Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

EXPERIMENTAL DESIGN ISSUES IN SIMULATION WITH EXAMPLES FROM SEMICONDUCTOR MANUFACTURING

Sarah J. Hood
Peter D. Welch

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

The contents of this paper reflect a portion of the conference tutorial on experimental design and simulation. In this paper we primarily focus on two level fractional factorial designs and their application to a discrete event simulation model of semiconductor manufacturing logistics. In the tutorial we will discuss a wider range of experimental design issues. However, in all cases, examples will be given of their application to the same semiconductor manufacturing model.

1 INTRODUCTION

There are two basic environments in which statistics are applied: one where the data are generated by processes out of the control of the investigator and the other where the investigator plans and hence in that sense controls the generation of the data. In the latter case, the investigator has some goal in mind and plans an experiment or sequence of experiments designed to take him in the direction of that goal. This planning is, in a broad sense, experimental design. In simulation the investigator has complete control over the generation of the data so all discussions of simulation output analysis are in some real sense discussions about design of experiments. This includes the estimation of output parameters with confidence intervals and confidence regions, the comparison of two systems, the ranking of multiple systems, etc.

We, however, will be concerned with what is more commonly considered experimental design. These are situations where one has a parametric simulation model such that it is impossible to exhaustively explore the parameter space. Hence the relationship between the output characteristics of the model and the parameters must be explored using regression type models and least squares techniques. In the tutorial talk we will review the basic notions of experimental design and present a number of examples of its application. In this paper we will discuss a single application. Hence it should not be considered a summary of the tutorial but rather a sample of the kinds of concepts and types of examples which will be considered.

2 DESCRIPTION OF THE MODEL

The discussion will use, as an example, a detailed, validated model of semiconductor manufacturing logistics. The model contains on the order of one hundred tool groups processing multiple products. The flow is highly re-entrant, that is, jobs feed back through sequences of the tool groups up to twenty times. The model includes tool setup, tool breakdown and repair, preventative maintenance, rework, test wafer send ahead, and detailed operator schedules. The primary purposes of the model are to study control rules proposed for the line [Hood et al. 1989] and to design new lines.

For the example discussed here, the line is under constant load and in a stationary state. That is, the rate of output of good product and scrap is equal to the rate at which orders are inputted. No portion of the system is saturated. We are interested in the cycle time, the time from beginning of manufacture to completion, of one of the products.

3 DESCRIPTION OF THE EXPERIMENT

3.1 Motivation

As was mentioned in section 2, interrupts for setup, preventative maintenance, and tool repair (after failure) are explicitly modeled in the simulation. They are all modeled in the same fashion, as highest priority sources of work for the tool groups. The interrupts (arrivals) are a renewal process and the interrupt durations (service times) are an i.i.d. sequence of random variables. The three interrupt processes operate independently.

These interruptions create backups of the actual semiconductor jobs and hence, contribute to the magnitudes of the cycle times. The purpose of the experiment described in this paper is the examination of the effect of these interrupt processes on the mean cycle time.

Each interrupt process is characterized by two distributions: one for the time between interrupts and one for the duration of the interrupt. In this investigation, two factors are associated with each distribution: the mean and the distribution type. Thus, there are a total of four factors associated with each interrupt process and, since there are three different interrupt processes, there are twelve factors in all.

The model under investigation is a model of a planned manufacturing line. The purpose of this study is to estimate the potential benefit, in terms of mean cycle time reduction, in effecting changes in the interrupt processes. Thus there is a "base" case, the planned parameters under which the line is expected to operate, and we are interested in the effect of changes made to this base case. Hence, in contrast to many situations in which design of experiments is used purely for sensitivity analysis, here there is a known specific starting state, the base case, and the desire to identify the improvement which can be achieved by moving to another specific state. This will be important when alternative experimental designs are considered in section 3.2.

Achieving cycle time reductions by effecting changes in the interrupt processes is particularly attractive since it involves less cost than reductions achieved by increasing the direct resources (tools and operators) involved in manufacturing.

