Factor Screening in Simulation Experiments: Review of Sequential Bifurcation

Jack P. C. Kleijnen

Abstract Factor screening means searching for the most important factors (or inputs) among the many factors that may be varied in an experiment with a real or a simulated system. This chapter gives a review of Sequential Bifurcation (SB), which is a screening method for simulation experiments in which many factors may be varied. SB is most efficient and effective if its assumptions are satisfied. SB was originally studied back in 1990. This review first summarizes SB. Then it summarizes a recent case study, namely, a supply-chain simulation with 92 factors where SB identifies a shortlist with 10 factors after simulating only 19 combinations. The review also references recent research. It ends with a discussion of possible topics for future research.

1 Introduction

Factor screening (or briefly "screening") means that the analysts are searching for the most important factors among the many factors that can be varied in their experiment. In practice, however, experiments with real-world systems usually can vary only a few factors, whereas experiments with simulation models can indeed vary hundreds or more factors (also see Kleijnen et al. 2005). In general, scientists assume that effects are "sparse"; they do not wish to report that "everything depends on everything else." The scientists' clients do not want to be "confused by details." Furthermore, philosophy of science exploits the *parsimony* principle or *Occam's razor*, which implies that a simpler explanation is preferred to a more complex explanation—all other things being equal. The psychologist Miller (1956) claims that people cannot handle more than "seven plus or minus two" factors when processing information. Many simulation modelers assume that the *Pareto* principle or *20-80* rule holds, i.e., only a few factors are really important (or "active", as some authors say). Many authors on simulation modeling (be it deterministic or random simulation) mention the *curse of dimensionality*; see, e.g., the panel discussion

J.P.C. Kleijnen (✉)
Department of Information Management, Tilburg University, 5000 LE Tilburg, Netherlands
e-mail: kleijnen@uvt.nl

C. Alexopoulos et al. (eds.), *Advancing the Frontiers of Simulation: A Festschrift in Honor of George Samuel Fishman*, International Series in Operations Research & Management Science 133, DOI 10.1007/b110059_8, © Springer Science+Business Media, LLC 2009

reported by Simpson et al. (2004). Altogether, screening is necessary in realistic simulation modeling. Unfortunately, practitioners do not yet apply screening methods; instead, they experiment with a few intuitively selected factors only. Hopefully, this chapter contributes to a change of attitude.

A practical example may illustrate the need for screening.Bettonvil and Kleijnen (1996) summarize a case study on the CO_2 greenhouse effect, using a deterministic simulation model with 281 factors (that simulation is also discussed by Kleijnen et al. 1992). The politicians wanted to take measures to reduce the release of CO_2; they realized that they should start with legislation for a limited number of factors. Another example will be summarized in Section 3.3.

A most efficient and effective screening method may be *Sequential Bifurcation (SB)*. After a slow start, SB has gained momentum in recent years—as the following overview shows.

1. Bettonvil's (1990) Ph.D. dissertation introduced SB.
2. Bettonvil and Kleijnen (1996) provide a summary of that dissertation (after experiencing rather much delay, finding a publication outlet).
3. Recently, SB has attracted the attention of several researchers in the UK and USA:

   - Cheng (1997) further explores SB for random simulation; also see Cheng and Holland (1999).
   - Wan et al. (2006a) improve SB's control of the type-I (or type-α) and type-II (or type-β) error probabilities in discrete-event simulation. Next, Wan et al. (2006b) extend their method to account for interactions between the factors of the simulation model. Shen and Wan (2006) combine this approach with classic fractional factorial designs. Finally, Xu et al. (2007) extend SB to nonnormal distributions, namely, binary responses (outputs) of simulation models for software reliability studies.
   - Ankenman et al. (2006) develop a more efficient but also more complicated SB variant based on "polytopes," requiring repeated solution of a sequence of Linear Programming (LP) problems.
   - Kleijnen et al. (2006) summarize SB and apply SB to a practical discrete-event simulation of a supply chain centered around the Ericsson company in Sweden, involving 92 factors; they identify a shortlist with 10 factors after simulating only 19 combinations. (This chapter updates the summary in Kleijnen et al. 2006, and summarizes their case study.)

The rest of this chapter is organized as follows. Section 2 summarizes several screening methods that may compete with SB. Section 3 summarizes SB, including its assumptions. Section 3.1 gives an outline of the simplest type of SB. Section 3.2 covers some mathematical details of this simplest SB. Section 3.3 summarizes a case study, namely, a supply-chain simulation for Ericsson in Sweden. Section 3.4 extends SB, accounting for two-factor interactions. Section 4 presents conclusions and possible topics for future research. Many references are given to enable further study of screening methods.

2 SB's Competitors

There are several types of screening designs. All these designs treat the simulation model as a *black box*; i.e., the simulation model transforms observable inputs into observable outputs, whereas the values of internal variables and specific functions implied by the simulation's computer modules are unobservable.

