Proceedings of the 1997 Winter Simulation Conference
ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

SEARCHING FOR IMPORTANT FACTORS: SEQUENTIAL BIFURCATION UNDER UNCERTAINTY

Russell C. H. Cheng

Institute of Mathematics and Statistics
The University of Kent at Canterbury
Canterbury, Kent CT2 7NF, UNITED KINGDOM

ABSTRACT

The problem of searching for important factors in a simulation model is considered when the simulation output is subject to stochastic variation. Bettonvil and Kleijnen (1996) give a method which they call sequential bifurcation which allows a large number factors to be considered using a relatively small number of simulation runs. They give the method under the assumption that the simulation response contains negligible random error, and show that when the number of important factors is small then the method is effective and efficient. In this paper the method is extended to handle simulations where the response is stochastic and subject to significant error. An attraction of the sequential bifurcation method is its flexibility in exploring the effects of different factors. The approach in this paper is to develop a clear but flexible framework in which the method is used as an exploratory tool. For illustration a numerical example is considered using a simulation metamodel involving 24 factors. The example is quite a testing one because the different factors cover a spread of values of differing importance. The results show that the method is capable of handling such situations.

1 INTRODUCTION

The output of interest in simulation studies of complex systems usually depends on a large number of factors. For example in a complex queue, the average customer delay may depend on a large number different arrival and service rates, and also on the capacity of different queues in the system, and on queueing disciplines. An important objective of the simulation study is to identify which factors are important and influence the output most strongly. Bettonvil (1990) and Bettonvil and Kleijnen (1996) point out that frequently there are only a small number of key factors which are important, even when the total number of factors is large. They review a number of *screening* methods (see Watson 1961, Jacoby and Harrison 1962, Li 1962, Morris 1987 and Cochran and Chang 1990, Welch et al. 1992). Most such methods are concerned with group screening in contexts where the experimental process is an involved one. In computer simulations however it is easy to modify the experimental runs actually during the process of accumulating the observations. It is thus possible to have a much more flexible experimental process which then determines the important factors efficiently in the sense of needing a relatively small number of simulation runs. Bettonvil and Kleijnen (1996) propose a novel method for doing this called *sequential bifurcation* (SB), which may be thought of as a generalisation of classical binary search (the objective of this latter technique being to find the single most important factor). They develop the method for deterministic simulations, which are often appropriate, for example, in studies involving systems dynamics and investment analysis.

In this paper we extend the the method to cover the situation where the simulation output is stochastic.

Usually such screening methods are used at an initial exploratory phase, where the effect of each given factor does not have to be determined too precisely. We adopt this viewpoint. We propose an explicit version of SB that is robust and flexible. The method divides the factors into two sets: those which are deemed important, and those which are not. One of the differences between a deterministic and a stochastic output is that in the latter case a disproportionate amount of effort can go into investigating factors which are borderline in importance. This seems particularly inappropriate in exploratory work, when importance might be somewhat inexactly defined. Our method therefore allows for an 'indifference-zone'. If the importance of a factor is estimated to fall inside this indifference zone, then no further effort is wasted in attempting to estimate this factor effect more accurately. Thus much computing effort is saved. For

simplicity we classify such factors as unimportant, though a more cautious approach could be taken. Our reasoning is that even if more accurate estimation might categorise them as important, they will only be marginally so. Thus this misclassification, if handled with due care, will be of little practical importance.

The main difficulty with sequential methods when observations are subject to uncertainty is establishing the level of significance of results. It is hoped to give a more detailed theoretical analysis of the methodology elsewhere. Here, the approach is to treat the method as an exploratory one where the flexibility of the method is not sacrificed for the sake of achieving a notionally precise level of significance. We adopt the approach typically used in practice in such problems and set a notional significance level for individual tests. This notional level of significance is then set sufficiently high to offset what is usually a degradation in the overall significance level through repeated use of the individual test. This is the approach used in stepwise regression for example.

