Proceedings of the 1998 Winter Simulation Conference
D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.

IDENTIFYING IMPORTANT FACTORS IN DETERMINISTIC INVESTMENT PROBLEMS USING DESIGN OF EXPERIMENTS

Willem J.H. Van Groenendaal
Jack P.C. Kleijnen

Department of Information Systems and Auditing/Center for Economic Research (CentER)
School of Management and Economics
Tilburg University
5000 LE Tilburg, THE NETHERLANDS

ABSTRACT

For large investment projects sensitivity analysis is an important tool to determine which factors need further analysis and/or can jeopardize the future of a project. In practice reliable information on the joint probability distribution of factors affecting the investment is mostly lacking, so a stochastic analysis is not possible. This paper analyzes how and to what extend statistical design of experiments in combination with regression meta modeling can be helpful in finding important factors in deterministic models. Information that is useful to decision makers.

1 INTRODUCTION

In practice, deciding on investment in infrastructure uses the Net Present Value (NPV); that is, a necessary condition to accept an investment proposal is that the NPV be not negative. In developing countries this criterion is used for investments financed by development aiding agencies (World Bank, Asian Development Bank). In this paper we address the problem of uncertainty in the model's inputs and parameters, further referred to as factors. In practice most models used to analyze investments are deterministic, because no or only limited information is available on the (joint) distribution of the factors.

An additional question is: Which factors can make a project go "wrong"; that is, which factors may cause $NPV < 0$. Decision makers ask for this type of information to support their decision making process; see Van Groenendaal (1998b).

Note that information on which factors affect the NPV is useful also to evaluate implementation progress after the decision to proceed has been taken.

In applied work sensitivity analysis is limited to one factor at a time in combination with a few scenarios. For this three data points per factor are required: the base case value, and a minimum and maximum value. The resulting information is, however, insufficient to meet the decision makers needs.

Van Groenendaal and Kleijnen (1997) and Van Groenendaal (1998a) suggest to apply the statistical theory on design of experiments in combination with regression meta-modeling (further referred to as DOE) for sensitivity analysis of deterministic models. This approach requires the same information on factors as the currently used methods.

DOE is typically applied in a constructive way; that is, one starts with a simple design and estimates a simple meta-model. For example, first use a design to identify important (main) effects and to see if there are possible interactions. Only if the estimation results indicate other effects, a more complicated design is introduced. This approach is chosen to minimize the amount of work required.

In this paper we explain the different steps of DOE for deterministic models and discuss some of the hazards. To keep our analysis manageable we use a rather simple deterministic investment model, based on work done for the Asian Development Bank (ADB, 1996), instead of a complicated one.

The remainder of this paper is organized as follows. Section 2 discusses DOE in more detail. Section 3 reviews the NPV model used as a test case. Section 4 applies DOE. Section 5 contains conclusions.

2 DESIGN OF EXPERIMENTS

As argued by Van Groenendaal (1998b), the NPV-analysis of an investment problem has a typical form. Many inputs need not to be analyzed separately, but can be combined. The way they are introduced in the NPV-analysis acts as a funnel. An example is the analysis of investment cost, which in most cases is based on many inputs. In the calculation of the NPV the aggregated cost is used. It is not necessary to vary all separate inputs affecting the investment cost, the variation in the total cost suffices. If

the investment cost is important, one can always analyze separately how the different factors affect the total investment cost. We therefore assume that the number of factors is limited.

Note that for models with many factors, screening is required before the procedures discussed here can be applied. For a discussion of screening we refer to Kleijnen (1998).

We propose to apply the design of experiments to obtain information at extreme points of the experimental area mentioned earlier. (This in contrast to a Monte Carlo approach where areas with high probability are sampled more often.) The simulation results at the extreme points, together with the design matrix, are the inputs for a regression meta-model. The parameter estimates of the meta-model indicate which factors are important.

Many designs are such that the regression matrix of the standardized factor values for the meta-model has nice properties. (For an excellent discussion see Montgomery (1991).) Such properties simplify the analysis and the interpretation of results (Kleijnen and Van Groenendaal, 1992, pp 177-8).

Let us review the commonly used approach to DOE. The aim is to obtain the required information with a minimum number of simulation runs, so the first step is to estimate the main effects. Although there are many designs to choose from, a 2^{k-p} fractional factorial design (with k the number of factors and p chosen so that $k-p \geq q$, with q the number of parameters) or a Plackett-Burman design (Plackett and Burman, 1946) are often used as a starting point.