The importance of factors depends on the *experimental domain*—also called the experimental area or experimental frame; see Zeigler et al. (2000). The users should supply information on this domain, including realistic ranges of the individual factors and limits on the admissible factor combinations (or "scenarios"); e.g., factor combinations are admissible only if they add up to 100% because these combinations represent chemical compositions (also called recipes). In practice, *user involvement* is therefore crucial for the application of screening methods.

This section summarizes several design types that may be considered to compete with SB (empirical comparisons of these various designs is beyond the scope of this chapter).

2.1 Classic Two-Level Factorial Designs

Classic two-level factorial designs are often considered to provide screening designs; these designs are detailed in many textbooks, e.g., Kleijnen (2008) and Myers and Montgomery (2002). In particular, so-called resolution-III designs are often called screening designs; see, e.g., Georgiou (2007) and Yu (2007). By definition, a *resolution-III* design gives unbiased estimators of all the main effects or first-order effects, provided a first-order polynomial is a "valid metamodel" ("adequate approximation") of the Input/Output (I/O) function that is implicitly determined by the underlying simulation model. These designs require the simulation of at least $n = k + 1$ factor combinations where k denotes the number of factors in the experiment. In such a design, each factor has two values or levels; these levels may denote quantitative or qualitative values. The case study reported in Section 3.3 has $k = 92$ factors, so at least 93 factor combinations would need to be simulated; simulating one combination takes 40 minutes (after modification of the simulation code, which originally took 3 hours per combination). Moreover, a random simulation model like this case study requires replication of each combination to obtain an estimate of the *signal/noise ratio*; i.e., if the *noise* (variance of the simulation output) is large compared with the *signal* (the simplest signal is the factor's main effect), then *replication* (simulation with non-overlapping PseudoRandom Numbers, PRNs) is unavoidable.

Note: A different type of resolution-III design changes only one factor at a time. In the supply-chain example, such a design still requires 93 combinations if not more than two values per factor are simulated (simulation practitioners often study three values per factor, when changing one factor at a time). Moreover, this approach is less efficient; i.e., the variances of the estimated main effects are larger than the variances resulting from the Design Of Experiments (DOE) literature.

Note: A full-factorial design is often used by practitioners when the number of factors is small. However, full factorials are impossible in screening studies; e.g., the supply-chain example with its 92 factors would require $2^{92} \approx 5 \times 10^{27}$ combinations.

Another class of designs called "conference designs" requires $n = 2k$ combinations; see Elster and Neumaier (1995). These designs are not practical screening designs if the simulation model is *expensive*; i.e., a single run with the model takes relatively much computer time.

2.2 Frequency Domain Experimentation (FDE)

Whereas classic designs keep the factor values constant during a simulation run, FDE oscillates these levels during a run. More precisely, each factor has its own oscillation frequency. FDE tries to find which input oscillations affect output oscillations. Originally, Schruben and Cogliano (1987) proposed this approach. Sanchez et al. (2006) apply FDE for second-order polynomial metamodels with an arbitrary number of factors; they give an example of a kanban simulation with 34 factors. Unfortunately, FDE requires rather complicated Fourier spectral analysis. Moreover, FDE has not yet been applied to simulation models with hundreds of factors.

2.3 Supersaturated Designs

By definition, supersaturated designs have fewer combinations than factors: $n < k$. These designs are not sequential, so they are relatively inefficient. Indeed, sequential statistical procedures are known to require fewer observations than fixed-sample (one-shot) procedures; see, e.g., Park et al. (2002). By definition, sequential designs imply that observations are analyzed—so the data generating process is better understood—before the next input combination is selected. This property implies that the design depends on the specific underlying process (simulation model); i.e., the design is customized (tailored or application-driven, not generic; also see Kleijnen and Van Beers 2004 and Van Beers and Kleijnen 2008). Moreover, computer experiments (unlike many real-world experiments) proceed sequentially. Nevertheless, sequential procedures may lose some efficiency; e.g., switching between the SB procedure and the simulation model may be awkward. Recent discussions of supersaturated designs are presented in Allen and Bernshteyn (2003), Gilmour (2006), Li and Li (2005), Wu and Hamada (2000), Yamada et al. (2006), and Zhang et al. (2007). Note that Tu and Jones (2003) also give a supersaturated design, but they use Moving Least Squares instead of classic linear regression analysis.

2.4 Group-Screening Designs

Group-screening designs aggregate (or confound) individual factors into groups so that k factors may be evaluated in $n < k$ combinations. Consequently, these designs are also supersaturated—but they are executed in two or more steps (stages).

