Evaluation of a New Screening Procedure

Steve Webb
(unaffiliated) 809 Grayling Bay, Costa Mesa, CA 92626
SteveWebb1@Compuserve.com

Abstract

The performance of a sequential screening procedure for finding a few active factors from a larger number is investigated. Called the Winnow, it was previously introduced and studied using dummy simulation models. Here we apply it to a legacy simulation using eight different baselines with from 25 to 78 inputs varied. Large fractional factorials are used to find "true" values of the effects. The largest were of the order of one standard deviation for these baselines. The sequential procedure is evaluated by finding how many real effects are identified and the accuracy of the resulting estimates. It is shown that an effect as large as 1 sigma will very likely be found and one as large as 0.5 sigma will usually be found. The winnow is found to be generally better than the alternatives of fixed supersaturated designs or a version of group screening.

Key Words: Screening designs, sensitivity analysis, factorial designs

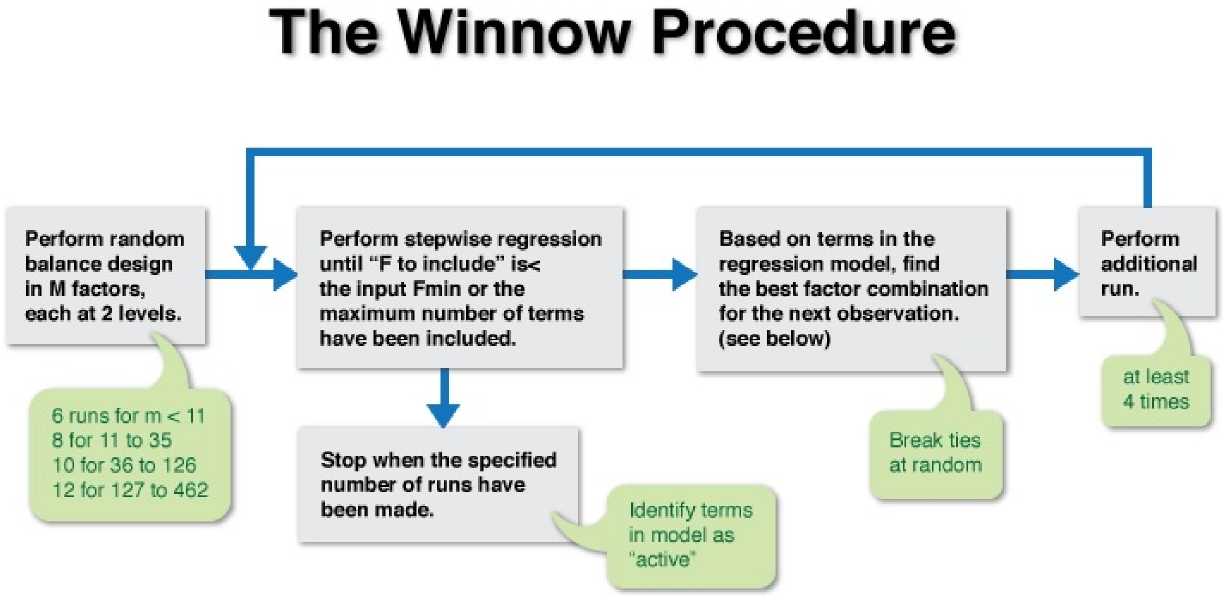
1. The Winnow Procedure

The problem that motivated the work reported here and the development of the new procedure was that of determining which inputs to a Monte Carlo simulation model are active in influencing the responses of interest. Such models are characterized by having a large number of inputs, the inputs are under the control of the computer running the model, and the results are stochastic, in that different outcomes result if different random number seeds are used (unlike some models for computer experiments). An important assumption is that only a few of the many factors have appreciable influences on the response, also referred to as factor sparsity. This problem is called factor screening (Dean & Lewis 2006) or sensitivity analysis. The procedure discussed is applicable to other experiments that share these features.

A practical approach to factor screening is to perform experiments using the model. Three standard approaches to experimentation are

- Orthogonal fractional factorials or Plackett-Burman (1946) designs
- Supersaturated designs
- Group screening

If the model has a large number of inputs (factors) but only a few are active, we would like for the experiment to have fewer runs than factors. Orthogonal factorials require more runs than the total number of factors, so will not be considered further. Supersaturated designs are available in the literature for some, but not all, combinations of number of runs and number of factors. Group screening is usually discussed in the literature in terms of finding a single factor. For finding possibly more an implementation of group screening will have a number of runs that depends on the relative positions of the active factors, so is a random variable.


