Case Study -- Estimating the Amount of Phosphorus Implanted in Silicon

Motivated in part by the fact that International SEMATECH (a consortium of semiconductor manufacturers) recently listed an SRM (standard reference material) implant of phosphorus in silicon as a high-priority need, the SIMS (secondary ion mass spectrometry) community in the United States performed a round-robin study to calibrate the implanted dose of phosphorus in a silicon wafer by consensus. The dose determinations among the participating laboratories varied by almost a factor of two, however, reflecting primarily the errors of the respective in-house standards. This demonstrated the need for a common phosphorus reference material to improve reference reproducibility.

In pursuit of a phosphorus standard, a radiochemical neutron activation analysis (RNAA) was developed by NIST researchers R. L. Paul and D. S. Simons, critically evaluated, and shown to have the necessary sensitivity, chemical specificity, matrix independence, and precision to certify phosphorus at ion implantation levels in silicon. The end result of this work was described in Paul, Simons, Guthrie, and Lu (2003), with the last two authors performing the necessary statistical analysis.

The work of the latter involved three “rabbits” (polyethylene irradiation vessels) and there were three observations made on each rabbit, in addition to a point that resembled a center point, although the positions were not identical across the rabbits because observations could not be taken at the same position. The rabbits were treated as being homogeneous as there was no evidence that the observations differed to an appreciable extent over the rabbits. The objective was to arrive at a single number to represent the phosphorus level, plus and minus two times the uncertainty.

The uncertainty results from a propagation of error computation involving 11 uncertainties that are classified as Type A, and 6 uncertainties that are classified as Type B. Specifically, the Type A uncertainties, expressed as a percentage of the measured quantity, are squared and summed and the square root of the sum is computed. The same computation is performed for the Type B uncertainties. The two numbers that result from the computations are then squared and summed, with the square root of the sum obtained. That result, which turns out to be 0.84 is then multiplied by 2 so as to obtain the "relative expanded uncertainty" of 1.68%. This percentage is then multiplied by the result from the regression analysis, 9.58, to obtain 9.58(.0168) = 0.16, so the final result was 9.58×10^{14} atoms/cm^2 $\pm$ 0.16×10^{14} atoms/cm^2. This is the result that was used in SRM® 2133, which was the end product of the study. We discuss this further in Section C.

In section A we discuss the design of the experiment and consider the analysis of the data.

A. Experimental Design

The following quote from Box, Hunter, and Hunter (1978, p. 298) is relevant.

> The basic problem of experimental design is deciding what pattern of design points will best reveal aspects of the situation of interest.... The question of where the points should be placed is a circular one in the sense that, if we knew what the response function was like, we could decide where the points should be. But to find out what the response function is like is precisely the object of the investigation. Fortunately, this circularity is not crippling, particularly when experiments may be conducted sequentially so that information gained in one set directly influences the choice of experiments in the next.

This study involved only one design so the placement of the design points is more critical than would be the case if a sequence of designs were used. When

nothing is known about a possible model for the factors under study, a space-filling design might be used. These are designs that are appropriate when an experimenter has no *a priori* model in mind. When this is the case, the points might as well be regularly spaced over the design region, which is essentially what is accomplished with these designs. These designs, especially when used in conjunction with nonparametric regression and possibly semiparametric regression, are potentially beneficial in many applications and industries, including the pharmaceutical, biotechnology, chemical and process industries. The construction of space-filling designs is not simple, however, and an algorithm, software, or a catalog should generally be used.

A more conventional and better-known design for investigating possible second order effects as well as linear effects is a central composite design. This is a design for k factors that consists of the 2^k factorial points plus the $2k$ axial (star) points, in addition to a selected number of center points, with the number of center points and the position of the axial points selected in accordance with the desired properties of the design. A central composite design in two factors is given in Figure 1.