The factor settings selected for the distribution of the time between interrupts were:

- distribution type (Factor A)
    1. exponential (base case)
    2. triangular (+ and - 100% of the mean)
- mean (Factor B)
    1. base case
    2. twice the base case

and for the duration of the interrupts

- distribution type (Factor C)
    1. exponential (base case)
    2. triangular (+ and - 100% of the mean)
- mean (Factor D)
    1. base case
    2. one half the base case

The non-base case settings would be expected to generate improvements. This is obvious in the case of the means. In the case of the distribution type, one would expect a reduction in the coefficient of variation to lead to a reduction in the mean cycle time. The exponential distribution has a coefficient of variation of 1, the triangular distribution selected has a coefficient of variation of 0.41. It is often necessary to set up the design is this way, with one level of the factor at the base case and the other level representing an *improvement* over the base case, when using design of experiments with discrete event simulation. This ensures that there is sufficient resource capacity represented in all of the experiments in the design. Otherwise, if the factor levels represent situations worse than the base case, then there is no guarantee that there is sufficient resource capacity for a steady-state result and the value of the performance measure may be infinity!

Hence, we have an experimental situation with 12 factors, each at two levels. Now one would expect that the effect of making a change in one of these factors would depend on the general level of congestion in the system. The greater the level of congestion, the larger the amount of change. Hence, a priori, one would expect a high level of interaction between the factors. For example, one would expect that the improvement in the mean cycle time resulting from the change in a factor from the base case would be greater if all the other factors were at the base case than it would be if other factors had been changed from the base case. Because we have chosen one level of the factors to represent the base case and the other level to represent an improvement over the base case, any significant change to a factor means there is less congestion in the manufacturing line and any additional changes are less likely to have as large an effect as they would when the line was running closer to saturation.

Because of this liklihood of significant higher order interactions we felt we had to consider resolution 5 or higher designs. We considered two possibilities. The first was a resolution 5 design on all 12 factors. It would have required 256 runs and would have left uncertainty about three way and higher interactions. Also, we wanted to generate replications so as to have an estimate of pure error so this choice would have required a very large amount of computing time. Since we were interested in changes from the base case to a state with improved performance, we decided instead on three separate, four factor, full factorial designs, with the idea in mind that line engineers would focus on improving the one interrupt process that improved performance the most and then that new state would serve as the base case for a second investigation. All designs were created and analyzed using the IBM software "A Graphical Statistical System" (AGSS). For a description of this system see Lane and Welch (1987).

The first experiment considers only the setup process with the factors for the other processes set at the base case. The second experiment considers only the preventative maintenance process with again all the factors of the other processes set at the base case. The third experiment, in a parallel fashion, considers only the failure-repair process. This alternative design is suitable for identifying a new state to move to from the base case state. If a base case does not exist and the purpose of the experiment is a more general sensitivity analysis, then another design might be more appropriate.

3.2 The Design Chosen

Each interrupt process is represented by four factors. As stated above one would assume, a priori, that there are interactions between the factors. To study the full detail of these interactions we chose a full factorial experiment, i.e. all 16 possible combinations of the four, two level factors.

The theory of 2 level designs provides a hierarchy of increasingly complex models leading to the full complexity of all possible interactions. The simplest model assumes that only the main effects are significant. That is, it assumes that

$$R = \mu + \sum_{i=1}^{4} \alpha_i x_i$$

where R is the mean cycle time, μ is the grand mean over the experiment, α_i is the coefficient associated with the *ith* factor and $x_i = -1$ if the *ith* factor is at the lower level (the base case) and $x_i = +1$ if the *ith* factor is at the upper level. This model has five parameters and assumes there are no interactions between the factors. It assumes each factor has an additive influence independent of the values of the other factors.