There are several types of screening designs. Examples are One-factor-At-a-Time (OAT), Morris's OAT, Cotter's design, Andres's Iterated Fractional Factorial Design (IFFD), multi-stage group screening, and SB; see Andres (1997), Andres and Hajas (1993), Campolongo et al. (2007), Campolongo et al. (2000), De Vos et al. (2006), Morris (2006), Saltelli et al. (2004, 2005), and Schonlau and Welch (2006). Note that Chipman (2006) gives a Bayesian analysis of screening experiments, but Bayesian approaches are not further considered in this chapter. The following web address gives access to a package (written in the R statistical software tool) that implements Morris's OAT.

```
http://cran.r-project.org/src/contrib/Descriptions/.
sensitivity.html
```

Different group-screening designs are based on different mathematical assumptions concerning the *characteristics* of their metamodels; e.g., their *smoothness* and *monotonicity*. Reviewing the assumptions and procedures of all these designs is beyond the scope of this chapter. This chapter focuses on SB because SB is a very efficient and effective method if its assumptions are satisfied.

3 Sequential Bifurcation

SB uses the following metamodel assumptions, which will be detailed in the next subsections; Assumption 1(b) may replace Assumption 1(a).

Assumption 1(a): a first-order polynomial is a valid metamodel.
Assumption 1(b): a first-order polynomial augmented with two-factor interactions is a valid metamodel.
Assumption 2: all main effects have known signs and are nonnegative.
Assumption 3: there is "strong heredity" if Assumption 1(b) holds.

3.1 Outline of Simplest SB

As its name suggests, SB is a *sequential* procedure. Its first step aggregates all factors into a single group, and tests whether or not that group of factors has an important effect (this statistical test will be presented in Section 3.2). If that group indeed has an important effect (which is most likely), then the second step *splits* the group into two subgroups—*bifurcates*—and tests each of these subgroups for importance. The next steps continue in a similar way; i.e., SB splits important subgroups into smaller subgroups, and discards unimportant subgroups. In the final steps, all individual factors that are not in subgroups identified as unimportant are estimated and tested—which terminates the procedure.

The simplest type of SB is based on Assumptions 1(a) and 2, which are now detailed.

Assumption 1(a) a valid metamodel is a *first-order polynomial* plus noise:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e. \tag{1}$$

In this equation, the factors x_j ($j = 1, \ldots, k$) are standardized such that they are either -1 or $+1$ (scaling in DOE is further discussed by Kleijnen 2008). This scaling implies that the factors may be ranked (sorted) by $|\beta_j|$; i.e., the most important factor is the one with the largest absolute value of its main effect; the least important factor is the one with the effect closest to zero. Note that the larger the range of an original (untransformed) factor is, the larger the response difference and hence the main effect of that factor is; also see the "unit cost" effects in Cheng and Holland (1999). The noise e in (1) arises from both approximation error and the use of PRNs. If the metamodel is *valid*, then this noise has zero mean: $E(e) = 0$.

To estimate the parameters in (1), it is most efficient to experiment with only two levels per factor. In practice, it is important that these levels are realistic *extreme* values; so the users of the underlying simulation model should provide these values. Readers are also referred to the discussion on scaling in Wan et al. (2006a) and the discussion on the experimental domain above, in Section 2.

Assumption 2 all main effects have *known signs* and are non-negative:

$$\beta_j \geq 0 \; (j = 1, \ldots, k).$$

Without Assumption 2, main effects within a (sub)group might cancel each other. However, if Assumption 2 holds, then the analysts can define the two levels of an individual factor such that changing the level from the standardized value -1 to $+1$ does not decrease the expected simulation output (i.e., that change either increases the output or does not change it at all). An example is the M/M/1 model: if the arrival rate increases, then the expected steady-state waiting time also increases; if the queuing discipline changes from First-In-First-Out (FIFO) to Shortest-Processing-Time-first (SPT), then the expected waiting time decreases; consequently, the level -1 should correspond to SPT and the level $+1$ to FIFO.

Assumption 2 is related to the *monotonicity* of the I/O function. By definition, a function $w = f(x_1, \ldots, x_k)$ is monotonically increasing if $\partial w / \partial x_j > 0$ for all j, for all values of $x_{j'}$ ($j, j' = 1, \ldots, k$; $j \neq j'$). Experience shows that Assumption 2 may be easily satisfied in practice; i.e., it is straightforward to define the upper and lower level of each factor such that changing a factor from its lower to its upper level does not decrease the expected response. For example, in the supply-chain case study, some factors refer to transportation speeds: the higher these speeds, the lower the Work In Process (WIP) and hence the lower the cost—which is the output of interest in the screening experiment. More examples are given by other authors; e.g., Lewis and Dean (2001) and Lim and Glynn (2006).

In unconstrained optimization, the function to be maximized or minimized is usually assumed not to be monotonically increasing (otherwise, the maximum or minimum would lie at the limits of the experimental area). This assumption may still be compatible with the known signs assumption; i.e., switching the standardized

factor values from -1 to $+1$ may increase the output, so this factor will be found to have an important effect. However, a counterexample is an I/O function that is not monotonic and happens to give roughly the same output values at the two observed input levels -1 and $+1$; in this example, the factor effect seems to be zero and SB eliminates this factor; also see Kleijnen (2008, pp. 162–163).