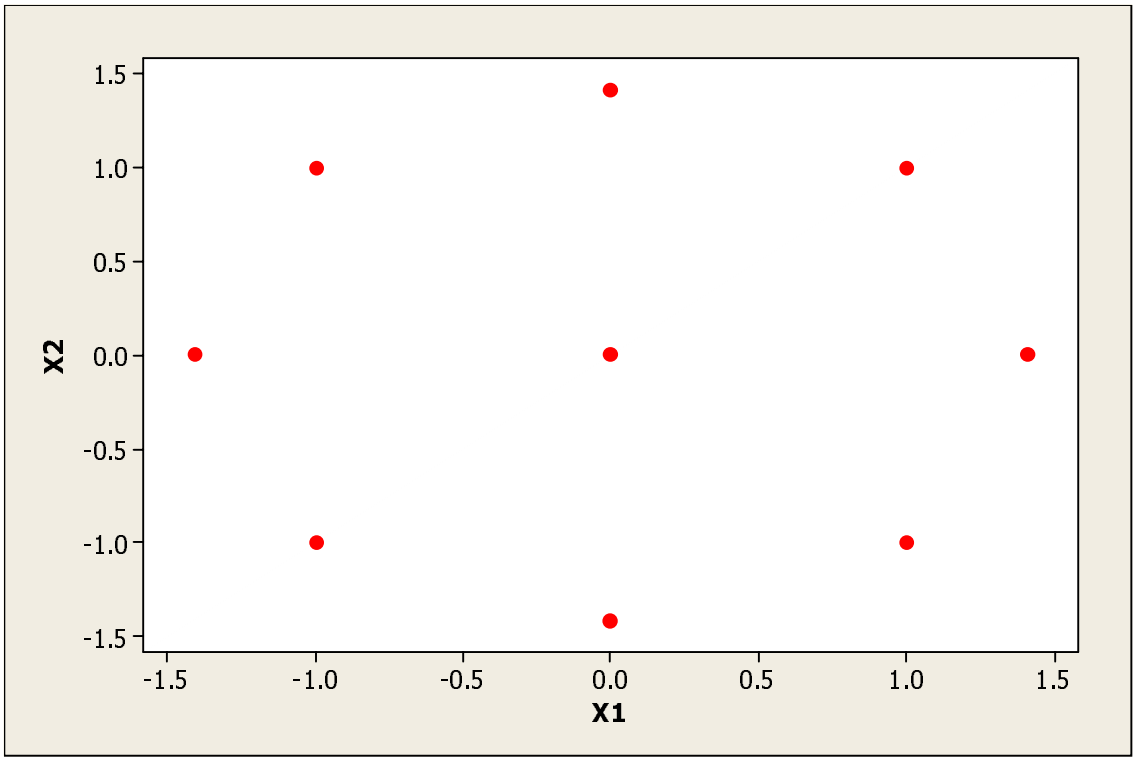

Figure 1. Central composite design in two factors

Notice that the design has 9 distinct points: center point(s), 2^2 factorial points, and the $2k = 4$ axial points that are a distance of α from the center, with the points having coordinates ($\pm \alpha$, 0) and (0, $\pm \alpha$). Orthogonality is almost always a desirable property of a design, and with two factors $\alpha = \sqrt{2}$ is needed to produce orthogonality.

The design points that were used in this study are shown, in the raw units, in Figure 2, and are listed below, along with the phosphorus concentration values.

X1	X2	Phosphorus Concentration
5	65	9.50
45	55	9.73
65	15	9.66
55	− 25	9.47
35	− 55	9.62
− 25	− 55	9.63
− 45	− 25	9.54
− 55	15	9.42
− 35	55	9.41
− 5	− 5	9.62
− 5	5	9.62
5	5	9.61

