International Journal of Architectural, Civil and Construction Sciences
ISSN: 2415-1734
Vol:1, No:2, 2007

Statistical Modeling of Accelerated Pavement Failure Using Response Surface Methodology

Anshu Manik, Kasthurirangan Gopalakrishnan, and Siddhartha K. Khaitan

***Abstract*—**Rutting is one of the major load-related distresses in airport flexible pavements. Rutting in paving materials develop gradually with an increasing number of load applications, usually appearing as longitudinal depressions in the wheel paths and it may be accompanied by small upheavals to the sides. Significant research has been conducted to determine the factors which affect rutting and how they can be controlled. Using the experimental design concepts, a series of tests can be conducted while varying levels of different parameters, which could be the cause for rutting in airport flexible pavements. If proper experimental design is done, the results obtained from these tests can give a better insight into the causes of rutting and the presence of interactions and synergisms among the system variables which have influence on rutting. Although traditionally, laboratory experiments are conducted in a controlled fashion to understand the statistical interaction of variables in such situations, this study is an attempt to identify the critical system variables influencing airport flexible pavement rut depth from a statistical DoE perspective using real field data from a full-scale test facility. The test results do strongly indicate that the response (rut depth) has too much noise in it and it would not allow determination of a good model. From a statistical DoE perspective, two major changes proposed for this experiment are: (1) actual replication of the tests is definitely required, (2) nuisance variables need to be identified and blocked properly. Further investigation is necessary to determine possible sources of noise in the experiment.

***Keywords*—**Airport Pavement, Design of Experiments, Rutting, NAPTF.

I. Introduction

DISTRESS modes normally considered in bituminous or Hot-Mix Asphalt (HMA) pavement analysis and design are fatigue cracking, rutting, and low-temperature cracking [1]. Rutting is a major load-related distress in airport flexible pavements [2]-[3]. It appears as longitudinal depressions in the wheel paths and may be accompanied by small upheavals to the sides. Permanent deformation in any or all of the pavement layers and/or subgrade under repeated traffic loading contributes to the total accumulation of pavement surface rutting. Depending on the magnitude of the load and the relative strength of the pavement layers, a significant portion of the total rutting can occur in the pavement foundation due to weak subgrade or the use of a low quality aggregate base [4]. Significant rutting can lead to major structural failure of the pavement.

Rutting in paving materials develop gradually with an increasing number of load applications, usually appearing as longitudinal depressions in the wheel paths accompanied by small upheavals to the sides [5]. To limit pavement surface rutting to acceptable levels, the various paving layers and the subgrade must be given careful attention during the design process.

Rutting makes it difficult to steer the wheels as the wheels would tend to follow the rutted path. This becomes dangerous in case of airport runways if the runway has even slight amount of rutting due to the large size and high speeds of the airplanes. A rut depth of 1 inch in airport pavements is considered to indicate functional failure due to the ponding it can cause in the presence of water resulting in hydroplaning of the aircraft wheels. Significant research has been conducted to determine the factors which affect rutting and how they can be controlled.

Design of Experiments (DoE) refers to experimental methods used to quantify indeterminate measurements of factors and interactions between factors statistically through observance of forced changes made methodically as directed by mathematically systematic tables. Thus, it is a structured, organized method for determining the relationship between factors (*Xs*) affecting a process and the output of that process (*Y*). Using the experimental design concepts, a series of tests can be conducted while varying levels of different parameters, which could be the cause for rutting in airport flexible pavements. If proper experimental design is done, the results obtained from these tests can give a better insight into the causes of rutting and the presence of interactions and synergisms among the system variables which have influence on rutting. This paper reports findings from a preliminary investigatory study conducted to quantify the effect of some of the critical system variables on airport pavement rutting performance using design of experiments.

II. Experimental Factors and their Levels

The concept of DoE was pioneered by R.A. Fisher in the 1920s [6]. Statistical DoE uses replication, blocking, randomization and orthogonality to acknowledge the

Manuscript received February 27, 2007.

Dr. Anshu Manik is a Research Associate with the Department of Civil Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail: manik@uiuc.edu).

Dr. Kasthurirangan Gopalakrishnan is a Research Scientist with the Department of Civil Engineering, Iowa State University, Ames, IA 50011, USA (e-mail: rangan@iastate.edu).

Mr. Siddhartha K. Khaitan is a Research Assistant with the Department of Electrical Engineering, Iowa State University, Ames, IA 50011, USA (e-mail: skhaitan@iastate.edu).

statistical interaction of variables, and utilizes statistics as an objective way of drawing conclusions in the presence of errors, noise, and unknown variables [7].