Nevertheless, if a particular case study does not satisfy Assumption 2 for a few specific factors, then these factors should be treated *individually*; i.e., none of these factors should be grouped with other factors in SB. For example, De Vos et al. (2006) create some subgroups of size one in a multi-stage group-screening design; their design is less efficient than SB, but it also uses aggregation. In general, treating such factors individually is safer than assuming that the probability of cancellation within a subgroup is negligible.

The *efficiency* of SB (measured by the number of simulated factor combinations and hence simulation time) may be improved in the following ways.

- The individual factors are labeled such that factors are placed in *increasing order of importance*; see Bettonvil (1990, p. 44). This labelling makes the important factors clustered. To realize this labelling, it is crucial to utilize prior knowledge of users and analysts about the real system being simulated. For example, if the analysts conjecture that environmental factors are most important, then they should place these factors at the end of their list of factors. Indeed, in the supply-chain case study, Kleijnen et al. (2006) place the environmental factor "demand" at the very end of the list with 92 individual factors; Section 3.3 returns to this labelling.
- *Similar factors* are placed within the same subgroup. In the supply-chain study, all "test yield" factors are grouped together; the conjecture is that if one yield factor is unimportant, then all yield factors are likely to be unimportant too.
- Subgroups are split such that the number of factors for the first new subgroup is a *power of two*; e.g., split the first 48 factors into a subgroup of 32 ($= 2^5$) factors and a subgroup with the 16 remaining factors (so the important factors are placed into the smallest subgroup, assuming the factors are sorted from unimportant to most important). This splitting, however, is not recommended if it implies splitting up a group of related factors. In any case, splitting a subgroup into subgroups of *equal* size (as some authors do) does not need to be optimal. Further discussion is found in Bettonvil (1990, pp. 40–43).

The way SB proceeds may be interpreted through the following *metaphor*. Imagine a lake that is controlled by a dam. The goal of the experiment is to identify the highest (most important) rocks (actually, SB not only identifies, but also measures the height of these "rocks"). The dam is controlled in such a way that the level of the murky water slowly drops. Obviously, the highest rock first emerges from the water. The most-important-but-one rock turns up next, etc. SB stops when the simulation analysts feel that all the "important" factors are identified; once SB stops, the analysts know that all remaining (unidentified) factors have smaller effects than the effects of the factors that have been identified. This property of SB seems quite important for its use in practice.

Some reflection shows that the aggregated effect of a given subgroup is an *upper limit* for the value of any individual main effect within that subgroup. Examples will be given in the supply-chain study in Section 3.3. If the analysts must terminate SB *prematurely* (e.g., because their computer breaks down or their clients get impatient), then SB still allows identification of the factors with main effects larger than the current upper limit.

SB is extended by Wan et al. (2006a), improving the control over the type-I error rates ("false positives"), using either a two-stage approach or a fully sequential approach. Theoretically, this control does not satisfy the classic statistical requirements concerning a prespecified experimentwise error rate and a prespecified power for the *final* results—after *all* stages have been executed. Nevertheless, the numerical results look very promising.

SB is also extended to the so-called *polytope* method by Ankenman et al. (2006). Their method is more efficient (requiring fewer combinations), but also more complicated (requiring the solution of a LP problem after each additional observation). Moreover they assume main effects only (no interactions). Note that the LP problem arises because this method computes the Ordinary Least Squares (OLS) estimate under the constraint stipulating that all regression coefficients be non-negative (see Assumption 2 above).

3.2 Mathematical Details of Simplest SB

To explain some mathematical details of SB, the following additional notation is used.

$w_{(j);r}$: observed simulation output with the factors 1 through j set to their *high* levels and the remaining factors set to their low levels, in replication r (with $j = 1, \ldots, k$ and $r = 1, \ldots, m$ with $m > 1$);

$\beta_{j'-j}$: *sum* of main effects of factors j' through j; i.e.,

$$\beta_{j'-j} = \sum_{h=j'}^{j} \beta_h. \tag{2}$$

A simple estimate (a complicated estimate is given by Ankenman et al. 2006) of this group effect based on replication r is

$$\widehat{\beta_{j'-j;r}} = \frac{w_{(j);r} - w_{(j'-1);r}}{2}. \tag{3}$$

Section 3.1 mentioned that SB starts with simulating the two most extreme scenarios; i.e., scenario 1 implies that all k factors are at their low levels, so $x_j = -1$; scenario 2 implies that all these factors are high, so $x_j = 1$. If the metamodel in (1) is valid, then

$$E(w_{(0)}) = \beta_0 - \beta_1 - \cdots - \beta_k \tag{4}$$

and

$$E(w_{(k)}) = \beta_0 + \beta_1 + \cdots + \beta_k, \tag{5}$$

so

$$E(w_{(k)}) - E(w_{(0)}) = 2(\beta_1 + \cdots + \beta_k), \tag{6}$$

which shows that the group effect estimator defined in (3) is *unbiased*.