In this paper we focus on presenting the practical methodology and discuss its application in an example involving a simulation experiment with a regression model where there are 24 factors. It is shown that the technique works effectively and is easy to apply.

2. THE SIMULATION MODEL

2.1. Model Structure

The simulation experiment is assumed to be built up sequentially using runs of equal length. The output of interest from a run is denoted by y. There are two quantities which affect the yalue of y that we focus attention on:

(i) a vector of decision variables

$$\mathbf{x} = (x_1,\ x_2,\ ...,\ x_K) \tag{1}$$

which are under the control of the experimenter, and

(ii) a set of random numbers

$$\mathbf{u} = (u_1,\ u_2,\ ...,\ u_n) \tag{2}$$

which (typically after a transformation that is not of concern here) form the stochastic input driving the simulation run. (Usually n will be a randomly varying quantity in its own right. However we shall consider only the case where y is a long run average, so that n will be approximately linearly dependent on the length of the simulation run. We therefore tacitly equate n with the run length and treat it as being essentially known.) Thus y can be regarded as a function of just $\mathbf{x}$ and $\mathbf{u}$:

$$y = y(\mathbf{x}, \mathbf{u}). \tag{3}$$

We can therefore adopt the following regression metamodel:

$$y(\mathbf{x}, \mathbf{u}) = \eta(\mathbf{x}) + \ e(\mathbf{x}, \mathbf{u}) \tag{4}$$

where $e(\mathbf{x}, \mathbf{u})$ is regarded as an 'error' term containing all the chance variation of y; and moreover assume that

$$E\left[e(\mathbf{x}, \mathbf{u})\right] = 0,$$
$$Var\left[e(\mathbf{x}, \mathbf{u})\right] = \sigma^2(\mathbf{x}).$$

Thus

$$\eta(\mathbf{x}) = E(y|\mathbf{x}) = \int y(\mathbf{x}, \mathbf{u})\mathbf{du} \tag{5}$$

is the expected value of y.

The precise distribution of $e(\mathbf{x}, \mathbf{u})$ is not known of course. However for stationary processes where y is a long term average then $e(\mathbf{x}, \mathbf{u})$ can be expected to be approximately normally distributed, and we assume this in what follows.

Our objective is to investigate how η depends on $\mathbf{x}$. We focus on the local behaviour of η about some notional operating point

$$\mathbf{x}^0 = (x_1^0,\ x_2^0,\ ...,\ x_K^0).$$

We confine our discussion to small changes of $\mathbf{x}$ about $\mathbf{x}^0$. Assuming that $\eta(\mathbf{x})$ is continuously differentiable, we can adopt the linear approximation

$$\eta(x) = \eta_0 + \sum_{i=1}^{K} \beta_i (x_i - x_i^0). \tag{6}$$

The coefficients $\beta_i = \partial\eta(\mathbf{x})/\partial x_i|_{\mathbf{x}^0}\ i = 1, 2, ..., K$ will be called the *sensitivity coefficients*.

Bettonvil and Kleijnen (1996) consider the situation where K is large, but only a few, k say, of the β_i are important in the sense of being large, and so influential on how $\eta(\mathbf{x})$ varies. The objective is to determine these k important factors without having to make too many runs. Bettonvil and Kleijnen also assume that the simulation error $e(\mathbf{x}, \mathbf{u})$ is small and can be neglected. We consider the situation where $e(\mathbf{x}, \mathbf{u})$ cannot be neglected. We shall however assume that locally the variability of $e(\mathbf{x}, \mathbf{u})$ does not depend on $\mathbf{x}$, that is we assume that

$$\sigma^2(\mathbf{x}) = \sigma^2,\ \ \textit{is independent of}\ \mathbf{x}.$$

The values of the β_i will be unknown, but in most situations it is quite realistic to assume that their *sign* is known. For example consider a queueing model where η is the average customer delay and the x_i are

the arrival and service rates; here it is fairly clear whether a change in x_i will increase or decrease η. Moreover, simply by reversing the sign of x_i itself where necessary, it may be assumed that:

$$\beta_i \geq 0 \; for \; all \; i. \tag{7}$$

We shall assume this from now on.