DoE can also be thought of as an experimental strategy that changes values of input variables with the purpose of studying the response of the system to these changes. In this, it is important to identify the experimental factors and their levels. A *factor* is an input variable that is controlled by the investigator and is manipulated to cause a change in the output. It is also sometimes called *independent variable*. Some of the factors that could affect the amount of flexible pavement rutting are:

(1) Amount of load that the pavement experiences
(2) Type of loading, i.e. characteristics of the wheels that are in contact with the pavement
(3) Frequency or speed with which the load travels on the pavement
(4) Type of material used in various layers of the pavements
(5) Amount and type of compaction done to different pavement layers
(6) Temperature and other environmental factors

The practice in the pavement industry today for testing in such cases is to carry out full scale testing of pavements constructed specially for this purpose. The scale of funding and organization required for full-scale testing puts practical constraints on the number of factors that can be even attempted for study. The set of experiments being reported here considered three of the factors expected to be more important in causing rutting namely, subgrade type, base type and loading type (loading wheel or gear configuration) based on available test results from the Federal Aviation Administration's National Airport Pavement Test Facility (NAPTF).

The NAPTF was constructed to generate full-scale traffic test data to support the development of advanced airport pavement design procedures. The NAPTF indoor test facility is located at the FAA's William J. Hughes Technical Center, near Atlantic City International Airport, New Jersey, USA. An inside picture of the NAPTF taken during the construction of test pavements is shown in Fig. 1 [8].

Housed within NAPTF is a 1.2 million-lb pavement testing machine spanning two sets of railway tracks that are 76 ft. apart. The vehicle is equipped with six adjustable dual-wheel loading modules with a total of twelve wheels. A hydraulic system applies the load to the wheels on the modules. The twelve test wheels are capable of being configured to represent two complete landing gear trucks having from two to six wheels per truck and adjustable up to 20 ft forwards and sideways. The wheel loads are adjustable to a maximum of 75,000 lbs per wheel (Hayhoe, 2004). During the first series of traffic tests (referred to as Construction Cycle 1), a dual-tridem or 6-wheel (simulated) Boeing 777 gear and a dual-tandem or four-wheel (simulated) Boeing 747 gear were tested on flexible pavements until they were deemed failed. The wheel loads were set at 45,000 lbs and the speed of the vehicle was 5 mph. A picture of the NAPTF test vehicle is shown in Fig. 2 [8].

An inertial profiling device was used to measure the transverse surface profiles periodically during the traffic testing to monitor the development of rut depths with repeated traffic repetitions. The NAPTF full-scale traffic testing rutting performance data used in this study is accessible for download at the FAA Airport Technology website: www.airporttech.tc.faa.gov and is analyzed and discussed in detail by Gopalakrishnan and Thompson [9].

Fig. 1 Inside view of NAPTF during construction of test pavements (Photo Courtesy: NAPTF)

Fig. 2 NAPTF test vehicle (Photo Courtesy: NAPTF)

Factorial designs are used in experiments involving several factors where it is necessary to study the joint effect of multiple factors on a response. A full factorial, which contains all possible combinations of factors (input variables) and levels (values of the factors), is necessary to avoid *aliasing* at any order. Alias means when the estimate of an effect also includes the influence of one or more other effects (such as by high order interactions). If it is assumed that the response is approximately linear over the range of factor levels chosen, then a two level study will define the response and a 2^k (k is the number of factors to be studied) design can be

implemented (Riter et al., 2005). In this study, two levels were identified for each of the three factors for 2^3 full factorial design as follows:

(a) Subgrade type (*X1*): subgrade refers to the original natural ground on which the pavement is constructed. Levels for this factor will be:

Low strength subgrade: (-)
Medium strength subgrade: (+)

At NAPTF, County Sand and Stone Clay (CSSC) with a target California Bearing Ratio (CBR) of 4 was used in the low strength subgrade, while DuPont Clay with a target CBR of 8 was used in the medium strength subgrade.

(b) Base type (*X2*): base course is the pavement layer constructed above subgrade (or above subbase, in some cases) using granular material or stabilized material.

Conventional base: (-)
Stabilized base: (+)

Conventional base is compacted unbound aggregate layer. Stabilized base generally has some cementitious material to give it extra strength. In the NAPTF flexible pavement sections, asphalt stabilized base courses were used.