Likewise it follows that the individual main effect β_j is estimated through the analogue of (3):

$$\widehat{\beta_{j;r}} = \frac{w_{(j)r} - w_{(j-1);r}}{2}. \tag{7}$$

The replicates enable the estimation of the mean and the variance for each (aggregated or individual) estimated effect; e.g., (7) gives

$$\overline{\widehat{\beta_j}} = \frac{\sum_{r=1}^{m} \widehat{\beta_{j;r}}}{m} \text{ and } s(\overline{\widehat{\beta_j}}) = \sqrt{\frac{\sum_{r=1}^{m} (\widehat{\beta_{j;r}} - \overline{\widehat{\beta_j}})^2}{m(m-1)}}. \tag{8}$$

This variance estimator allows variance heterogeneity of the simulation outputs, as well as Common Random Numbers (CRN); also see the discussion on variance heterogeneity and CRN in Kleijnen (2008).

To *test* the importance of the estimated (either aggregated or individual) main effects statistically, SB uses a classic t statistic. Different scenarios probably produce observations with different variances, and may use CRN. SB applies a one-sided test because SB assumes that all individual main effects are nonnegative. SB uses a prespecified type-I error rate (e.g., $\alpha = 0.05$) per test; i.e., SB does not adjust for multiple testing. (Response Surface Methodology or RSM is also a sequential procedure that does not control the type-I and type-II error rates over the whole procedure, but is much applied; see Kleijnen 2008 and Myers and Montgomery 2002.) However, Wan et al. (2006a) do use multiple testing procedures in SB.

To *verify* (or validate) the shortlist resulting from SB, the effects of the "unimportant" factors may be tested through the following two scenarios, each simulated m times:

i. Set all factors that SB declared to be unimportant at their low levels, while keeping the important factors fixed (e.g., at their base levels).
ii. Switch all these unimportant factors to their high levels, still keeping the important factors fixed.

Obviously, these two scenarios are not used in SB if verification fixes the important factors at base values (coded as 0) that are not extreme values (coded as either -1 or 1, which are used in SB). The difference between the outputs of

these two scenarios may be tested through a t statistic; this difference is expected not to differ significantly from zero if the factors are actually unimportant.

How SB proceeds sequentially is illustrated in the following case study. A formal computer procedure for the SB steps is given by Wan et al. (2006a).

3.3 Case Study: Ericsson's Supply-Chain Simulation

An example of the application of SB to a simulation with many factors is the following case study. Originally, Persson and Olhager (2002) developed a supply-chain simulation for the Ericsson company in Sweden, and simulated only nine factor combinations. Kleijnen et al. (2006), however, revisit this simulation model and distinguish $k = 92$ factors! (Moreover, they study two other variants of this supply chain with fewer factors, which are not reported in this chapter.)

They replicate each combination $m = 5$ times. The first extreme scenario with all 92 factors at their low levels gives the average output $\overline{w_{(0)}} = 3{,}981{,}627$. The other extreme scenario with all factors at their high levels gives the average output $\overline{w_{(92)}} = 34{,}013{,}832$. So, the estimated group effect of all 92 factors is obtained from (2), (6), and (8), and is $\widehat{\overline{\beta_{1-92}}} = (34{,}013{,}832 - 3{,}983{,}627)/2 = 15{,}016{,}102$. The standard error of this estimated group effect follows from (8), and turns out to be $s(\widehat{\overline{\beta_{1-92}}}) = 42{,}051.18$. So this effect is very significant; and in hindsight, fewer replicates might have been simulated at this early stage; e.g., only $m = 2$ replicates would have shown that one or more factors among the 92 factors must be important.

Next, SB divides the current group of 92 factors into two subgroups. The first subgroup consists of all the 79 "decision" (or "controllable") factors (labeled from 1 through 79); the other subgroup contains all 13 "environmental" factors. Simulation of this scenario must give an expected output between the expected outputs of the preceding extreme scenarios. Comparison of the simulation observations $\overline{w_{(79)}}$ and $\overline{w_{(0)}}$ gives the estimated (sub)group effect $\widehat{\overline{\beta_{1-79}}}$. Similarly, comparison of $\overline{w_{(92)}}$ and $\overline{w_{(79)}}$ gives $\widehat{\overline{\beta_{80-92}}}$. Thus, this step splits the total estimated effect $\widehat{\overline{\beta_{1-92}}}$ into its two additive components. This step decreases the upper limit for any individual effect in the first subgroup and the second subgroup respectively.

Kleijnen et al. (2006) give details on the successive SB steps for this case study. SB does not split a subgroup any further when its estimated aggregated main effect is not significantly positive; e.g., the estimated aggregated main effect of factors 50 through 79 turns out to be a small negative value.