The next simplest model assumes that

$$R = \mu + \sum_{i=1}^{4} \alpha_i x_i + \sum_{i,j} \beta_{ij} x_i x_j$$

where β_{ij} is the coefficient associated with i,j th two way interaction. It assumes that the factors do interact but only in pairs and additively. There are six β_{ij} coefficients hence this model has 11 parameters. The existence of two way (and higher) interactions changes the interpretation of main effects. The significance of interactions and the interpretation of the response variable when they exist will be discussed in section 5. The complete model, which has no restrictions, is

$$R = \mu + \sum_{i=1}^{4} \alpha_i x_i + \sum_{i,j} \beta_{ij} x_i x_j$$

$$+ \sum_{i,j,k} \gamma_{ijk} x_i x_j x_k + \lambda_{1234} x_1 x_2 x_3 x_4$$

The hierarchical formulation has two advantages. First, it creates a sequence of models from the simple to the complex. Second, the components of the models are orthogonal so the significance of the terms can be judged independently. That is, the significance of any single term is independent of the model in which it is imbedded and the other terms of that model.

However, to judge the significance of a term in the model or to test a submodel we must have an estimate of error. Now error estimates are of two varieties. There are pure error estimates which come usually from replications and (non-pure) error estimates which depend upon model assumptions. Pure error estimates are highly desirable. In simulation applications it is always possible to get pure estimates of error independent of any assumptions of the model. These pure error estimates can then be compared with error estimates obtained through the fitting of a model to give a test of the goodness of fit of the model.

To get a pure estimate of error the experimental points were replicated using the method of batch means. The batch sizes were chosen to be 200 observations in length based on autocorrelation functions which went to zero at about 50 lags. An initial transient of length 250 was removed. Thus we had 3 experiments each with 80 runs, the full 2^4 factorial replicated 5 times. A sample series of length 1250 (with the initial transient) is shown in Figure 1.

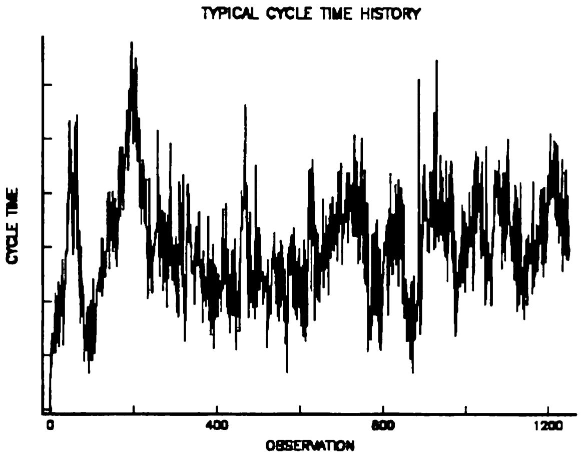


Figure 1. A Cycle Time History

The sample correlation function corresponding to this series (with the initial transient removed) is shown in Figure 2.

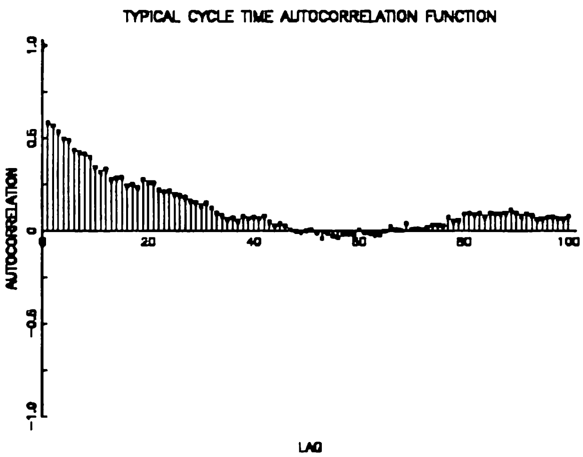


Figure 2. The Autocorrelation Function of Cycle Time History

4 MODEL SELECTION AND DIAGNOSTICS

4.1 Impact on Capacity

The interrupt processes result in a reduction of the capacity of the line by an amount which is the proportion of time in the interrupted state. This proportion is the mean interrupt time divided by the mean time between interrupts. These reductions can be used to rank the interrupt processes. They are given below:

Setups	0.157
Preventative Maintenance	0.029
Failure - Repair	0.015

Now, as discussed in section 3.2, three independent experiments of 80 runs each were conducted. In each experiment, the effect on mean cycle time of changes in a particular interrupt process was investigated. The parameters of the other interrupt processes were set at values representing the base case. We will now discuss the results of these three experiments, starting with model selection and diagnosis. Then, in section 5, the interpretation of the selected model is discussed with emphasis on the interpretation of interactions.

4.2 The Setup Process

The main effects are represented as A,B,C, and D where

A: inter-interrupt distribution type
B: mean time between interrupts
C: interrupt distribution type
D: mean interrupt time

The two way interactions are represented by AB, AC, AD, ... , the three way interactions by ABC, ABD, ... , etc. The values of the main effects and interactions are twice their corresponding coefficients. The standard t-tests on the effects gave the 95% confidence intervals indicated in Table 1.
Here, the estimate of error is the pure error estimate given by the 64 degrees of freedom associated with the replications. The model suggested by this table includes the effects A, B, C, D, AB, AD, BD and ABD. Thus, it is of the form

$$R = \mu + \sum_{i=1}^{4} \alpha_i x_i + \beta_{12} x_1 x_2 + \beta_{14} x_1 x_4 + \beta_{24} x_2 x_4 + \gamma_{124} x_1 x_2 x_4 \qquad (1)$$

This is a relatively complex model. The meaning of the interaction terms is discussed in the next section.

Here, we concentrate on the problem of model selection and diagnostics. The preceding model was generated by choosing the effects which were indicated as significant in Table 1, however the signif-

```
                              TABLE OF COEFFICIENTS

80 OBSERVATIONS    R-SQUARED     = 0.94239   STANDARD ERROR = 28.214
15 VARIABLES       ADJ R-SQUARED = 0.92889


                                                             0.95 CONFIDENCE LIMITS
EFFECT  ESTIMATE   STD ERR    T STAT       SIG LEVEL       LOWER          UPPER
 A       -49.732   6.3088     -7.8829     5.1861E-11      -62.346       -37.117
 B      -100.04    6.3088    -15.857      2.2204E-16     -112.65        -87.422
 C       -18.959   6.3088     -3.0052     3.7866E-3       -31.574         -6.3449
 D      -145.12    6.3088    -23.003      1.3878E-16     -157.73       -132.5
 AB       22.479   6.3088      3.5631     6.9870E-4         9.8643       35.093
 AC      -10.43    6.3088     -1.6532     1.0319E-1       -23.044          2.1848
 AD       50.286   6.3088      7.9708     3.6290E-11       37.672        62.901
 BC        1.2096  6.3088      0.19173    8.4856E-1       -11.405        13.824
 BD       60.795   6.3088      9.6365     4.4825E-14       48.18         73.409
 CD        6.5902  6.3088      1.0446     3.0013E-1        -6.0243       19.205
 ABC      12.538   6.3088      1.9874     5.1166E-2        -0.07668      25.152
 ABD     -25.693   6.3088     -4.0726     1.3040E-4       -38.308       -13.079
 ACD      -1.9588  6.3088     -0.31049    7.5720E-1       -14.573        10.656
 BCD      -6.1841  6.3088     -0.98023    3.3066E-1       -18.799          6.4304
 ABCD      5.085   6.3088      0.80601    4.2322E-1        -7.5295       17.699
```

Table 1. Coefficient Estimates and Confidence Intervals

icance tests of Table 1 are only approximately valid because there is a selection process taking place. If one picks an effect before looking at the table and then looks at its confidence interval, the corresponding significance test is valid at the 0.05 level. However, if one looks at the table and selects an effect whose confidence interval does not contain zero, then the test is not valid at the 0.05 level. For example, suppose none of the effects were significant. Then, since there are 15 of them, one of them would appear to be significant by chance with a probability greater than 0.05.