(c) Loading gear configuration (*X3*): two different types of loading wheel configurations were used and they were:

Dual tandem (Boeing 747 gear): (-)
Dual tridem (Boeing 777 gear): (+)

III. Presence of Nuisance Variables

There are several variables which are not included in the test. Therefore it is required that they are either randomized or blocked. *Randomization* is a procedure that randomly determines the allocation of the experimental material, and the order of the experimental runs. *Blocking* is a restriction on the randomization of the schedule for conducting experiments such that any effects on the experimental results due to a known change that can not be controlled (i.e., nuisance variable) become concentrated in the levels of the blocking variable, in order to isolate a systematic effect and prevent it from obscuring the main effects [7]. Following are some of the nuisance variables that were identified in relation to this experimental design:

(1) Wheel load: since the variable being used in the experiment is wheel configuration and not wheel load, it was fixed constant to block its affect from entering the effects of other variables.
(2) Loading frequency: frequency of loading affects the way the pavement material responds to that loading. When the load passes with high speed over the pavement the effective HMA dynamic modulus is higher than that if load passes with slower speed, if the temperature is held constant. Therefore, to block the effect of this parameter same speed and hence same loading frequency was used on all the test pavements.
(3) Construction material and equipment: The P-401 asphalt concrete, P-209 and P-154 pavement geomaterials were used throughout. The construction equipment used during construction of the flexible test pavements remained the same.
(4) Mix formula: it is one of the most important factors that affects the performance of the pavement. But this factor is not being studied in this experiment, therefore it was blocked as the mix formula remained the same for all flexible test pavements.
(5) Time of construction: time of construction also has appreciable affect on the pavement quality because weather conditions affect HMA mixing temperature, subgrade compaction etc. So, the test pavements were constructed under similar weather conditions.
(6) Environmental factors during testing: The NAPTF is an indoor test facility and the test pavements during Construction Cycle 1 were constructed next to each other. Therefore, environmental effects during testing on all of them can be assumed to be similar.

There could still be many factors affecting the process which could not be quantified or foreseen. Any of such factors can introduce a bias in the test. To avoid this, the tests should be randomized. Unfortunately because of the practical constraints, it was not possible to randomize the tests and they were run in standard order.

IV. Response Variable

The selection of the response variable is critical for a successful experiment. It must be established that this parameter actually provides useful information on the system being studied. Most often the average or the standard deviation of the measured characteristic will be the response variable. In this study, the rutting performance of airport flexible test pavements is the response variable in terms of maximum rut depth measured on the pavement surface.

It is best to limit the allowable surface rut depth on airport pavements for smooth and safe aircraft operations. Typically, different degrees of severity (low, medium, high) are defined based on rut depth magnitudes which are used in triggering appropriate maintenance and rehabilitation activities. In this study, it is reasonable to consider the pavement to be failed, at least functionally, when the surface rut depth exceeds 1.5 inches. However, this definition is applicable if the pavement does not fail in any other distress mode like fatigue before experiencing rutting failure.

Two rutting measurements at two different pavement locations were taken for each test pavement section considered in this study. The pavement however, was constructed with very strict quality control and it was intended that each section is as uniform as possible. Using this data, two response variables can be defined. In one case, the response variable could be the amount of rutting for a fixed

number of load repetitions; and second could be the number of load repetitions to reach specific rut depths. Both of these response variables have their significance in pavement design. But in this study, only the first response variable (maximum surface rut depth) was considered.

Note that surface course of the test pavements were constructed with Hot Mix Asphalt (HMA) which can hardly be called as homogeneous material in the strict sense of the term. This indicates that variation can always be expected in the way the material behaves even when produced using similar material and equipment etc. One of the reasons for taking two readings for each number of passes is to be able to estimate this variation.

V. Experimental Design

Experimental design specifies values of the input variables $x_1 \ldots x_k$ at which one measures the response y. A 2^3 full factorial design was planned for this experiment. Therefore, four different types of test pavements need to be considered for each combination of subgrade and base types. Two identical test pavements were constructed for each type. These two pavements were loaded using one type of wheel configuration each. The experimental design matrix is shown in Table 1.

TABLE I
Design Matrix for the Experiment

Test no.	Subgrade Type (*X1*)	Base Type (*X2*)	Loading gear configuration (*X3*)
1	-1	-1	-1
2	+1	-1	-1
3	-1	+1	-1
4	+1	+1	-1
5	-1	-1	+1
6	+1	-1	+1
7	-1	+1	+1
8	+1	+1	+1

VI. Experimental Data

The tests were conducted as planned on all the pavements. The aim was to allow as many number of passes as possible to be run on each pavement and then use the highest common number of passes (before pavement failure) for comparing rut depth. Two measurements were taken from two different locations on each of the test pavements. This would be similar to the concept of repetition in experimental design. Table 2 shows portions of rutting measurements collected from NAPTF full-scale traffic tests. The rutting measurements are presented pictorially for other cases in Figs. 3 to 5 (measurement 1), where the portions of measurements used in statistical analysis are highlighted.