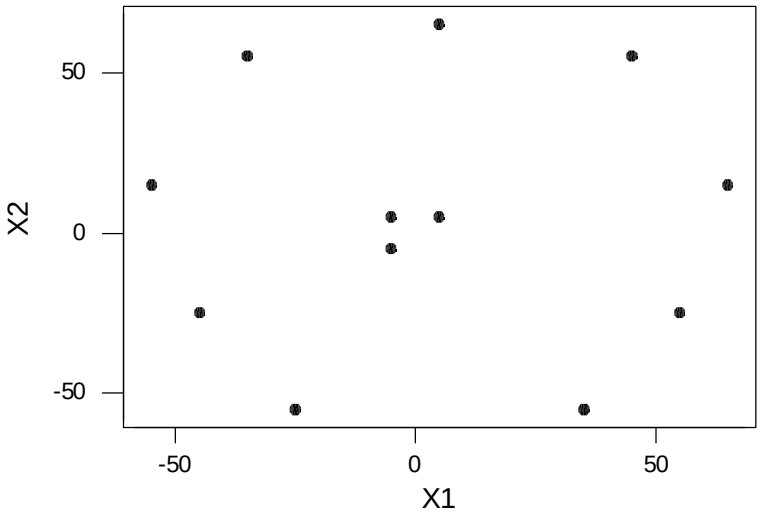

Figure 2. Design points used in the study

Notice that the design configuration in Figure 2 differs from the central composite design in certain ways. In particular, it was not possible to make repeated

runs in exactly the same position, so the three points in the center are not true center points but rather are as close to being center points as was physically possible. Notice also that there are 9 additional points whereas there are 8 additional points for a central composite design in two factors. We also note that the design in Figure 2 does not have the symmetry of the central composite design. This was because of some physical limitations. In general, it won't always be possible or practical to use the exact points as specified by a central composite design or spherical or radial-type designs because such designs will often call for fractional values of the factor to be used, and this could be odd fractions such as 55.365. A radial design for two factors is one for which the points would be equidistant from the origin. Given below is one of many possible configurations for a radial design in two factors, with the levels in coded units.

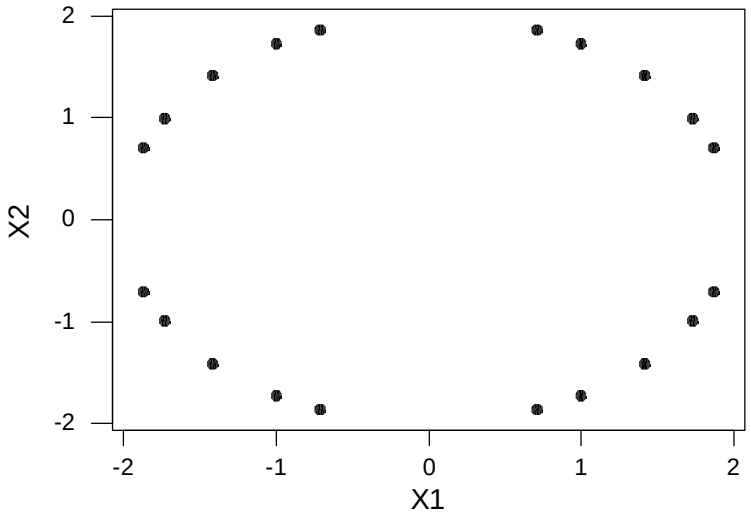

Figure 3. A radial design in two factors

With this design configuration, X_1 and X_2 are orthogonal, and the interaction term X_1X_2 is orthogonal to the linear term, as are the quadratic terms X_1^2 and X_2^2. The latter are also each orthogonal to X_1X_2, but they are not orthogonal to each other, and in fact are perfectly correlated. This is the general set of conditions that one would have with a central composite design if center points were not used, except that although the quadratic terms are correlated, they are not perfectly correlated (Of course one could also use center points in a radial design, which would reduce the correlation between X_1^2 and X_2^2 and drive the correlation toward zero as the number of center points is increased.)

In this study the objective was to have points that were 60 units from the center. That objective wasn't completely met, however, as a few points deviated slightly from this distance.

Nevertheless, close inspection of Figure 2 suggests that, ignoring the inner points, the correlation between the X_1 and X_2 values should be zero or practically zero, and. in fact, the correlation is zero. We examine more than first-order (linear) effects with designs such as this (which is approximately a radial design, given the physical limitations), so we would like to have the second-order effects estimated orthogonally to the first-order effects.