In general both these designs are Resolution III designs; that is, the estimates of the main effects are not aliased with other main effects, but they are with two-factor interactions. However, by carefully choosing the generators of a design, it is possible to minimize the work required, or even start with a Resolution IV design. In a Resolution IV design the estimates of the main effects are no longer aliased by interactions, but interactions are still aliased.

To check whether a meta-model based on input from a Resolution III design needs to be extended, Box and Wilson's (1951) fold-over can be applied. Let $\boldsymbol{D}$ be the design matrix used. The fold-over of $\boldsymbol{D}$ is defined as $-\boldsymbol{D}$, so the number of simulation runs is doubled. The effect of the fold-over is that a Resolution III design becomes a Resolution IV design. If the coefficients change it is clear that interactions are present. Because we now have twice as many simulation runs, we can estimate a limited number of interactions also. In some cases this might be sufficient; see Van Groenendaal (1998a) for an example. In case the model is inadequate we have to add more simulation runs. This can be done gradually or one can proceed to a Resolution V design, in which no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with higher order factor interactions.

In practice interactions between three or more factors are often assumed to be zero. In all our applications we never experienced significant three-factor interactions.

To evaluate the estimation result the adjusted R-square, R^2_{adj}, is normally used. However, at the start of the procedure to construct an adequate meta-model the number of data points is often close to the number of coefficients to estimate, in which case the R^2_{adj} is of limited use only.

Note that the distinction between the lack-of-fit error and the experimental error that is often used to test for "goodness of fit" can not be used here, because we have only one observation for each factor combination.

Another way to determine if the model is correct is by applying cross-validation; that is, delete one of the simulated NPV values from the data set and re-estimate the model for the reduced set. The re-estimated model is then used to forecast the deleted value. In case of (not detected) interactions the model will not be stable and the forecasts will be poor; see Kleijnen and Van Groenendaal (1992). The advantage of this approach is that no new simulation runs are required to test the model. Note that cross-validation can be applied at any stage of the procedure.

A final problem that has to be addressed is the fact that the error term in the meta-model will in general not be normally distributed, because we use the extreme points in the experimental area. To test normality of the residues we apply the Wald statistic on skewness, on kurtosis, and a combined test (Greene, 1993, pp. 309-311). All three are χ^2 distributed. If the statistics are significant, the assumption of normality of the residues has to be rejected and we cannot use the F-test on model reduction (Kleijnen, 1987, pp. 155-57). To test for model reduction; that is, $H_0 : R\beta = 0$, we use the limiting distribution of the Wald statistic $W = (R\hat{\beta})^T [R\, \hat{\sigma}^2 (X^T X)^{-1} R^T] R\hat{\beta}$, which converges to a χ^2 distribution with degrees of freedom equal to the rank of the matrix R (Greene, 1993, pp. 300-301). This Wald statistic on model reduction assumes homoscedasticity. Because we simulate extreme points, this assumption may not hold. Therefore, we tested the model reduction also assuming heteroscedasticity.

Let $(e_1^2, \ldots, e_i^2, \ldots, e_m^2)$ be the vector of squared residues, with m the number of observations, and let $\hat{\Omega}$ denotes a covariance matrix with $(e_1^2, \ldots, e_i^2, \ldots, e_m^2)$ on the diagonal and zero otherwise. The Wald statistic for the heteroscedastic model is $W = (R\hat{\beta})^T [R(X^T X)^{-1} (X^T \hat{\Omega} X)(X^T X)^{-1} R^T]^{-1} R\hat{\beta}$, which has the same χ^2limiting distribution as the homoscedastic model. Next we briefly introduce the case study.

3 STAR FARM CASE STUDY

The Chinese government sees large-scale biogas production as an opportunity to solve several problems simultaneously, namely: (i) the lack of energy in rural areas, (ii) the pollution of the environment by large breeding farms, and (iii) the lack of fertilizer for the agricultural sector. Large-scale biogas digesters using the manure of one or more breeding farms plus some crop residues help to solve these problems; they (i) produce a convenient form of energy (biogas), (ii) there is no more uncontrolled pollution by dumping manure, and (iii) the residuals of biogas production can be used as fertilizer in the production of vegetables and as an addition to fodder for other stock, such as, pigs, fish, and prawns. There are, however, a number of factors that affect the profitability of investing in large scale biogas plants. (For a complete description of the problem we refer to Van Groenendaal and Kleijnen (1998).) These factors and their base values are:

1. The shares of the different inputs in the total Z, for which the vector of base values is $(0.808,\ 0.114,\ 0.078)^T$.
2. The total amount of annual input $\sum_{i=1}^{3} Z_i$; base value is 31,000 metric ton.
3. The total investment costs $P_I{}^T\ I$; base: 4,961,000 Yuan and a building time of one year.
4. Environmental benefits A; 564,900 Yuan per year.
5. The prices of labor P_L; 4,200 Yuan per year, and the intermediary inputs water and de-sulfurizer $P_M^T = (P_W,\ P_D)$; (0.48;2,034) Yuan per unit.
6. The price of biogas P_{Q_1}; 0.8 Yuan/m_g^3, and the prices of the energy inputs electricity, diesel oil, and coal $P_E^T = (P_{electricity},\ P_{diesel\ oil},\ P_{coal})$; (0.375, 1780, 285) Yuan per unit.
7. The prices of the post-processing output liquid sludge (Q_2), fertilizer (Q_3), and fodder (Q_4) (P_{Q_2} , P_{Q_3} , P_{Q_4}); (1.627, 813.7, 537.0) Yuan per unit.
8. The efficiency τ of the biogas digesters; the base value is 1.029 m_g^3/m_d^3.

The possible changes in the base values listed above are as follows. For the factors 1, 2, 5, and 7 we set the maximum changes at ± 20%. For factor 3 the change is ± 25%, based on previous experience. Factor 4 contains 209,900 Yuan per year of avoided damages, but these are highly uncertain. Therefore, we set the change of avoided damages at ± 50%. Given the current law, the indemnities are assumed fixed. For factor 6 (energy prices) we vary the price of biogas ± 25%, whereas we vary the other energy prices ± 20%. We vary the efficiency of the bio-digesters (factor 8) ± 17%.

Note that factor 1 actually comprises two factors, factors 1a and 1b (the share of chicken dung (say) α_2 and the share of industrial waste α_3 in the total annual input (the sum of all shares (α_1 , α_2 , α_3) equals 1). We vary α_2 and α_3 in the same way; that is, if α_2 is at its maximum (minimum) than so is α_3; hence in the DOE analysis the two components have to be treated as a single factor. In the same way factors 5, 6, and 7 represent combined factors. In case of a stochastic analysis these factors would be strongly correlated.

4 SENSITIVITY ANALYSIS THROUGH DOE

For the deterministic model we denote the eight factors mentioned in Section 3 by X_i ($i = 1, \ldots, 8$); for these X_i we consider only three values: -1, 0, and +1: $X_i = -1$ indicates the low value of the range, $X_i = 0$ denotes the base case value, and $X_i = 1$ denotes the high value of the range.

Since we assume three-factor and higher order factor interactions to be zero, our meta-model is at most a second-order approximation in X. In the sequel β_0 denotes the grand mean, β_i ($i = 1, \ldots, 8$) main effects, $\beta_{i,j}$ two-factor interactions, and $\beta_{i,i}$ the quadratic effects in the approximation.

First, we apply a 2_{IV}^{8-4} design and estimate a first-order polynomial; that is, β_0 and β_i ($i = 1, \ldots, 8$) (Table 1, column 2). This Resolution IV design is obtained by choosing the generators of the design such that the best possible alias relationship is obtained (Montgomery, 1991, pp. 358-60). The first four factors are identified with the four columns, (say) d_1, d_2, d_3, and d_4, of the 2^4 full factorial design. Factors 5 till 8 are defined as: $d_5 = d_2 * d_3 * d_4$, $d_6 = d_1 * d_3 * d_4$, $d_7 = d_1 * d_2 * d_3$, and $d_8 = d_1 * d_2 * d_4$. The result is a Resolution IV design.

Table 1: First Order Regression Model

design	2_{IV}^{8-4}	2_{III}^{8-4}	$2_{IV}^{(8+1)-4}$
β_0	2574627	2575971	2575971
β_1	309834	300206	353296
β_2	913001	919721	919721
β_3	-1184953	-1026669	-1235252
β_4	650248	679051	644872
β_5	-258503	13824	-265450
β_6	1084935	1090311	1090311
β_7	317222	310502	310502
β_8	869229	876176	876176
R^2_{adj}	0.87	0.86	0.85

All main effects have the signs expected by experts. Their absolute values indicate their relative importance (because we standardized: $-1 \leq X_i \leq 1$), assuming the experimental area (the combination of factor ranges) is chosen correctly (see Kleijnen and Van Groenendaal, 1992, pp. 177-178).

We also included in Table 1 the estimation result for a case where the generators were chosen rather arbitrarily (column 3) (so the design is a 2_{III}^{8-4}) and applied a fold-over (column 4).

If we compare column 3 and column 4, the β_3 and β_5 change considerably, and β_1 to a lesser extent, indicating possible interactions. Comparing columns 2 and 3 indicates that estimates of β_3 and β_5 may be aliased.