TABLE II
Rutting Results for Low Strength Subgrade (-) and Stabilized Base (+)

N # of Passes	Measurement 1 (mils) Dual Tridem	Measurement 1 (mils) Dual Tandem	Measurement 2 (mils) Dual Tridem	Measurement 2 (mils) Dual Tandem
5566	31	125	31	125
5866	16	125	16	125
6176	31	125	31	125
6518	16	109	16	109
6862	-16	125	-16	125
7156	16	109	16	109
7466	16	125	16	125
7744	31	141	31	141
7994	31	141	31	141
8038	31	141	31	141
8396	16	141	16	141
8678	16	219	16	219
8986	16	125	16	125
9264	0	141	0	141
9518	16	156	16	156
9836	0	125	0	125
10156	0	141	0	141

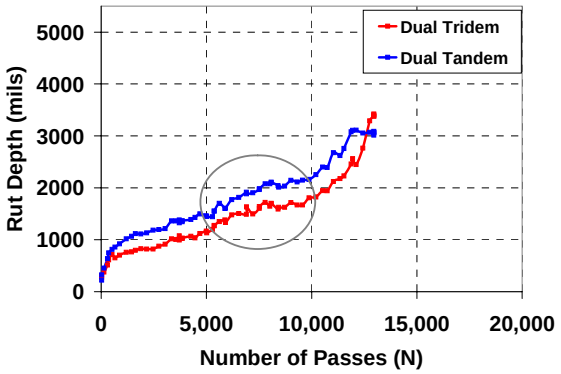

Fig. 3 Rutting results for Medium strength subgrade (+) and Conventional base (-)

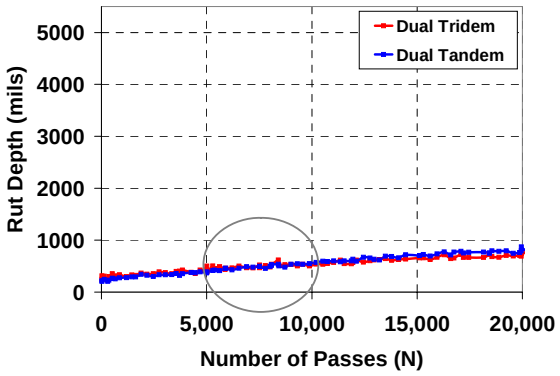

Fig. 4 Rutting results for Low strength subgrade (-) and Conventional base (-)

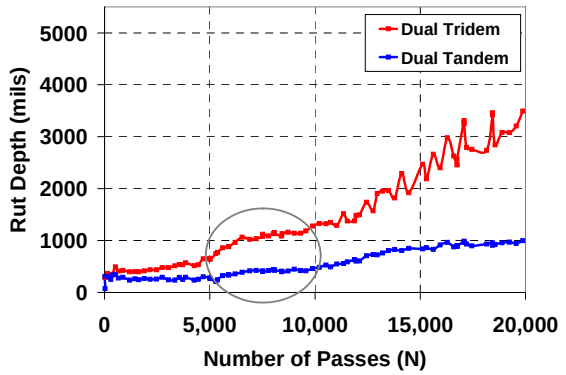

Fig. 5 Rutting results for Medium strength subgrade (+) and Stabilized base (+)

In this analysis, the response variable has been chosen to be the rut depth after completion of 10,156 passes. The values of response variable at different combinations of treatment levels are summarized in Table 3.

TABLE III
RUT DEPTH AT 10,156 PASSES (RESPONSE) BASED ON MEASUREMENT 1

Test No.	X1	X2	X3	Rut Depth – Measurement 1- Y1 (mils)	Rut Depth – Measurement 2 - Y2 (mils)
1	-1	-1	-1	250	188
2	+1	-1	-1	1625	2000
3	-1	+1	-1	140	203
4	+1	+1	-1	375	438
5	-1	-1	+1	47	172
6	+1	-1	+1	1125	1063
7	-1	+1	+1	0	78
8	+1	+1	+1	688	313

VII. STATISTICAL MODELING

The most important information that can be obtained from the measurements collected is the effect of the factors being studied in this experiment. Table 7 shows the calculations in this regard. The grand mean of the test results along with the second measurement values was found to be **544** mils.