2.2- Estimating the Sensitivity Coefficients

In SB the decision variables are assumed to be listed in some fixed order. This order will be maintained throughout the ensuing calculations.

Each run is made by selecting a value of j and then making the run with $\mathbf{x}$ set at $\mathbf{x}^{(j)}$ with components

$$\begin{aligned} x_i^{(j)} &= x_i^0 + \delta, \quad i = 1,\ 2,\ ...,\ j \\ x_i^{(j)} &= x_i^0, \qquad i = j+1,\ j+2,\ ...,\ K \end{aligned}$$

We denote the output from the run made at this value of $\mathbf{x}$ by

$$y(\mathbf{x}^{(j)}, \mathbf{u}) = y^{(j)}; \tag{8}$$

with $y^{(0)}$ the value obtained from a run made with $\mathbf{x} = \mathbf{x}^{(0)}$. For convenience we refer to $y^{(j)}$ as a run made at *level* j.

If $j < k$ then the scaled difference

$$D(j,k) = [y^{(k)} - y^{(j)}]/\delta$$

has expectation

$$E[D(j,k)] = \sum_{i=j+1}^{k} \beta_i,$$

so that $D(j,k)$ can be regarded as an estimator of the sum of the sensitivity coefficients running from $j+1$ to k. If the runs $y^{(j)}$ and $y^{(k)}$ are independent, then

$$Var[D(j,k)] = 2\tau^2,$$

where $\tau^2 = \sigma^2/\delta^2$.

The selection of δ is an interesting problem in its own right, but will not be discussed here. We shall set it to a (small) value equal to the value that we would wish to vary x_i by in the practical application. This at least partially addresses the problem of non-linearities in the response not being properly allowed for, in that it considers y at settings of x_i that are of direct interest.

In the following it will sometimes be necessary to improve the accuracy $y^{(j)}$. This will be done simply by replication. In general we may have $r^{(j)}$ observations at level j : $y_i^{(j)}$ $i = 1, 2, ..., r^{(j)}$. If we use

$$\bar{D}(j,k) = (\sum_{i=1}^{r^{(k)}} y_i^{(k)}/r^{(k)} - \sum_{i=1}^{r^{(j)}} y_i^{(j)}/r^{(j)})/\delta$$

then this has improved variance

$$Var[\bar{D}(j,k)] = \tau^2(1/r^{(j)} + 1/r^{(k)}).$$

3- SB UNDER UNCERTAINTY

3.1- The Proposed Method

We shall only consider one specific, prototype, version of the problem. The primary objective is to divide the sensitivity coefficients into just two groups: those for which $\beta_j > B$ and those for which $0 \leq \beta_j \leq B$:

$$I = \{\beta_j :\ \beta_j > B\},\ \ U = \{\beta_j :\ \beta_j \leq B\}.$$

Here B is some prescribed value selected by the user, depending on the problem. Bettonvil and Kleijnen consider the case $B = 0$, calling those $\beta_j > 0$ important. We extend this terminology by calling those $\beta_j \in I$ *important*, and those $\beta_j \in U$ *unimportant*.

As mentioned in the Introduction, much simulation effort can be spent on borderline cases, where β_j is close to B, to determine if β_j just falls above or below B. We avoid doing this by allowing for an indifference-zone $(0, B+a)$. If β_j is estimated as being located within this zone, it is automatically classified as being in U. This greatly saves on computing effort for such cases. If a is set small relative to B, then a misclassification where β_j is wrongly put in U, when actually $B < \beta_j$, can be assumed to be of little practical consequence, as we will know that $\beta_j < B + a$.