Note that the grand mean $\hat{\beta}_0$ is for all cases almost equal to the base case value (namely NPV = 2,557,937). We now continue with the 2_{IV}^{8-4} design.

Nine coefficients are estimated and significant. However, we have 16 data points, so estimation of the combined interactions is possible without adding new simulation runs. In the design two-factor interactions are aliased with each other and there are 7 sets of aliased two-factor interactions (Montgomery, 1991, p. 631). We can estimate 16 coefficients from 16 data points, but this is stretching the use of the available information to the limit. Statistical testing is not possible.

Unless information besides the estimation results is available, we cannot become more specific about which two-factor interactions are actually important. Sometimes such information is available from earlier experiments (the model is not a black box to the experimenter (Van Groenendaal, 1998a)), or experts can rule out certain

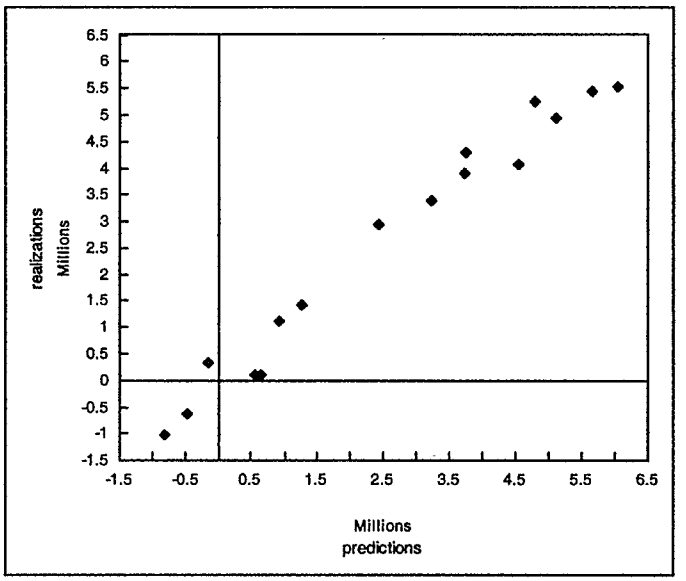

Figure 1 NPV Meta-model Predictions versus Simulation Realizations

coefficients. In this example $\beta_{2,8}$, $\beta_{2,6}$, and $\beta_{6,8}$ are expected to be significant, but it is not certain that all the other effects are zero. In case no further information is available, extra simulation runs are required to identify significant interactions.

If we assume that only $\beta_{2,8}$, $\beta_{2,6}$, and $\beta_{6,8}$ are significant we obtain: $\beta_{2,6}$ = 216987, $\beta_{2,8}$ = 173846, and $\beta_{6,8}$= 251468. Because the design is orthogonal, the estimates for the grand mean and the main effects do not change when interaction terms are added, so the estimation results in Table 1, column 2, remain the same.

To test the stability of the estimation results (including $\beta_{2,8}$, $\beta_{2,6}$, and $\beta_{6,8}$) we apply cross-validation. The result is in Figure 1, which shows that the result is acceptable, given the limited information available.

In case we had used the 2_{III}^{8-4} design to start our analysis, the fold-over would result in sufficient data to estimate a number of two-factor interactions. In this case we need to identify the exact the alias structure; see Kleijnen (1975, pp. 320-28) on how to proceed. We will not elaborate on this approach, but go to the next step in case the result is inadequate; that is, more information is required.

To further analyze the two-factor interactions we select a central composite design including a 2_V^{8-2} design. For the star design we added 10% to (subtracted 10% of) the high (low) value of the range. This design has 81 data points (the base case, plus the 64 points of the 2_V^{8-2}design, and the 16 points of the star design).

The final result of this analysis is in Table 2. All main effects remain significant; there are ten significant two-factor interactions, and no significant quadratic effects. The most important two-factor interactions are the ones already identified previously.