The next step is to derive a model or response function relating the input variables and outcome, which could be used to the study the effects of various factors on the outcome. This is called response surface modeling in DoE terminology. A nice description of the theory of experimental design in the context of response surface modeling is provided by Riter et al. [7] which is also included here for the reader's benefit. The relationship between input variables and the outcome y can be expressed as:

$$y = f(\xi_1, \xi_2 \ldots \xi_k) + \varepsilon \quad (1)$$

where f is an unknown function, which may be very complicated, and ξ represents other non-systematic sources of variability not accounted for in f, such as measurement error. The goal is to approximate f by a relatively simple analytical function on the basis of experimental data. The variables ξ_1, $\xi_2, \ldots \xi_k$ are natural variables, that is they are expressed in the terms of the units used in the experiment. It is useful to convert the natural variables to coded variables, denoted x_1, x_2, . . . x_k., which are defined as dimensionless with a mean of zero and the same standard deviation, generally on a scale of -1 to +1. This conversion is accomplished by:

$$x_k = \frac{\xi_i - (\xi_{(-1)} + \xi_{+1})/2)}{(\xi_{(+1)} + \xi_{-1})/2)} \quad (2)$$

This brings all the variables on a common scale and allows the evaluation of influence that each of the variables (and their interactions) has on the output function regardless of its measurement units. The relationship between k coded input variables to the response y can be described in the form:

$$y = g(x_1, x_2 \ldots x_k) + \varepsilon \quad (3)$$

Because the true form of the response function is unknown, it must be approximated. The simplest approximation of the response surface can be expressed by a multiple linear regression model with k input variables:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon \quad (4)$$

The linear regression coefficients, β_j , will be estimated from the data. This is often referred to as model fitting. This equation postulates that the change in the response due to a change in an input variable x_j is constant regardless of the values of the remaining input variables. As such, the model describes the response as a hyperplane lying above the k-dimensional space of the independent variables. The model is sometimes called the *main effect model*. When the change in the response due to a change in an input variable x_j depends on the values of the other variables, this can be expressed by a model:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{i=2}^{k} \sum_{j=1}^{i-1} \beta_{ij} x_i x_j + \varepsilon \quad \sum \quad (5)$$

The term $\beta_{ij} x_i x_j$ is called a *statistical interaction* and represents a non-additive effect of the variables x_i and x_j. The introduction of the interaction term adds curvature to the response function. Both these are first-order models and are typically used in the initial (screening) stages of the analysis of the response surface.

The last row in Table 4 shows the main effects and interaction effects for all the factors. It should be noted that the second set of measurements (measurement 2) as collected in this experiment might count as repetition rather than replication. This is because the second measurement was obtained by measuring rut depth at a different location on the same pavement. So, all other conditions for pavement construction and testing were identical. In reality the measurements could have been called as replicates if a second pavement was constructed in each case and tested. That would have been much closer to reality because each pavement is constructed from scratch and has nothing to do with any other pavement constructed earlier.

TABLE IV
CALCULATIONS FOR DETERMINING MAIN AND INTERACTION EFFECTS

E1	*E2*	*E3*	*E1E2*	*E2E3*	*E1E3*	*E1E2E3*
-219	-219	-219	219	219	219	-219
1813	-1813	-1813	-1813	1813	-1813	1813
-172	172	-172	-172	-172	172	172
406	406	-406	406	-406	-406	-406
-109	-109	109	109	-109	-109	109
1094	-1094	1094	-1094	-1094	1094	-1094
-39	39	39	-39	39	-39	-39
500	500	500	500	500	500	500
818	**-529**	**-217**	**-471**	**197**	**-96**	**209**

Since the second measurement is just a repeated measurement, the variation that would be expected will most probably be smaller than that if the tests were actually replicated. Note that replication refers to performing the same treatment combination more than once in order to form an estimate of the random error independent of any lack of fit error. From statistical point of view, true error cannot be obtained by repeated test results and therefore, in this experiment we cannot find out true error of the experiment.

Therefore, to be able to establish which effects are significant and which ones are not, we need to use normal probability plots. It should also be noted, however, that this is a 2^3 factorial design, which means that use of normal probability plots will not be very effective, especially if many of the effects are significant. The Normal Probability Plot showed that there is not a definite trend through which a line could be drawn to get the significant effects. This means that the line that is drawn could vary appreciably while being similarly close to the points. To begin with, only effects *E1*, *E2* and *E123*, which were farthest from the line, were considered as significant. The logic for not choosing the other effects as significant is that in that case *E23*, *E12* and *E13* also will be significant which does not leave any effects on the plot which fall on the normal probability line. This would mean that use of normal probability plot will not be advisable.