In this case study, it turns out that SB stops after only 19 replicated observations (combinations). The upper limit for the main effect of any remaining individual factor is then reduced to 87,759 (whereas $\widehat{\overline{\beta_{1-92}}} = 15{,}016{,}102$). SB produces a shortlist with only 10 factors; its most important factor is factor 92. Section 3.1 mentioned that the SB efficiency improves when factors are labeled from least important to most important; indeed, factor 92 turns out to be the most important factor and

no factor labelled smaller than 43 is declared to be important. The most important individual factor (namely, factor 92) has already been identified and estimated after only six replicated observations.

3.4 SB with Two-Factor Interactions

This section summarizes SB for situations in which Assumption 1(a) is replaced by Assumptions 1(b) and 3, which are detailed now.

Assumption 1(b) a valid metamodel is a first-order polynomial augmented with two-factor interactions $\beta_{j';j}$ ($j' < j$; $j' = 1, \ldots, k-1$; $j = 2, \ldots, k$) and noise:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{1;2} x_1 x_2 + \cdots + \beta_{k-1;k} x_{k-1} x_k + e. \tag{9}$$

The signs of these interactions are irrelevant (see below).

Assumption 3 if a factor has no important main effect, then this factor does not interact with any other factor.

Assumption 3 is called the *strong heredity* assumption; see Wu and Hamada (2000) and also Saltelli et al. (2005). Strong heredity is related to *functional marginality*, which is discussed by Tsai et al. (2007).

Let $w_{-(j)}$ denote the *mirror* observation of $w_{(j)}$; i.e., $w_{-(j)}$ is the simulation output with the factors 1 through j set to their *low* levels and the remaining factors set to their high levels. For example, the analogues of (4) and (5) are for $j = 48$:

$$\begin{aligned} E(w_{-(49)}) = \beta_0 &+ (-\beta_1 - \cdots - \beta_{49}) + (\beta_{50} + \cdots + \beta_{92}) \\ &+ (\beta_{1;2} + \cdots + \beta_{48;49}) + (-\beta_{1;50} - \cdots - \beta_{49;92}) \\ &+ (\beta_{50;51} + \cdots + \beta_{91;92}) \end{aligned}$$

and

$$\begin{aligned} E(w_{(49)}) = \beta_0 &+ (\beta_1 + \cdots + \beta_{49}) + (-\beta_{50} - \cdots - \beta_{92}) \\ &+ (\beta_{1;2} + \cdots + \beta_{48;49}) + (-\beta_{1;50} - \cdots - \beta_{49;92}) \\ &+ (\beta_{50;51} + \cdots + \beta_{91;92}), \end{aligned}$$

so subtracting these two equations *cancels all interactions*. The analogue of (3) gives the unbiased group estimator

$$\widehat{\beta_{j'-j;r}} = \frac{(w_{(j);r} - w_{-(j);r}) - (w_{(j'-1);r} - w_{-(j'-1);r})}{4}. \tag{10}$$

The analogue of (7) gives the unbiased individual estimator

$$\widehat{\beta_{j;r}} = \frac{(w_{(j);r} - w_{-(j);r}) - (w_{(j-1);r} - w_{-(j-1);r})}{4}. \tag{11}$$

In other words, SB enables the estimation of first-order effects unbiased by two-factor interactions provided SB simulates the mirror combinations besides the original combinations. Hence, the number of simulated combinations *doubles*. Wan et al. (2006b) point out that—in the case of mirror observations—fewer replications per combination may be needed. They further state that the SB efficiency may increase when applying CRN separately to all positive levels and negative levels respectively.

SB with mirror scenarios may still give misleading results if (say) two factors have unimportant main effects but their interaction is important. Therefore SB assumes *strong heredity* (Assumption 3). If the analysts suspect that this assumption is violated for a specific factor, then they should investigate that factor after the screening phase.

SB with mirror observations does not enable estimation of *individual* interactions, but it does show whether interactions are important—as follows. Estimate the main effects from the original scenarios—ignoring the mirror scenarios. If the analyses of the mirror observations and of the original observations give the same conclusions, then interactions are unimportant. This happened, e.g., in the ecological simulation reported in Bettonvil (1990) and Bettonvil and Kleijnen (1996). In that study, the factor values change relatively little (larger changes give unrealistic simulation outputs), so a first-order polynomial is adequate. In the supply-chain study, however, interactions turn out to be important; see Kleijnen et al. (2006). (In a follow-up experiment with the factors declared to be important in SB, the sizes of the individual interactions are estimated from a resolution-V design, which by definition enables the unbiased estimation of all the individual two-factor interactions; details on resolution-V designs are given by Kleijnen 2008.) Note that the mirror observations and the original observations may give different SB paths through the list of individual factors.

