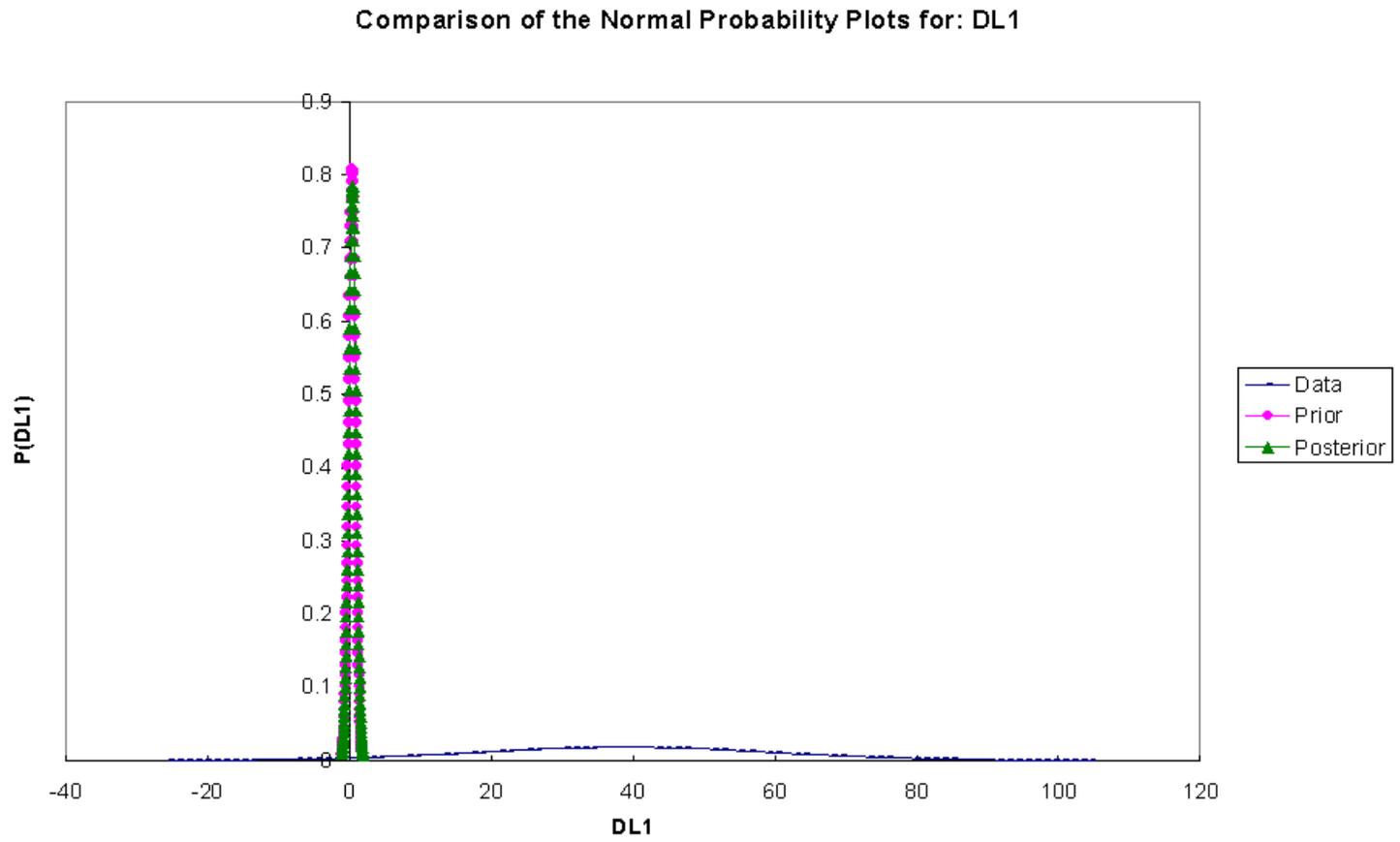


Figure 5.12 PDF Plot for Distress Level 1 for PDBIT Pavements

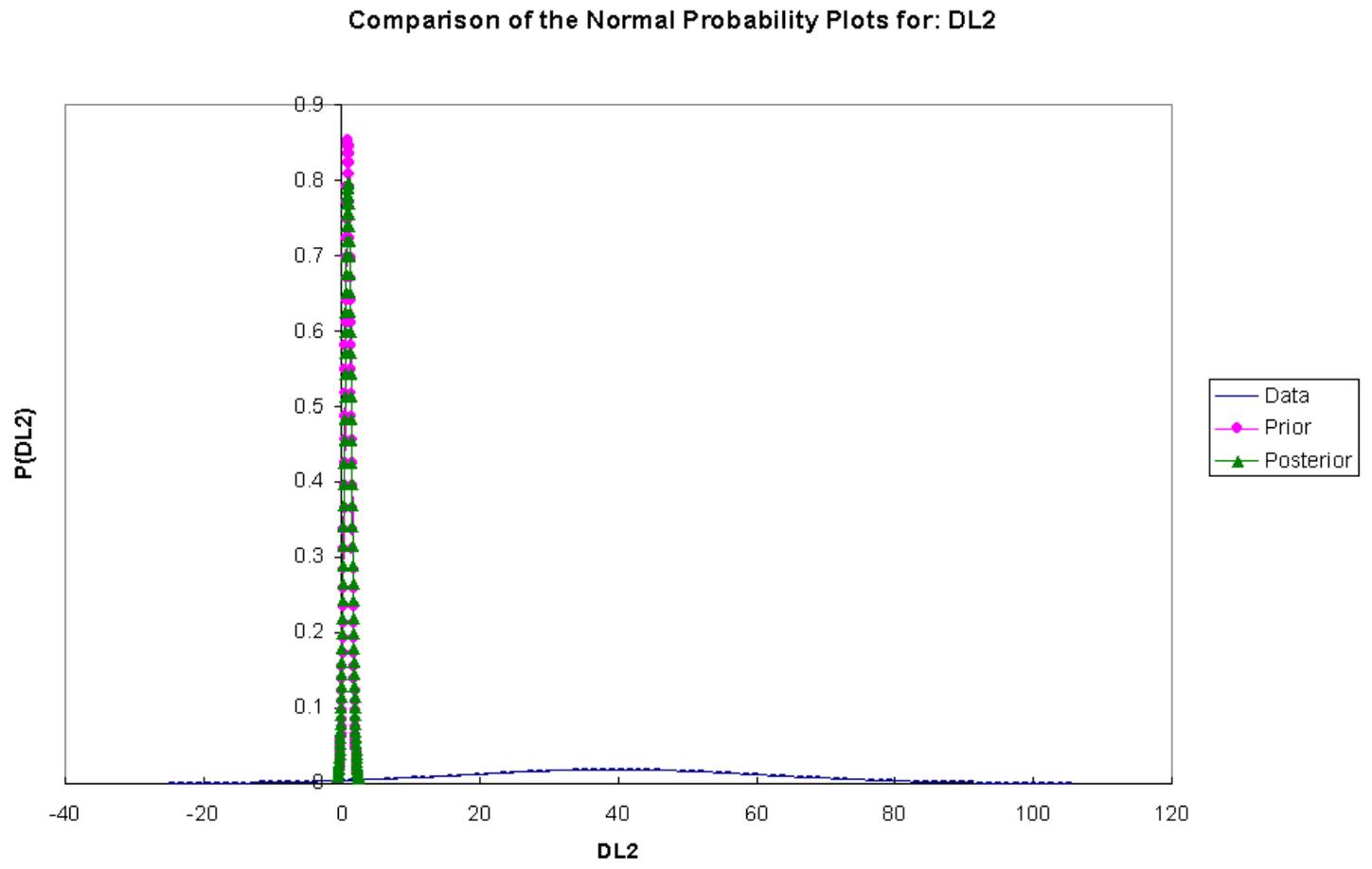


Figure 5.13 PDF Plot for Distress Level 2 for PDBIT Pavements

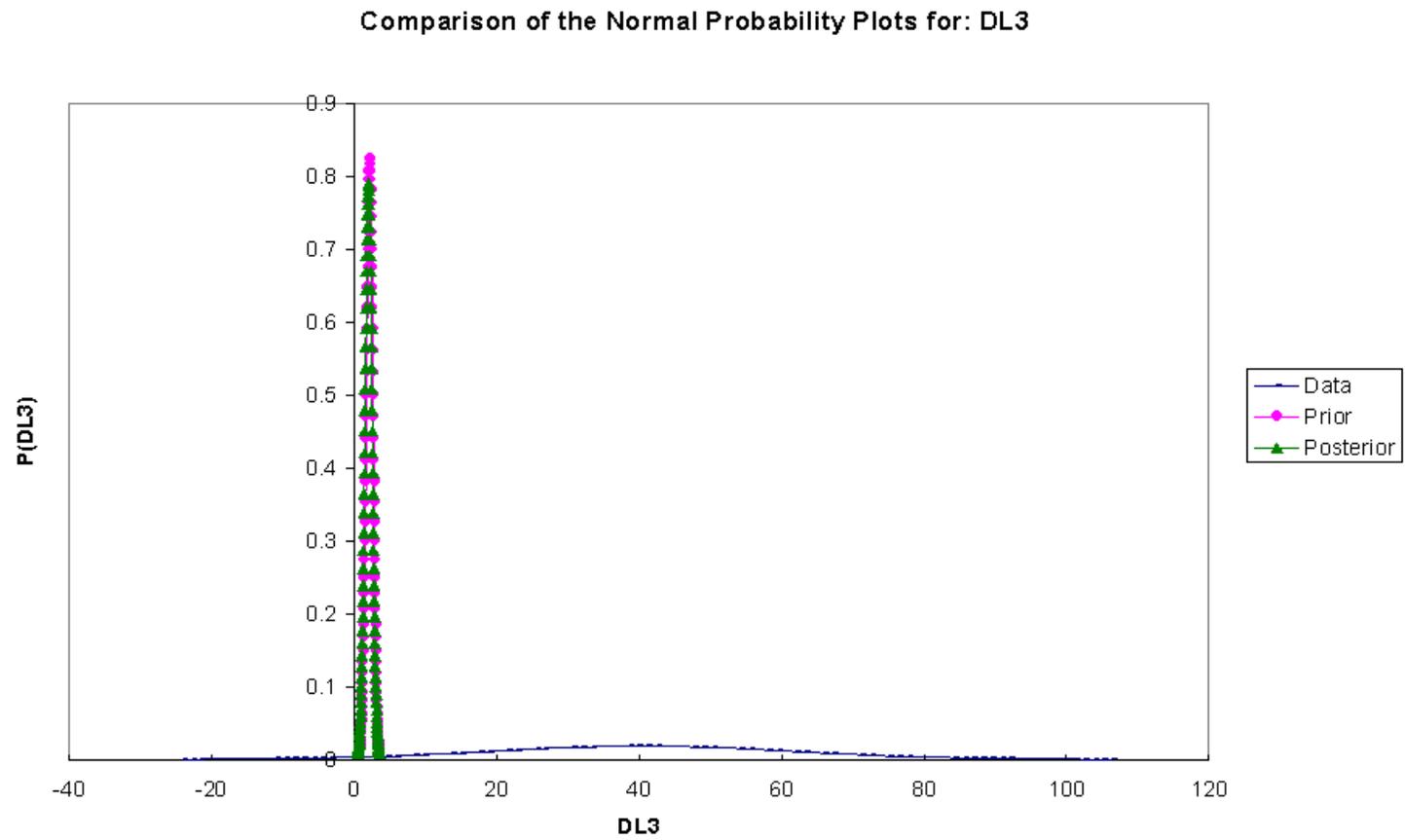


Figure 5.14

PDF Plot for Distress Level 3 for PDBIT Pavements

At 5% level of significance, where the number of degrees of freedom is very large (i.e., the t distribution is approximately the same as the normal distribution), the critical value of t is ± 1.96 . If the t -value is greater than 1.96 or less -1.96, the null hypothesis is rejected and it is accepted that the estimate of b_n is statistically significant. The higher the value of t , the more is the confidence about its value and significance. If the t -value is between 1.96 and -1.96, the null hypothesis is accepted and it is concluded that the estimate of b_n is not statistically significant. The values calculated for the coefficients may only be different from zero due to chance. If the regression coefficients in the prior and posterior are not statistically significant it may be useful to re-run the analysis after excluding the variable in question. If the standard error term does not increase significantly, the excluded variable may not be a statistically significant contributory variable.