Therefore, the model will be as follows. It should be noted that *E3* also is included in the model because *E123* is significant.

$$Rut\ Depth\,(at\,10{,}156\ passes) = 544 + 409 * X_1 - 265 * X_2 - 108 * X_3 + 99 * X_2 * X_3 \quad (6)$$

This model is a trial model only because of the reasons mentioned before. To test if this model is acceptable, residual analysis was done. A residual is the difference between the observed value of a response measurement and the value that is fitted under the model. Parameters β_k in the linear models are typically estimated from the experimental data by the method of least squares, and are denoted $\hat{\beta}$. Therefore, one can compute the values $\hat{y}_i$ of the outcome that are predicted by the model. For example, in the case of the model in Eqn (5), the values $\hat{y}_i$ predicted by the model for the input values $x_{i1} \ldots x_{ik}$ are computed as:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j + \sum_{i=2}^{k} \sum_{j=1}^{i-1} \hat{\beta}_{ij} x_i x_j \qquad \sum \quad (7)$$

The differences between the observed and the predicted values are called the residuals of the fitted model:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad (8)$$

The residuals are useful for making inference regarding the parameters and to assess the adequacy of the model. Table 5 shows the actual measurement, predicted response value and the residual. Before drawing conclusions on the basis of the fitted model, it is necessary to examine the adequacy of the fitted model and to determine that (1) it represents a satisfactory approximation of the response surface and (2) the assumption of normal error distribution with constant variance is plausible. In presence of replicates, these can be formally assessed using a lack-of- fit test and a Normal probability plot [10]-[11].

TABLE V
CALCULATIONS FOR DETERMINING MAIN AND INTERACTION EFFECTS

Test No.	*X1*	*X2*	*X3*	Average (mils)	Predicted (mils)	Residual
1	-1	-1	-1	219	403	-185
2	+1	-1	-1	1813	1431	382
3	-1	+1	-1	172	83	89
4	+1	+1	-1	406	693	-286
5	-1	-1	+1	109	395	-286
6	+1	-1	+1	1094	1005	89
7	-1	+1	+1	39	-343	382
8	+1	+1	+1	500	685	-185

The Normal Probability Plot showed that the points were

widely scattered around the straight line. This is an unhealthy sign because the residuals do not follow normal distribution.

The other option could be to accept that all the factors namely *E1*, *E2*, *E3*, *E12*, *E13*, *E23* and *E123* are significant. The residual obtained by following this model however will be close to zero because the same 8 response values were used to get seven effects and the grand mean which appear in the model as well. But in that case the use of normal probability plot for determining significance of the effects is not justifiable. We need to look at the experiment and determine if these tests are giving any meaningful data.

Because the study of response surfaces involves data that are subject to experimental error, observed differences in the response can be either due to the true effects of the input variables, or to the artifacts of the random variation. Obviously, it is important to distinguish between these two situations, and the statistical approach is the only objective way of making such conclusions from the data [7]. Formally, one tests the hypothesis:

$$H_0 : \beta_j = 0 \;\; against \;\; H_1 : \beta_j \neq 0 \tag{9}$$

If the null hypothesis, H_0 is not rejected, there is no evidence that levels of the input variable x_j result in systematic differences in the response, and the term can be excluded from the model. On the other hand, if H_0 is rejected, the corresponding term has a statistically significant effect and it should be kept. The hypothesis is tested using the estimated parameters $\hat{\beta}$ of the empirical model and from the estimate of the experimental error. If the magnitude of $\hat{\beta}_j$, as compared to its variance,

$$\frac{\hat{\beta}_j}{\sqrt{\hat{V}\hat{\beta}_j}} \tag{11}$$

is 'large', the hypothesis H_0 can be rejected. To determine which values of the ratio are 'large', additional distributional assumptions regarding the experimental error ε are needed. Typically, one assumes that $\varepsilon_1 \ldots \varepsilon_n$ are mutually independent and normally distributed, have a mean of 0 and a constant (but possibly unknown) variance σ^2. The estimation of the unknown variance σ^2 requires replicates of measurements of the response y for at least one set of levels of the input variables [7].

As an alternative approach to check the observations above, it is being assumed that the repeated measurements are replications for the sake of analysis. The motivation for this is as follows. Just looking at the data in the response table, it seemed that there is much more variation in the response than what was expected. If this variation itself is very significant then it certainly hints towards an even larger variation if the tests were replicated rather than repeated. This is because the two locations that were used for the two measurements were expected to be almost identical with very little variation arising from production variability. With the above assumption, we can try to get the error and find out which effects come out to be significant by this method. A summary of calculation of variances for each test are shown in Table 6.