To counter this selection problem, it is common to generate a probability plot of the effects and to view them in the context of their distribution. Then if only a few are significant they will stand out against the distribution of the remainder which should be normally distributed and fall approximately along a straight line when plotted on a normal probability scale. Figure 3 shows a probability plot of the 15 effects in this case.

Such plots have to be viewed with care particularly when there is a pure estimate of error available from the replications as in this case. If one looks at Figure 3, one would be inclined to choose the simpler model containing the effects A, B, D, AD and BD. If these effects are removed as has been done in Figure 4, one sees that the remaining effects lie on a straight line indicating they are consistent with the assumption of a normal distribution. However, the normal distribution that they are consistent with does not have a variance estimate which is consistent with that obtained from the 64 degrees of freedom associated with pure error. On Figure 4, we have plotted the straight line of the normal distribution which insignificant effects should have if they are consistent with the pure error estimate.

Comparison of the plotted points and the straight line in Figure 4 is analogous to the comparison of the error estimates obtained from the lack of fit sum of squares and the pure error sum of squares in the usual analysis of variance table. That comparison is made rigorous through the standard F test. The results of this test for the simple model were highly significant indicating a lack of fit and, hence, an inadequate model. In the case of the more complex model including AB and ABD, the F-test did not indicate a lack of fit. In the case of the more complex model, the distribution of the 7 effects not included in the model was consistent with the straight line in Figure 4.

A second important diagnostic test is a check on the distribution of the residuals. The methodology assumes that there is an error term which is normally distributed with a constant variance. In Figure 5, we show a set of standard graphics tests of this assumption. The distribution of the residuals is compared with a fitted normal distribution in histogram-density, cdf and probability plots; and the residuals are plotted against the fitted values. Often the variance of the residuals increases with the level of the response. There is only a slight indication of that in this case, not enough to consider any remedial steps. For a discussion of more intensive testing of this assumption of a common variance see Hood and Welch (1990).

Hence, we accept the model suggested by Table 1 and given by equation (1). Remember, the coefficient values in equation (1) are one-half the effect estimates of Table 1.

4.3 The Preventative Maintenance Process

In the case of the preventative maintenance interrupt process the standard t-tests on the effects indicate that only the main effects B and D are significant. The estimate of the B effect is -27 and the D effect is -79. Hence the response equation is

$$R = \hat{\mu} - 13.5x_2 - 39.5x_4$$

where $\hat{\mu}$ is the estimate of the grand mean.

4.4 The Failure-Repair Process

The analysis of the data from the failure-repair interrupt process generates a model with only one effect, the main effect D. The estimate of D is -34. Hence, the response equation is