Our search process, though varying in detail, follows the basic precept of the SB process given by Bettonvil and Kleijnen. The decision variables are kept in their initial order throughout. The key idea is that if we find that

$$\sum_{j=k_1}^{k_2} \beta_j < B \tag{9}$$

then, because $\beta_j > 0$ for all j, we must have $\beta_j < B$ $for\ j = k_1,\ k_1 = 1,\ ..., k_2$. Thus it is not necessary to test for the importance of individual β_j when (9) is true. The overall method is to partition the coefficients into contiguous sets G_i, $i = 1, 2, ..., p$ where all the coefficients of each set have the same classification of unimportant or important.

The method is as follows.

Algorithm for SB under Uncertainty

(0) To initiate the process we set the step counter, s say, to $s = 1$. We then make $r^{(0)} > 1$ and $r^{(K)} > 1$ (typically $r^{(0)} = r^{(K)} =$ a small number, between 2 and 5, say) runs at levels 0 and K respectively. Thus at the start of Step 1 we have two sets of observations:

$$\{y_j^{(k_{i1})} :\ j = 1, ..., r^{(k_{i1})}\},\ \ i = 0, 1,$$

where $k_{01} = 0$ and $k_{11} = K$ denotes the initial two levels at which runs have been made. We also place all the coefficients in the single,unclassified, set $G_{1s} = \{\beta_1, \beta_2, ..., \beta_K\}$. Thus initially the number of sets is $p\ (= p_1) = 1$.

We now progressively select sets to partition and classify in the following way.

(1) At the start of Step s we have $p_s + 1$ sets of observations:

$$\{y_j^{(k_{is})} :\ j = 1, ..., r^{(k_{is})}\},\ \ i = 0, 1, 2, ..., p_s, \quad (10)$$

(where all the observations $\{y_j^{(k_{is})} :\ j = 1, ..., r^{(k_{is})}\}$ of the ith set are made at the same level k_{is}, i.e. at the same decision variable setting $\mathbf{x}^{(k_{is})}$). Moreover the decision variables are partitioned into p_s contiguous sets:

$$G_{is} = \{\beta_j : k_{(i-1),s} < j \leq\ k_{is}\},\ \ i = 1, 2, ..., p_s.$$

Some of the sets are already classified, some are unclassified. If all sets have been classified the algorithm ends. Otherwise we select any unclassified set (the one with largest cardinality, say): call this G_{is}.

(2) If G_{is} is not a singleton set we see if all the coefficients can be classified as unimportant. This is done by considering $\bar{D}(k_{(i-1),s},\ k_{is})$. From the assumption that $e(\mathbf{x}, \mathbf{u})$ is normally distributed, $\bar{D}(\ k_{(i-1),s},\ k_{is})$ is also normally distributed, with mean and variance

$$\mu_{is} = \sum_{i=k_{(i-1),s}+1}^{k_{is}} \beta_i, \quad v_{is} = \tau^2(1/r^{(k_{(i-1),s})} + 1/r^{(k_{is})}).$$

We can estimate τ^2 from each set (10):

$$t_{is}^2 = \sum_{j=1}^{r^{(k_{is})}} (y_j^{(k_{is})} - \bar{y}^{(k_{is})})^2/[(r^{(k_{is})} - 1)\delta^2],$$

and pooling these estimates gives an overall estimator of τ^2 at the start of the Step s as:

$$t_s^2 = [\sum_{i=0}^{s}(r^{(k_{is})} - 1)t_{is}^2]/(\sum_{i=0}^{s} r^{(k_{is})} - s - 1).$$

If z_α is the upper α-quantile of the standard normal distribution, then with probability approximately $(1-\alpha)$:

$$\mu_{is} < \bar{D}(k_{(i-1),s},\ k_{is}) + z_\alpha t_s(1/r^{(k_{(i-1),s})} + 1/r^{(k_{is})})^{\frac{1}{2}}.$$

If therefore

$$\bar{D}(k_{(i-1),s},\ k_{is}) < B - z_\alpha t_s(1/r^{(k_{(i-1),s})} + 1/r^{(k_{is})})^{\frac{1}{2}} \quad (11)$$

then with confidence $(1 - \alpha)$:

$$\mu_{is} < B, \quad i.e. \sum_{i=k_{(i-1),s}+1}^{k_{is}} \beta_j < B.$$

As the β_j are all positive this means

$$\beta_j < B, \quad j = k_{(i-1)s} + 1,\ k_{(i-1),s} + 2,\ ...,\ k_{is}.$$

and the entire set of sensitivity coefficients, G_{is}, can thus be classified as unimportant.