The ideal result is for the data and prior to reinforce each other, resulting in a posterior coefficient that has a smaller standard error than either one individually. This is not always the case, however, and the posterior may in fact have a larger standard error. Irrespective of how much the variance has changed, it is desirable that the coefficients in the posterior model all be statistically significant.

The t -statistics and the standard deviations of different coefficients are presented in Table 5.8. It is observed that the t -statistics of all selected variables are outside the range of 1.96 and -1.96 which means that the null hypothesis is rejected in all cases. Thus, the variables used in the models are significant at 5% level of significance.

5.3.3 Standard Error of the Residuals (S_e)

The standard error of the residuals, S_e , is a basic measure of regression model performance. The standard error (or standard deviation) of the residuals is simply the square root of the residual variance, S_e^2 . The lower the S_e , the closer the predictions made by the model are to the actual

Table 5.2 Standard Deviation and t-Statistic of the Posterior Coefficients

Pavement type	Variable	Std. Deviation	t-value	Res. Var. (S_e^2)
FDBIT	(Age) ^{1.5}	0.034	3.620	0.329
	Thickness	0.041	2.547	
	Exp[(SN)]	4.240	-2.200	
	PSE	0.107	3.486	
	DL1	2.979	1.98	
	DL2	2.876	2.101	
	DL3	2.424	2.670	
PDBIT	(Age) ^{1.5}	0.008	2.349	0.203
	Exp[(SN)]	0.500	-3.746	
	PSE	0.038	7.850	
	DL1	0.196	1.990	
	DL2	0.383	2.301	
	DL3	0.466	4.234	

observations of the dependent variable, and therefore, the better the model.

Under the assumptions of regression, the residual has a mean of zero and is normally distributed. Thus the confidence interval for the forecasts made by the model can be calculated using a table of areas under the standard normal curve. For example, 95% confidence interval for a forecast corresponds to the mean forecast plus or minus 1.96 times the standard deviation of the residual. Therefore, the selected models will predict the (PSE) values within ± 1.1 units of actual ratings for FDBIT and ± 0.88 units for PDBIT pavements with 95% confidence.