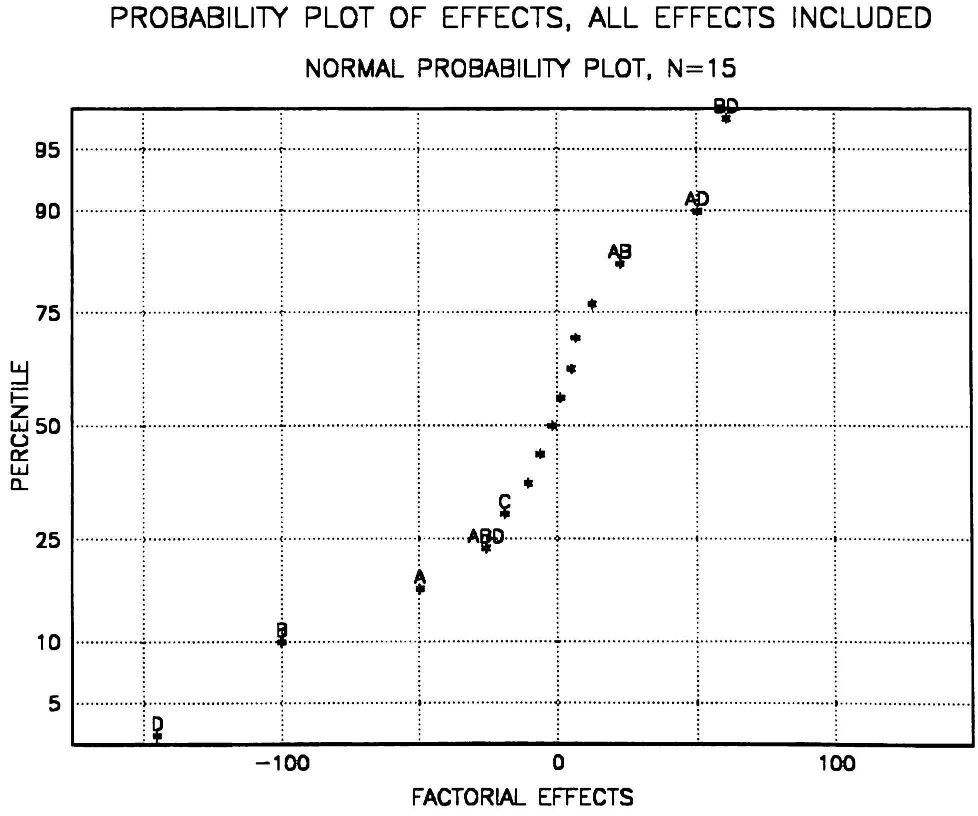
$$R = \hat{\mu} - 17x_4$$

5 MODEL INTERPRETATIONS, INTERPRETATION OF INTERACTIONS

5.1 The Failure-Repair Process

The simplest model is that corresponding to the failure-repair process. This is the interrupt process which generates the smallest reduction in the capacity of the overall system. The only effect is the main effect for the mean repair time. The results indicate that the only change that has a detectable effect in this experiment is to reduce the mean time to repair. The improvement in cycle time will be by an estimated 34 hours.

It is interesting to note that doubling the time between failures results in the same improvement in capacity as halving the time to repair but does not produce a statistically significant change in the mean