4 Conclusions and Future Research

This chapter may be summarized as follows. There are different screening designs, including resolution-III, supersaturated, and group-screening designs. This chapter, however, focused on SB, and stated the various assumptions of SB. These assumptions may not be too restrictive in practice, as the Ericsson case-study illustrated. If its assumptions are satisfied, then SB is a most efficient and effective screening method.

There is a need for more research:

- It is a challenge to derive the number of replicates that control the *overall* probability of correctly classifying the individual factors as important or unimportant. So far, SB applies a statistical test to each subgroup individually. (Furthermore, SB may terminate "prematurely," and yet estimate the most important factors—instead of classifying all factors with effects that exceed a prespecified threshold.)

- It might be that the simulation of *mirror* factor combinations can be stopped as soon as it seems that no interactions are important.
- After SB stops, the resulting shortlist of important factors should be *validated.* (A procedure was proposed above.)
- *Software* needs to be developed that implements sequential screening of simulation experiments. This software should generate an input file, once a particular design type (e.g., SB) has been chosen. Such a file can then be executed sequentially and efficiently in batch mode; i.e., no human intervention is required while the computer executes the sequential design. Furthermore, good computer programming avoids fixing the inputs at specific numerical values within the code; instead, the computer reads input values so that the program can be run for many combinations of these values. (Of course, the computer should check whether these values are admissible; i.e., do these combinations fall within the experimental domain?) Such a practice can automatically provide a long list of potential factors.
- A contest may be organized that challenges the experts in the different screening methods to estimate the most important factors in a set of simulation models. Such "testbeds" are popular in Mathematical Programming. Note that nobody is expert in *all* screening methods.
- *Multivariate* output may consist of univariate outputs that require different SB paths. This problem has not yet been touched in the literature!

Acknowledgments I thank Bert Bettonvil and Wim van Beers (both at Tilburg University) for their comments on an earlier version.

References

Allen, T. T., and M. Bernshteyn. 2003. Supersaturated designs that maximize the probability of finding the active factors. *Technometrics* 45(1):1–8.

Andres, T. H. 1997. Sampling methods and sensitivity analysis for large parameter sets. *Journal of Statistical Computation and Simulation* 57(1/4):77–110.

Andres, T. H., and W. C. Hajas. 1993. Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model. In *Proceedings of the Joint International Conference on Mathematical Methods and Supercomputing in Nuclear Applications*, Karlsruhe, Germany, April 19–23, 2:328–337.

Ankenman, B. E., R. C. H. Cheng, and S. M. Lewis. 2006. A polytope method for estimating factor main effects efficiently. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, pp. 369–375. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

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

Bettonvil, B., and J. P. C. Kleijnen. 1996. Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research* 96:180–194.

Campolongo, F., J. Cariboni, and A. Saltelli. 2007. An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software* 22(10):1509–1518.

Campolongo, F., J. P. C. Kleijnen, and T. Andres. 2000. Screening methods. In *Sensitivity Analysis*, eds. A. Saltelli, K. Chan, and E. M. Scott, pp. 65–89. Chichester, England: Wiley.

Cheng, R. C. H. 1997. Searching for important factors: Sequential bifurcation under uncertainty. In *Proceedings of the 1997 Winter Simulation Conference*, eds. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, pp. 275–280. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Cheng, R. C. H., and W. Holland. 1999. Stochastic sequential bifurcation: Practical issues. In *Proceedings of UK Simulation Society Conference*, eds. D. Al-Dabass and R. C. H. Cheng, pp. 74–80. Nottingham: UKSIM.

Chipman, H. 2006. Prior distributions for Bayesian analysis of screening experiments. In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. A. Dean and S. Lewis, pp. 235–267. New York: Springer-Verlag.

De Vos, C., H. W. Saatkamp, M. Nielen, and R. B. M. Huirne. 2006. Sensitivity analysis to evaluate the impact of uncertain factors in a scenario tree model for classical swine fever introduction. *Risk Analysis* 26(5):1311–1322.

Elster, C., and A. Neumaier. 1995. Screening by conference designs. *Biometrika* 82(3):589–602.

Georgiou, S. D. 2007. New two-variable full orthogonal designs and related experiments with linear regression models. *Statistics & Probability Letters* 77(1):25–31.

Gilmour, S. G. 2006. Factor screening via supersaturated designs. In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. A. Dean and S. Lewis, pp. 169–190. New York: Springer-Verlag.

Kleijnen, J. P. C. 2008. *Design and Analysis of Simulation Experiments.* New York: Springer Science + Business Media.

Kleijnen, J. P. C., B. Bettonvil, and F. Persson. 2006. Screening for the important factors in large discrete-event simulation: sequential bifurcation and its applications. In: *Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. A. Dean and S. Lewis, pp. 287–307. New York: Springer-Verlag.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. State-of-the-art review: A user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing* 17(3):263–289.

Kleijnen, J. P. C., and W. C. M. van Beers. 2004. Application-driven sequential designs for simulation experiments: Kriging metamodeling. *Journal of the Operational Research Society* 55(9):876–883.