The correlations for the design in Figure 2 are given below, with correlations involving X_2^2 not shown because this is not a good candidate term.

Correlations: X_1, X_2, X_1X_2, X_1^2

	X_1	X_2	X_1X_2
X_2	0.005		
X_1X_2	0.102	0.148	
X_1^2	0.323	-0.019	0.025

Of concern here is the correlation between X_1 and X_1^2, which would be zero with a radial design or a central composite design. This moderate correlation could cause problems in trying to determine what terms to use in the model because certain methods for making this determination, such as *t*-tests, are undermined by non-zero correlations between candidate terms, with the extent to which the tests are undermined related to the size of the correlations.

B. Model Selection

A matrix scatterplot is helpful for initially seeing two-dimensional relationships, but this should be viewed as only preliminary work since we are interested in seeing the contribution of each term when the other terms are in the model, and such relationships cannot be seen from the matrix scatterplot.

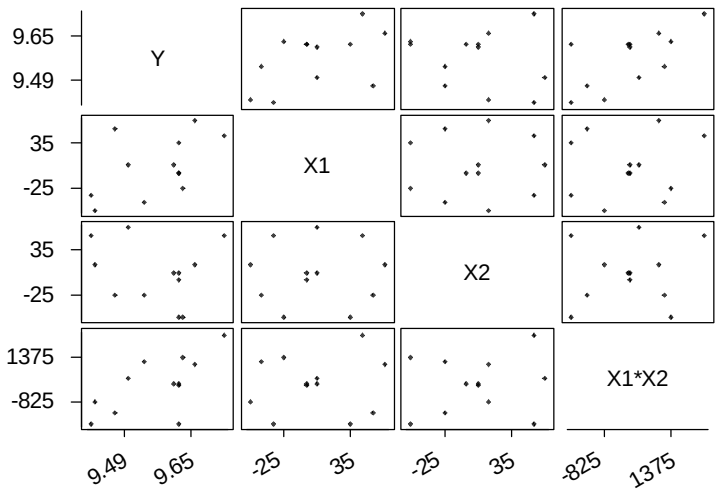

Figure 4. Matrix scatterplot

The terms displayed in Figure 4 were the terms that the experimenters used in the model. Because of the correlations between the predictors, various model selection methods should be used and the results compared. Probably the most common method is to simply look at the *t*-statistics when all of the candidate predictors are in the model, but *t*-statistics can be misleading when non-zero correlations exist between the predictors, as stated previously.

Nevertheless, the *t*-statistics are given below when all of the candidate predictors are used in the model.

```
Regression Analysis: Y versus X1, X2, X1*X2, X1**2, X2**2
```

```
The regression equation is
Y = 9.61 + 0.00142 X1 -0.000654 X2 +0.000048 X1*X2 -0.000025
X1**2 -0.000006 X2**2

Predictor        Coef     SE Coef          T        P
Constant      9.61051     0.03445     278.96    0.000
X1          0.0014187   0.0004823       2.94    0.026
X2          -0.0006536   0.0004631      -1.41    0.208
X1*X2       0.00004797  0.00001338       3.59    0.012
X1**2      -0.00002458  0.00001383      -1.78    0.126
X2**2      -0.00000564  0.00001204      -0.47    0.656

S = 0.05933     R-Sq = 80.9%     R-Sq(adj) = 64.9%
```

These results suggest that only X_1 and X_1X_2 should be used in the model, but, again, the results could be misleading because of the predictor correlations. It would be rather impractical to fit every possible model in these predictors and compare the results since there are $2^5 - 1 = 31$ such models, so variable selection methods should be used and the results compared.

It is worth noting that the correlation structure between the predictors produces some unusual results, as the X_1 term has a *p*-value of .10 when only X_1 and X_2 are used in the model -- thus suggesting that it shouldn't be used in the model --- whereas it has a much smaller *p*-value when all 6 terms are in the model. (The correlation between X_1 and X_2 is .005, so correlation is not the problem.)