PROBABILITY PLOT OF EFFECTS, ALL EFFECTS INCLUDED

NORMAL PROBABILITY PLOT, N=15

Figure 3. Probability Plot: All 15 Effects

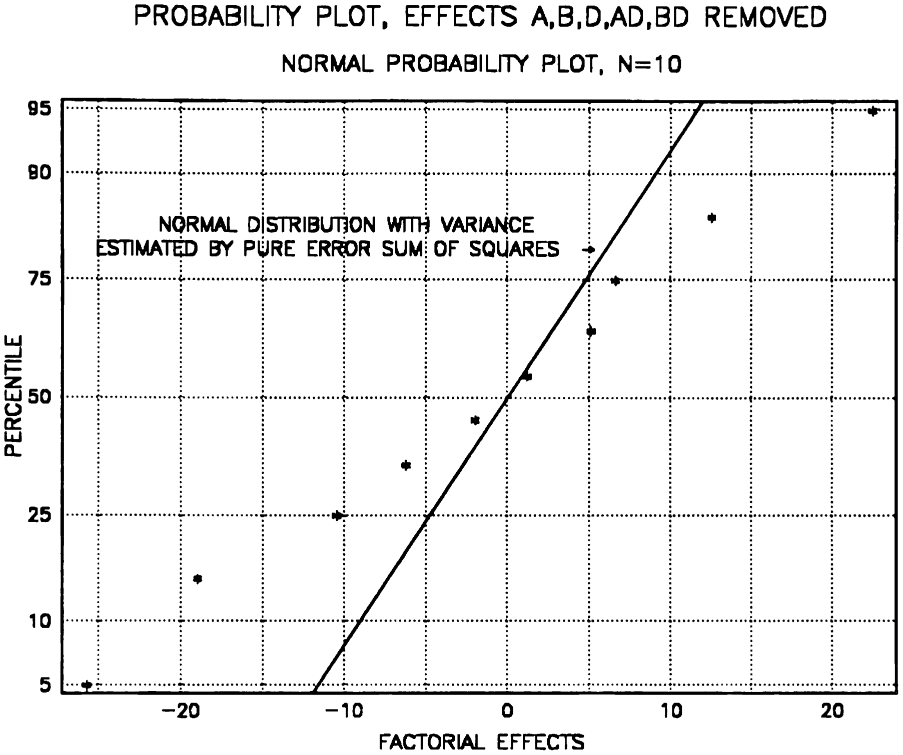


Figure 4. Probability Plot: Effects of Simple Model Removed

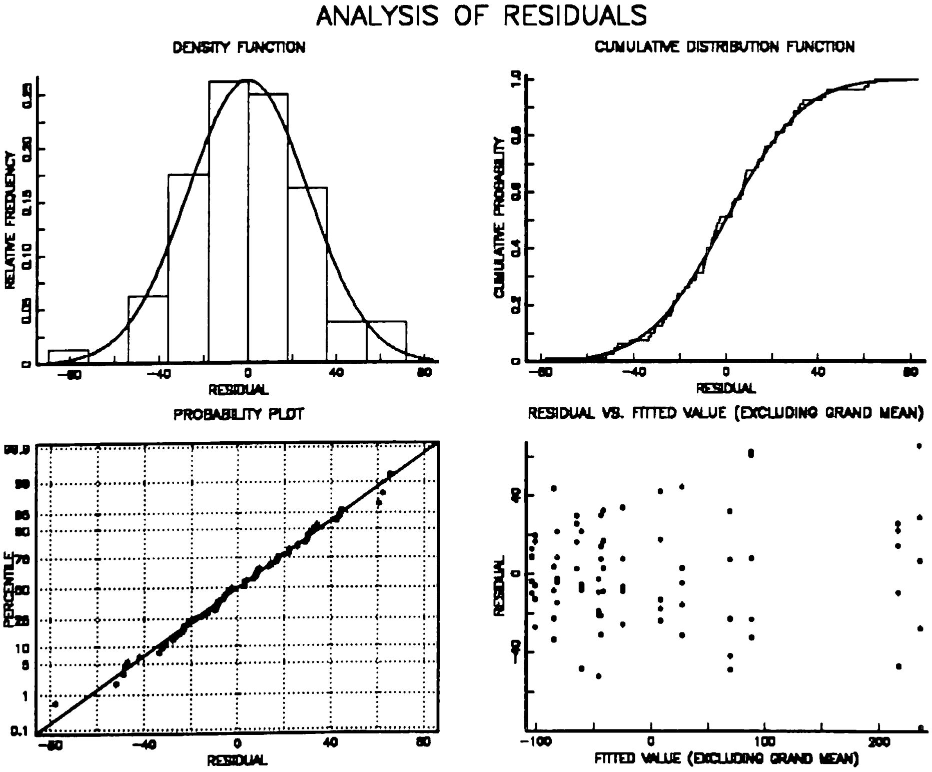


Figure 5. Residual Analysis Plots

cycle time. This importance of the interrupt time over the time between interrupts is consistent over all the interrupt processes. In this case it suggests that equipment should be specified in terms of time to repair as well as time to failure and that special attention should be paid to repair procedures.

5.2 The Preventative Maintenance Process

The next more complex model corresponds to the preventative maintenance process, the process having the next smallest reduction in the overall capacity of the system. The only two significant effects are B and D, the main effects corresponding to the mean time between preventative maintenance and the mean interrupt time for preventative maintenance. There is no significant interaction between the two effects. Hence, by doubling the time between interrupts we can improve the mean cycle time by 27 hours, by halving the mean interrupt time we improve it by 79 hours and by doing both we can improve it by 106 hours. Again, even though doubling the mean time between interrupts creates the same increase in system capacity as the halving of the mean interrupt time, its effect on the mean cycle time is much less. Long interrupts have a very deleterious effect on mean cycle time even when the overall degradation to capacity is small.