Kleijnen, J. P. C., G. van Ham, and J. Rotmans. 1992. Techniques for sensitivity analysis of simulation models: A case study of the CO_2 greenhouse effect. *Simulation* 58(6):410–417.

Lewis, S. M., and A. M. Dean. 2001. Detection of interactions in experiments on large numbers of factors (including discussion). *Journal Royal Statistical Society B* 63:633–672.

Li, L., and W. Li. 2005. Tabu search and perturbation methods in the construction of supersaturated designs. *American Journal of Mathematical and Management Sciences* 25(1&2):189–205.

Lim, E., and P. W. Glynn. 2006. Simulation-based response surface computation in the presence of monotonicity. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, pp. 264–271. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Miller, G. A. 1956. The magical number seven plus or minus two: Some limits on our capacity for processing information. *The Psychological Review* 63:81–97.

Morris, M. D. 2006. An overview of group factor screening. In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. A. Dean and S. Lewis, pp. 191–206. New York: Springer-Verlag.

Myers, R. H., and D. C. Montgomery. 2002. *Response Surface methodology: Process and Product Optimization Using Designed Experiments,* 2nd edition. New York: Wiley.

Park, S., J. W. Fowler, G. T. Mackulak, J. B. Keats, and W. M. Carlyle. 2002. D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. *Operations Research* 50(6):981–990.

Persson, J. F., and J. Olhager. 2002. Performance simulation of supply chain designs. *International Journal of Production Economics* 77:231–245.

Saltelli, A., M. Ratto, S. Tarantola, and F. Campolongo. 2005. Sensitivity analysis of chemical models. *Chemical Reviews* 105(7):2811–2827.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto. 2004. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models.*Chichester, England: Wiley.

Sanchez, S. M., F. Moeeni, and P. J. Sanchez. 2006. So many factors, so little time… Simulation experiments in the frequency domain. *International Journal of Production Economics* 103(1):149–165.

Schonlau, M., and W. J. Welch. 2006. Screening the input factors to a computer code via Analysis of Variance and visualization. In *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. A. Dean and S. Lewis, pp. 308–327. New York: Springer-Verlag.

Schruben, L. W., and V. J. Cogliano. 1987. An experimental procedure for simulation response surface model identification. *Communications of the ACM* 30(8):716–730.

Shen, H., and H. Wan. 2006. A hybrid method for simulation factor screening. In *Proceedings of the 2006 Winter Simulation Conference*, eds. L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, pp. 382–389. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Simpson, T. W., A. J. Booker, D. Ghosh, A. A. Giunta, P. N. Koch, and R.-J. Yang. 2004. Approximation methods in multidisciplinary analysis and optimization: a panel discussion. *Structural and Multidisciplinary Optimization* 27(5):302–313.

Tsai, P.-W., S. G. Gilmour, and R. Mead. 2007. Three-level main-effects designs exploiting prior information about model uncertainty. *Journal of Statistical Planning and Inference* 137(2):619–627.

Tu, J., and D. R. Jones. 2003. Variable screening in metamodel design by cross-validated Moving Least Squares method. In *Proceedings of the 44th AIAA / ASME / ASCE / AHS / ASC Structures, Structural Dynamics, and Materials Conference*, AIAA-2003-1669, Norfolk, Virginia, April 7–10, 2003.

Van Beers, W. C. M., and J. P. C. Kleijnen. 2008. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research* 186(3):1099–1113.

Wan, H., B. E. Ankenman, and B. L. Nelson. 2006a. Controlled sequential bifurcation: A new factor-screening method for discrete-event simulation. *Operations Research* 54(4):743–755.

Wan, H., B. E. Ankenman, and B. L. Nelson. 2006b. Simulation factor screening with controlled sequential bifurcation in the presence of interactions. Working Paper, Purdue University, West Lafayette, Indiana.

Wu, C. F. J., and M. Hamada. 2000. *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley.

Xu, J., F. Yang, and H. Wan. 2007. Controlled sequential bifurcation for software reliability study. In *Proceedings of the 2007 Winter Simulation Conference*, eds. S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, pp. 281–288. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Yamada, S., M. Matsui, T. Matsui, D. K. J. Lin, and T. Takahashi. 2006. A general construction method for mixed-level supersaturated design *Computational Statistics & Data Analysis* 50(1):254–265.

Yu, H.-F. 2007. Designing a screening experiment with a reciprocal Weibull degradation rate. *Computers & Industrial Engineering* 52(2):175–191.

Zeigler, B. P., H. Praehofer, and T. G. Kim. 2000. *Theory of Modeling and Simulation*, 2nd edition. San Diego: Academic Press.

Zhang, Q.-Z., R.-C. Zhang, and M.-Q. Liu. 2007. A method for screening active effects in supersaturated designs. *Journal of Statistical Planning and Inference* 137(6):2068–2079.