TABLE VI
CALCULATION OF VARIANCES FOR EACH TEST

Test No.	*X1*	*X2*	*X3*	*Y1*	*Y2*	Y_{avg}	Variance
1	-1	-1	-1	250	188	219	1953
2	+1	-1	-1	1625	2000	1813	70313
3	-1	+1	-1	140	203	172	1992
4	+1	+1	-1	375	438	406	1953
5	-1	-1	+1	47	172	109	7813
6	+1	-1	+1	1125	1063	1094	1953
7	-1	+1	+1	0	78	39	3052
8	+1	+1	+1	688	313	500	70313

From Table 6 it can be seen that,

Sum of variances = 159341
Pooled variance:

$$s_p^2 = \frac{s_1^2 + s_2^2 + ... + s_8^2}{8} = \frac{159341}{8} = 19917.63 \tag{12}$$

Also,

$$s_{effect}^2 = \frac{4 s_p^2}{N} = \frac{4 * 19917.63}{16} = 4979.4 \tag{13}$$

Therefore,
Standard Error (S.E.):

$$s_{effect} = \sqrt{4979.4} = 71 \; mils \tag{14}$$

And, variance of the average:

$$Var(average) = \frac{s_p^2}{N} = \frac{19917.63}{16} = 1244.85 \tag{15}$$

Therefore,

$$s_{avg} = \sqrt{1244.85} = 35 \tag{16}$$

The next step would be to check the significance of the effects calculated earlier. For *E1*:

$$s_{effect} = 71 \; mils \tag{17}$$

Effect estimate = 818.44

Associated *t*-value:

$$\frac{E1 - 0}{s_{effect}} = \frac{818.44 - 0}{71} = 11.599 \tag{18}$$

Similarly, associated *t*-values for all other main and interaction effects were calculated and are listed in Table 10.

Degrees of freedom = (# of replicates-1) * # of tests
= (2-1) * 8 = 8

$t_{\alpha, 0.025} = -2.306$
$t_{\alpha, 0.975} = 2.306$

Comparing the *t* value associated with *E1* with *t* from the table, it can be concluded that the effect *E1* is significant. Table 7 lists the calculations for significance of all the effects.

TABLE VII
DETERMINATION OF STATISTICALLY SIGNIFICANT EFFECTS

Effect	Effect Estimate	Assoc. *t* value	t_{table}	Significance
E1	818.44	11.599	± 2.306	Significant
E2	-529.38	-7.502	± 2.306	Significant
E3	-216.72	-3.071	± 2.306	Significant
E12	-470.63	-6.670	± 2.306	Significant
E23	197.34	2.797	± 2.306	Significant
E13	-95.78	-1.357	± 2.306	Not Significant
E123	208.91	2.961	± 2.306	Significant

It was found that E1, E2, E3, E12, E23, E123 had significant effect on the response. Therefore, the model for the response with the three factors being considered would be as follows:

$$Rut\ Depth(at\ 10{,}156\ passes) = 544 + 409 * X_1 - 265 * X_2 - 108 * X_3 - \\ - 235 * X_1 * X_2 + 99 * X_2 * X_3 \\ + 104 * X_1 * X_2 * X_3 \qquad (19)$$

A 95% confidence level was used to determine the significance of main and interaction effects in this model. Although, it may not be a sound practice to use the repeated measurement and conduct variance analysis with it, a variance analysis has been attempted here to get a better idea of the variations considering the repeated measurements as replicated measurements. The results for the test of variance are summarized in Table 8. The *F-critical* value reported from the statistical table was 6.39.

From Table 8, it is noted that variance introduced by *E1*, *E23* and *E123* is significant. Fig. 5 shows the response verses run order of tests. It can be concluded from this plot that the run order does not seem to have any affect on the response. This is particularly important because the tests were not randomized because of practical constraints.

TABLE VIII
TEST OF VARIANCE

Effect	Sum (var(+))	Sum (var(-))	*F*-value	Significance
E1	144531.3	-14809.8	9.76	Significant
E2	77309.8	-82031.3	1.06	Not Significant
E3	83129.9	-76211.1	1.09	Not Significant
E12	82031.3	-77309.8	1.06	Not Significant
E23	145629.9	-13711.1	10.62	Significant
E13	76211.1	-83129.9	1.09	Not Significant
E123	150429.9	-8911.1	16.88	Significant

Fig. 5 Rut depth Vs run order

VIII. CONCLUSIONS

Rutting is a major load-related distress in airport flexible pavements. Significant research has been conducted to determine the factors which affect rutting and how they can be controlled. This paper presents findings from a preliminary investigatory study conducted to quantify the effect of some of the critical system variables on airport pavement rutting performance using the response surface methodology. In contrast to the one-factor-at-a-time approach, experimental design methods or Design of Experiments (DoE) both address the issue of interaction of variables and are generally more efficient.