If (11) is not satisfied we split G_{is} into two at $k = \lceil (k_{(i-1),s} + k_{is})/2 \rceil$ (if the cardinality of the set is odd, the split gives the set with the smaller indices one more member than the other). The two new sets:

$$\{\beta_{k_{(i-1),s}}, ..., \beta_k\}\ and\ \ \{\beta_{k+1}, ..., \beta_{k_{is}}\}$$

replace G_{is}.

Finally a run is made at k. We then increment s and repeat the process from (1).

(3) If however $G_{is} = \{\beta_k\}$, i.e. it is a singleton set, then we proceed to fully classify β_k. This is done by calculating a two-sided $(1-\alpha)$ confidence interval for β_k with upper and lower limits given by:

$$\beta_k^{\pm} = \bar{D}(k-1,\ k) \pm z_\alpha t_s(1/r^{(k_{(i-1),s})} + 1/r^{(k_{is})})^{\frac{1}{2}}. \quad (12)$$

If B is contained in this interval we make additional runs at the levels $k-1$ and k. If $r^{(k_{is})}$ and $r^{(k_{(i+1),s})}$ are not initially equal we add the runs at the level with the smaller number of runs, until $r^{(k_{is})} = r^{(k_{(i+1),s})}$, and then add runs at both levels $k-1$ and k, keeping $r^{(k_{is})} = r^{(k_{(i+1),s})}$. As runs are added the length of the confidence interval (12) decreases. We stop when either

(i) $B < \beta_k^-$ when β_k is classified as important,

or

(ii) $\beta_k^+ < B$ when β_k is classified as unimportant,

or

(iii) $\beta_k^+ < B + a$, (where a is small relative to B) when β_k is deemed determined to be sufficiently close to B to be classified as unimportant.

Once β_k is classified we increment s and repeat from (1).

It will be seen that at the end of the process every β_j will have been classified as either important or unimportant. Moreover all β_j classified as important will have a corresponding confidence interval calculated.

3.2- Properties

The final probability of correctly classifying all β_j is not easily determined as the procedure is sequential.

However the procedure used in the calculation is not unusual for such processes, and should be sufficiently robust to give a satisfactory practical procedure. An analogous case occurs in polynomial regression where the degree of the polynomial to be fitted is unknown and has to be estimated. Our methodology operates in essentially the same type of way as standard forward or backward stepwise selection methods which use single step probability calculations to decide on whether to proceed to the next step or not.

The main property of the method is that it allows large groups of variables to be eliminated from consideration if they are unimportant. The key controlling factor is the setting of the probability level in the test given in equation (11). A high level (i.e. small α value) makes high the probability of correctly determining the important and unimportant sets I and U but at the expense of additional observations being needed. A low level (i.e. large value of α) makes the overall determination process faster but with a higher risk of variables being assigned to the incorrect set.

Overall however the technique should be an effective and efficient one for exploratory purposes. We illustrate with a detailed example.

4- REGRESSION EXAMPLE

Bettonvil and Kleijnen give a deterministic regression metamodel example involving 24 factors. We modify the problem to include stochastic variation. Our model is given in equation (4), where $e \sim N(0, 30^2)$ and η is as given in equation (6) with $K = 24$. The individual β_j used in the experiment are given in Table 1.