Since there are only 31 possible models, we could use *all possible subsets* regression, which implicitly considers all possible models, and examine the results. The latter indicate that the model with terms $\{X_1, X_2, X_1X_2, X_1^2\}$ is a good candidate model with an R^2 value of 80.1 and a C_p value (see, e.g., Mallows, 1973) of 4.2. This is the smallest C_p value for the various models. Models should not be chosen on the basis of R^2 and/or C_p alone, however, and indeed the C_p statistic was intended to be used to identify a subset of good models, rather than one particular model.

If we use an *extra sum of squares test* to determine whether X_1^2 should be added to a model that has the terms $\{X_1, X_2, X_1X_2\}$, which as stated previously is what the experimenters used as their model, then we obtain the same result as just looking at the model *t*-statistics since the latter indicate the contribution of a particular term when the other terms are in the model. The *t*-statistics are given below.

Predictor	Coef	SE Coef	T	P
Constant	9.59972	0.02416	397.31	0.000
X1	0.0013886	0.0004506	3.08	0.018
X2	-0.0007076	0.0004228	-1.67	0.138
X1*X2	0.00004801	0.00001261	3.81	0.007
X1**2	-0.00002279	0.00001254	-1.82	0.112

On the basis of the *t*-statistics, we would probably not add X_1^2 to the model since the *p*-value is greater than .05. If we applied that logic consistently, however, we would also not add X_2 to a model that contained $\{X_1, X_1X_2\}$ as the *p*-value for adding that term is .187, which is considerably higher than the *p*-value for adding X_1^2. If we did so, however, we would have a non-hierarchical model as X_2 would appear in an interaction without there being a linear term in X_2. Whereas most statistical experts would argue for a hierarchical model, routine insistence on such models can cause problems because large interactions can cause main effect estimates to be small (i.e., the coefficient of the linear term will be small), which will give a misleading impression of the effect of the factor. We may wish to use numerical results such as those given above as a red flag that further investigation is needed.

Model selection is far from being an exact science and the use of different tools will often lead to the selection of different models.

C. Certified Value

It is of interest to see how the different models would produce different certified values. The certified value that was given in SRM 2133 was 9.58 x 10^{14} atoms/cm^2 $\pm$ 0.16 x 10^{14} atoms/cm^2. The 9.58 is the fitted value from the regression analysis with $\{X_1, X_2, X_1X_2\}$ as terms in the model, and the 9.58 being the fitted value at (X_1, X_2) = (7.191836, $-$ 11.68867), with the latter being halfway between the point (0,0) and the stationary point on the P (phosphorus) concentration surface, which was (14.38367, $-$ 23.37735). The determination of the uncertainty of 0.16 x 10^{14} atoms/cm^2 was explained in the 4th paragraph of this case study. The standard error of the fitted value, which was 0.0199, is included in the 0.16 value as "measurement uncertainty", which was one of the Type A uncertainties.

Adding more terms to the model would result in a different fitted value (9.61 when the four-term model is used) but the uncertainty would also increase as the standard error of the fitted value evaluated at the same (X_1, X_2) point is 0.0254, which is slightly greater than the 0.0199 obtained with the three-term model.

Arguments can be made for using the four-term model instead of the three-term model, just as one could argue for the two-term model with only $\{X_1, X_1X_2\}$ based on significance tests. (The fitted value using the latter model is 9.57 and the standard error is 0.0196. These numbers obviously differ very little from the results obtained using the three-term model.)

The standard errors for the three models clearly differ only slightly, and although there is a noticeable difference in the fitted values of 9.61 for the 4-term model and 9.58 for the 3-term model, the scientists involved in the experimentation considered this difference to be trivial for their application.

References

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters.* New York: Wiley.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

Paul, R. L., D. S. Simons, W. F. Guthrie, and J. Lu (2003). Radiochemical neutron activation analysis for certification of ion-implanted phosphorus in silicon, *Analytical Chemistry*, 75(16), 4028-4033.