The test results do strongly indicate that the response (rut depth) has too much noise in it and it would not allow determination of a good model. From a statistical DoE perspective, two major changes proposed for this experiment are: (1) actual replication of the tests is definitely required, (2) nuisance variables need to be identified and blocked properly. Practically speaking, it is very costly to run more number of tests in such full-scale traffic test scenario. But to be able to derive a good experimental model and meaningful

conclusions, it is necessary to run additional tests as proposed above.

It appears that the experimental results have too much of noise in them. This is not a healthy sign because this means that effect estimates are not very realistic, rather they are prone to high error. This means that at least some of the nuisance variables were not blocked properly. Further investigation is necessary to determine possible sources of noise in the experiment. Traditionally, laboratory experiments are conducted in a controlled fashion to understand the statistical interaction of variables using experimental design concepts. This study was a preliminary attempt to identify the critical system variables influencing airport flexible pavement rut depth from a statistical DoE perspective using real field data from a full-scale test facility. Thus, it would be advantageous to formulate the test factorial using experimental design concepts before conducting such expensive full-scale traffic tests to derive the maximum benefit from the analysis of data.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Federal Aviation Administration (FAA) Airport Technology Branch researchers, Dr. David Brill and Dr. Gordon Hayhoe, for providing the NAPTF rutting results and for the photographs included in this paper. Ms. Patricia Watts is the FAA Program Manager for Air Transportation Centers of Excellence and Dr. Satish Agarwal is the Manager of the FAA Airport Technology R & D Branch. The contents of this paper reflect the views of the authors who are responsible for the facts and accuracy of the data presented within. The contents do not necessarily reflect the official views and policies of the Federal Aviation Administration. This paper does not constitute a standard, specification, or regulation.

REFERENCES

[1] M.R. Thompson and D. Nauman, "Rutting Rate Analysis of the AASHO Road Test Flexible Pavements," *Transp. Res. Record 1384*, TRB, Washington, DC, 1993.
[2] R.P. Rawe, T.A. Ruhl, and R.J. Sunta, "Results of the 1989 ASCE Airfield Pavement Survey," in *Proc. Airfield Pavement Specialty Conference*, ASCE, New York, 1991.
[3] E. Guo and J. Rice, "General Statistic Performance Analysis of Asphalt Airport Pavements," in *Proc. Federal Aviation Administration Airport Technology Transfer Conference*, Atlantic City, NJ, 1999.
[4] U. Seyhan and E. Tutmluer, "Anisotropic Modular Ratios as Unbound Aggregate Performance Indicators," *Journal of Materials in Civil Engineering*, ASCE, Vol. 14, No. 5, 2002, pp. 409-416.
[5] J.B. Sousa, J. Craus, and C.L. Monismith, "Summary Report on Permanent Deformation in Asphalt Concrete," SHRP-A/IR-91-04, Strategic Highway Research Program, National Research Council, Washington, DC, 2001.
[6] R.A. Fisher, *Statistical Methods for Research Workers*. 9th ed., Oliver and Boyd, London, 1944.
[7] L.S. Riter, O. Vitek, K.M. Gooding, B.D. Hodge, and R.K. Julian, Jr., "Statistical design of experiments as a tool in mass spectrometry," *Journal of Mass Spectroscopy*, Vol. 40, 2005, pp. 565-579.
[8] G.F. Hayhoe, "Traffic Testing Results from the FAA's National Airport Pavement Test Facility," in *Proc. 2nd International Conference on Accelerated Pavement Testing*, University of Minnesota, Minneapolis, MN, 2004.
[9] K. Gopalakrishnan and M.R. Thompson, "Severity Effects of Dual-Tandem and Dual-Tridem Repeated Heavier Aircraft Gear Loading on Pavement Rutting Performance," The International Journal of Pavement Engineering, Vol. 7, No. 3, 2006, pp. 179-190.
[10] D.C. Montgomery, *Design and Analysis of Experiments*. 5th ed., Wiley: New York, 2000.
[11] R.H. Myers and D.C. Montgomery, *Response Surface Methodology: Process and Product Optimization using Designed Experiments*. 2nd ed., Wiley: New York, 2002.