The Winnow Procedure
Perform random balance design in M factors, each at 2 levels.
6 runs for m < 11
8 for 11 to 35
10 for 36 to 126
12 for 127 to 462
Perform stepwise regression until "F to include" is< the input Fmin or the maximum number of terms have been included.
Stop when the specified number of runs have been made.
Identify terms in model as "active"
Based on terms in the regression model, find the best factor combination for the next observation. (see below)
Break ties at random
Perform additional run.
at least 4 times

Best next run: for the terms in the regression model, find that factor combination that yields largest variances of a prediction. For factors in the regression model, select level 0 or 1 that is better when considered in conjunction with the combination selected for terms in model.

Figure 1: Schematic description of the Winnow procedure

The new Winnow procedure (Webb 2011) starts with a smallest possible randomly chosen balanced design in the m factors. Let 0 represent a nominal level for each factor and 1 represent an excursion level. If m is 10 or fewer this design will be a random selection of the 10 possible columns starting with 0 and containing two more 0's and three 1's, for a total of 6 runs N. Values for N for other m are in Figure 1. Observations are obtained for the runs in the initial design.

The second step is to perform stepwise regression on the results, stopping when the "F to include" is less than an input limit Fmin, or a maximum number of terms is included. Thus a small model is formed.

The third step is to augment the experiment with a single new run. For terms in the small model, levels are set at the point that maximizes the variance of a predicted value. For other factors the level is selected that is better when run with the chosen values for the terms in the small model. The new point is run and the procedure loops back to step 2, stepwise regression.

If factors appear in several successive small models, the procedure includes their interaction among those terms available to the stepwise regression. They are included only if both parents are already placed in the model and they are the next stepwise selection.

Currently the Winnow procedure terminates when a prespecified number of augmentation iterations have been made. Other stopping rules might look at the stability of the models selected in successive steps or the sizes of the regression coefficients.

2. Evaluation Methodology

Previous work (Webb 2011) suggests that Winnow is effective when looking for effects that are of the order of magnitude of the standard deviation, but that work dealt with very simple dummy models. The work reported here is an empirical investigation using a large legacy military simulation that has a history of use by many agencies and contractors. Called Osprey, it is used to study the overall performance of an antisatellite system in shooting down hostile military satellites of various types (see Appendix).

The first step is to establish baselines defined by the nominal values of all inputs, the selection of which factors will be varied, and the excursion level for each of these. Eight baselines were used. Next large fractional factorials were run to determine “true” values for effects and interactions. For this purpose designs from Xu (2009) of resolution IV or V were used, with 1024 runs for four of the eight baselines, 512 for two, and 256 for two. Common random numbers were used to reduce the variability of the estimates of effects. This was repeated four times using different random seeds to initialize the random number generators used in the target simulation.

Separate experiments were run to obtain estimates of the standard deviation of the response of interest. These used independent random number strings rather than common random numbers. These used factorial designs in a few of the largest effects found in the baselines to check for homogeneity of error.

The effects from the large factorials were categorized into four classes. Class A includes effects larger than one standard deviation; class B includes those less than one standard deviation but larger than one-half the standard deviation; class C includes effects between σ/3 and σ/2. The fourth class is everything else, assumed to be not of interest or not active. Note that a given effect may be in different classes for the four target simulation random seeds.

The evaluation of the Winnow will be done by comparing results with those obtained from supersaturated designs and from group screening. Each of these three procedures contains a random mechanism for assigning factors to columns of the designs used, and the Winnow also uses random numbers to break ties when performing augmentation steps. For each procedure four random seeds were used, and crossed with the four target simulation seeds for a total of 16 cases.

3. Implementation Details for the Procedures

3.1 Winnow Procedure

The Winnow procedure is implemented as a C program running under Linux. It is called to set up data sets for each run of the target simulation. Its control parameters are

- File names
- Maximum number of terms in the model (current limit 12)
- Minimum F to include in the stepwise regression (2.0 works well)
- Number of runs (current limit 100)
- Random seeds
- List of variables to be considered (current limit 128).

The procedure is also called after the last simulation run to provide a final assessment of the terms found to be active and their regression coefficients.

3.2 Supersaturated Designs

The supersaturated designs used were obtained from Booth & Cox (1962) and Lin (1995). Design matrices from the references were entered with the assignments of factors to columns made at random. Observations from the experiment were analyzed using forward stepwise regression. The regression stopped when the F to include was less than an input level Fmin or a maximum number of terms were included. The values used were 2.0 and 12, the same as for Winnow. The designs used had 12, 18, or 24 runs. Four different random assignments of factors to columns were used.

3.3 Group Screening

The variant of group screening evaluated was adapted from Kleijnen (1987). The factors are randomly assigned to one of two groups of as nearly the same size as possible. Four runs are made: all factors at their 0 level, those in group A at 0 and group B at 1, those in group A at 1 and group B at 0, and all factors at level 1. Estimates are obtained of the two group factors. Any group whose estimated group effect exceeds $\sigma/3$ is subdivided, with factors not in the group held at their 0 level. Each string of subdivisions terminates when a group contains a single factor or neither of the groups has estimated effect greater than $\sigma/3$. Note that the number of runs is a random variable that will have different values for different assignments of the factors to groups and subgroups. As with Winnow and supersaturated designs, four different randomizations were used.