Table 2: Meta-model Based on a Central Composite 2_V^{8-2} Design

Coef.	Estimate	Coef.	Estimate
$\hat{\beta}_0$	2558301	$\hat{\beta}_{1,2}$	61967
$\hat{\beta}_1$	309817	$\hat{\beta}_{1,5}$	43461
$\hat{\beta}_2$	912900	$\hat{\beta}_{1,6}$	34736
$\hat{\beta}_3$	-1235250	$\hat{\beta}_{1,7}$	34178
$\hat{\beta}_4$	644872	$\hat{\beta}_{1,8}$	26877
$\hat{\beta}_5$	-264147	$\hat{\beta}_{2,5}$	-46371
$\hat{\beta}_6$	1084915	$\hat{\beta}_{2,6}$	216987
$\hat{\beta}_7$	310418	$\hat{\beta}_{2,7}$	62100
$\hat{\beta}_8$	869229	$\hat{\beta}_{2,8}$	173846
		$\hat{\beta}_{6,8}$	217307
R^2_{adj} = 0.98			

The Wald statistic on skewness, on kurtosis, and the combined test were highly significant, so the assumption of normality of the residues of our meta-model has to be rejected. Therefore, we tested the model reduction assuming homo- as well as heteroscedasticity. The Chi-square values are: χ^2_{26} = 6.86 and χ^2_{26} = 17.91 respectively, so the model reduction is accepted. The test results indicate that assuming homoscedasticity is permitted. Further reduction leads to significant W-values; that is, a loss of information.

Table 2 gives the impression that the previous model with 12 significant factors is inadequate. We should, however, keep in mind that with more data we are able to identify more effects also; effects which are not necessarily important for our goal: identify the most important threats to our investment.

5 CONCLUSIONS

In practice the NPV of investment problems is often analyzed through the use of deterministic models, because no information on the joint probability distribution function of factors is available. Sensitivity analysis of the NPV is required to help decision makers understand what can make a project go wrong. For sensitivity analysis practitioners use one factor at a time and a few scenarios. We base our sensitivity analysis on experimental design and regression meta-modeling. Our approach uses the same information about the experimental area as current practices and is relatively simple. It results, however, in better information to support decision makers.

REFERENCES

ADB. 1996. Rural energy study in the People's Republic of China. Final Report TA 2100, Asian Development Bank, Manila, Philipines.

Box, G.E.P. and K.B. Wilson. 1951. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society*, Series B, 13(1):1-38.

Greene, W.H. 1993. *Econometric Analysis.* New York: Macmillan Publishing Company.

Kleijnen, J.P.C. 1975. *Statistical Techniques in Simulation, Part II.* New York: Marcel Dekker.

Kleijnen, J.P.C. 1987. *Statistical Tools for Simulation Practitioners.* New York: Marcel Dekker.

Kleijnen, J.P.C. 1998. Experimental design for sensitivity analysis, optimization, and validation of simulation models. In *Handbook of Simulation.* Editor J. Banks. New York: Wiley.

Kleijnen, J.P.C. and W.J.H. Van Groenendaal. 1992. *Simulation: a Statistical Perspective.* Chichester: John Wiley & Sons.

Montgomery, D.C. 1991. *Design and Analysis of Experiments, 3rd Edition.* New York: Wiley.

Plackett, R.L. and J.P. Burman. 1946. The design of optimum multifactorial experiments. *Biometrika* 33:305-25.

Van Groenendaal, W.J.H. 1998a. Estimating NPV variability for deterministic models. *European Journal of Operations Research* 107(1):202-13.

Van Groenendaal, W.J.H. 1998b. *The Economic Appraisal of Natural Gas Projects.* Oxford: Oxford University Press.

Van Groenendaal, W.J.H. and J.P.C. Kleijnen. 1997. On the assessment of economic risk: factorial design versus Monte Carlo methods. *Journal of Reliability Engineering and Systems Safety* 57: 91-102.

Van Groenendaal, W.J.H. and J.P.C. Kleijnen. 1998. Deterministic versus stochastic sensitivity analysis in investment problems. Paper presented at the Second SAMO conference held in Venice on April 19-22.

AUTHOR BIOGRAPHIES

WILLEM J.H. VAN GROENENDAAL is Associate Professor in the Department of Information Systems and Auditing, and Fellow at the Center for Economic Research (CentER). Both the Department and the Center are within the School of Management and Economics of Tilburg

University in Tilburg, The Netherlands. He is also a consultant for several international development agencies. He has a Ph.D. from Tiburg University. His research interests are in simulation, decision support, and investment analysis.

JACK P.C. KLEIJNEN is Professor of Simulation and Information Systems in the Department of Information Systems and Auditing; he is also associated with the Center for Economic Research (CentER). Both the Department and the Center are within the School of Management and Economics of Tilburg University (Katholieke Universiteit Brabant) in Tilburg, Netherlands. He received his Ph.D. in Management Science at Tilburg University. His research interests are in simulation, mathematical statistics, information systems, and logistics. He published six books and more than 130 articles; lectured at numerous conferences throughout Europe, the USA, Turkey, and Israel; he was a consultant for various organizations; and is a member of several editorial boards. He spent some years in the USA, at different universities and companies. He was awarded a number of fellowships, both nationally and internationally.