5.3 The Setup Process

The setup process has the most complex model with all four main effects, three two way interactions and one three way interaction. The effects, A, B, C, D, AB, AD, BD and ABD are all significant. The model is

$$R = \hat{\mu} - 25x_1 - 50x_2 - 9.5x_3 - 72.5x_4$$
$$+ 11x_1x_2 + 25x_1x_4 + 30.5x_2x_4 - 13x_1x_2x_4 \quad (2)$$

Now because of the existence of the interaction terms the result of making a set of specific changes cannot be inferred from the main effects alone. Furthermore, the results of making changes is very dependent on the order in which the changes are made. For example, if we consider the response matrix for the two mean factors B and D (represented by the variables x_2 and x_4) with the distribution factors A and C (represented by the variables x_1 and x_3) fixed at the base case (i.e. $x_1 = x_3 = -1$). Working through equation (2) we obtain the following matrix for the effect (relative to the grand mean) of the two factors at the two levels.

$$\begin{array}{cccc} & + & -45.5 & -80.5 \\ \mathrm{D} & & & \\ & - & 236.5 & 27.5 \\ & & - & + \\ & & \multicolumn{2}{c}{\mathrm{B}} \end{array}$$

Hence consider the mean setup time, factor D. The main effect indicates that changing it will result in an improvement of 145 whereas because of the interactions if we change it from the base case the improvement will be 282, almost twice as much. To illustrate the importance of the sequence in which changes are made consider the mean time between setups, factor B. The main effect is 100. If we change it from the base case we get an improvement of 209. But if we change it after we have changed the mean setup time then we only get an improvement of 35. Thus, when there are interaction terms they must be taken into account when considering the effect of specific changes. Looking only at the main effects can be very deceiving.

6 SUMMARY

This has been a brief discussion on the application of experimental design to simulation with an example from semiconductor manufacturing. The emphasis has been on the value of a "pure" estimate of error, the process of model selection and diagnostics, and the importance and interpretation of interaction terms.

REFERENCES

Hood S.J., A.E. Amamoto, and A.T. Vandenberge. 1989. A modular structure for a highly detailed model of semiconductor manufacturing. *Proceedings of the 1989 Winter Simulation Conference*. E.A. MacNair, K.J. Musselman, and P. Heidelberger, Eds. IEEE, Piscataway, NJ, 811-817.

Hood S.J. and P.D. Welch. 1990. The application of experimental design to the analysis of semiconductor manufacturing lines. Proceedings of the 1990 Winter Simulation Conference. O. Balci, R.P. Sadowski, and R.E. Nance, Eds. IEEE, Piscataway, NJ, 303-309.

Lane, T. and P.D. Welch. 1987. The integration of a menu-oriented graphical statistical system with its underlying general purpose language. Computer Science and Statistics: Proceedings of the 19th Symposium of the Interface. American Statistical Association, 267-273.

AUTHOR BIOGRAPHIES

SARAH J. HOOD is a Research Staff Member in the Manufacturing Research Department of the IBM Thomas J. Watson Research Center in Yorktown Heights NY. She received a Ph.D. in Mechanical Engineering from the U. of California, Davis. She is currently working in the discrete event system domain exploring various methods and tools for decision support of semiconductor manufacturing lines. She has also simulated environmental, physiological, and electro-mechanical systems. She is a member of SCS and IEEE.

PETER D. WELCH is a Research Staff Member in the Computer science Department at the IBM Thomas J. Watson Research Center in Yorktown Heights NY. He received a Ph.D. in Mathematical Statistics from Columbia University. His research interests are graphical-statistical software and simulation output analysis. He is a member of ORSA.