Table 1: Coefficients β_j of the Regression Metamodel

j	1	2	3	4	5	6
β_j	5.7-	0.6-	4.5-	19.3-	9.1-	28.4
j	7	8	9	10	11	12
β_j	51.0-	34.6-	23.3-	39.7-	45.4-	93.0
j	13-	14-	15-	16-	17-	18
β_j	73.7-	144.6-	130.4-	42.5-	313.0-	166.2
j	19-	20-	21-	22-	23-	24
β_j	76.6-	345.4-	195.7-	188.8-	148.0-	206.4

The settings used for the other factors were as follows: $\delta = 1$, $\alpha = 0.05$, $a = 25.0$. The initial run settings were $r^{(0)} = r^{(24)} = 2$. Table 2 gives the results for three different settings of B. Also shown are the results from an equivalent experiment where SB was not used. Instead two runs were initially made at all levels 0 through 24. The confidence interval calculation of stage (3) in the algorithm above was then applied to each β_j to determine if each is important or not, using the same criterion as applied in the SB method. Examination of the true values of the β_j shows that none of the misclassifications are serious errors. It will be seen that the SB method requires proportionately fewer runs, relative to the basic method as the importance threshold B is raised. At the highest level $B = 200$, the basic method needed over two and a half times as many runs to carry out the classification. This is very much as expected, as the SB method is designed to handle situations where there are only a few important factors.

Table 2: Classification of β_j in the Regression Metamodel

		SB Method		Basic Method	
B	Import. Coeffs	Mis-class	No. Runs	Mis-class	No. runs
100	β_{14} β_{15} β_{23} + below	None	48	None	68
150	β_{18} β_{21} β_{22} + below	β_{14} β_{18}	35	β_{18}	86
200	β_{17} β_{20} β_{24}	β_{24}	25	None	64

5 CONCLUSIONS

The paper has focused on a specific implementation of the method of sequential bifurcation under uncertainty. Variations of the method suggest themselves, and a number of interesting properties need to be investigated. Most obvious is the robustness of the method. The efficiency clearly depends on order in which the sensitivity coefficients are placed in the list, and on their relative sizes. The best case scenario occurs where there is a sharp distinction between important and unimportant coefficients, and where the important coefficients are placed close together in the initial list. In this case large numbers of unimportant coefficients will be rapidly removed in just a few steps. The worst case scenario is where the important coefficients are evenly spread over the initial list, and where there are are many of them. In this latter situation the SB method reduces in effect to the basic method where each coefficient is individually estimated.

One possibility of interest is to use prior information or knowledge in assigning the coefficients their positions in the list, and it is hoped to discuss this elsewhere.

The example is a reasonable test in that the values of the coefficients vary over a wide range, and their position in the list used for the experiment reported previously might be typical of an initial ordering based on prior knowledge. Thus the order is roughly correct but not exactly right.

The ease of implementation and its generality of application are the most attractive features of the method and the potential efficiency gains make it worth serious consideration.

REFERENCES

Bettonvil, B. 1990. *Detection of important factors by sequential bifurcation.* Tilburg: Tilburg University Press.

Bettonvil, B., and J.P.C. Kleijnen. 1996. Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operations Research.* (To appear.)

Cochran, J.K., and J. Chang. 1990. Optimization of multivariate simulation output models using group screening method. *Computers in Industrial Engineering* 18:95–103.

Li, C.H. 1962. A sequential method for screening experimental variables. *Journal of the American Statistical Association* 57:455–477.

Morris, M.D. 1987. Two-stage factor screening procedures using multiple grouping assignments. *Communications in Statistics: Theory and Methods* 16:3051–3067.

Welch, J.W., R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, and M.D. Morris. 1992. Screening, predicting and computer experiments. *Technometrics* 34:15–25.

Watson, G.S. 1961. A study of the group screening method. *Technometrics* 3:371–388.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Professor of Operational Research in the Canterbury Business School at the University of Kent at Canterbury. He has an M.A. from Cambridge University, England. He obtained his Ph.D. from Bath University. He is Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He is Joint Editor of the IMA Journal on Mathematics Applied to Business and Industry, and an Associate Editor for *Management Science.*