3.4 Adapting the Target Simulation

The modifications required to form an interface between the screening procedures and the target simulation are minimal. The values of the inputs are read from a special text file, one line per input, which may be a vector quantity. The value of the output response is written to another special text file. All communication is done through these files, so that it is not necessary for the programs to call one another or even be written in the same language (Osprey is in Fortran, other programs in C). Evaluations using Winnow and supersaturated designs were done in a fully automatic mode using shell scripts to alternately call the screening procedure and the target simulation. Group screening was evaluated in a manual mode because the next two experiments to be run depend on the results of the previous runs.

4. Synopsis of Experiments Run

The 8 baseline data sets differed in the nominal values of parameters and the number of factors varied.. The number varied ranged from 25 to 78 in the experiments reported here. They were also intentionally constructed so that different numbers of factors would be included in the three classes A, B, and C. Table 1 gives some summary data for what was obtained. The third through fifth columns of the table give the number of effects found to be in the three classes. Listed are the total number found from the four data sets that differ only in the random seeds used to drive Osprey. For example the entries for Osp1 represent one factor, call it P, that appears as a class A effect in one of the four data sets and class B effect in the other three, and a second factor Q that is class B in one data set and class C in a second.

Table 1: Summary of Experiments Run

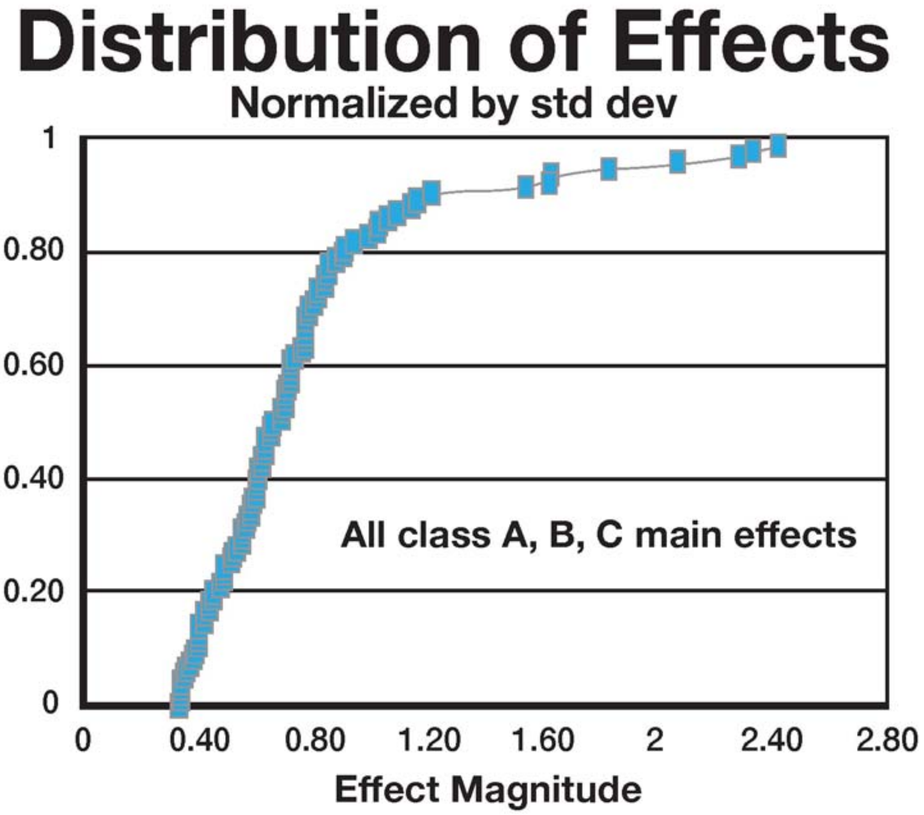
Baseline Number	Factors Varied	# effects by class			# runs for Winnow & Supersat			# runs with Group screen
		A	B	C	12	18	24	
Osp1	25	1	4	1	X	X		16.625
Osp2	28	5	13	6	X	X	X	27.5
Osp3	28	1	8	2		X	X	28.0
Osp4	33	4	3	1	X	X	X	
Osp5	45	0	1	3		X		
Osp6	45	1	14	7		X		
Osp7	54	1	2	4		X		
Osp8*	78	2	10	0		X	X	35.625
	Totals	15	55	24				

* Supersaturated not run for Osp8

The number of effects found to be in the three classes varies from 4 to 24. A plot of the distribution of the 94 effects found is shown as Figure 2.

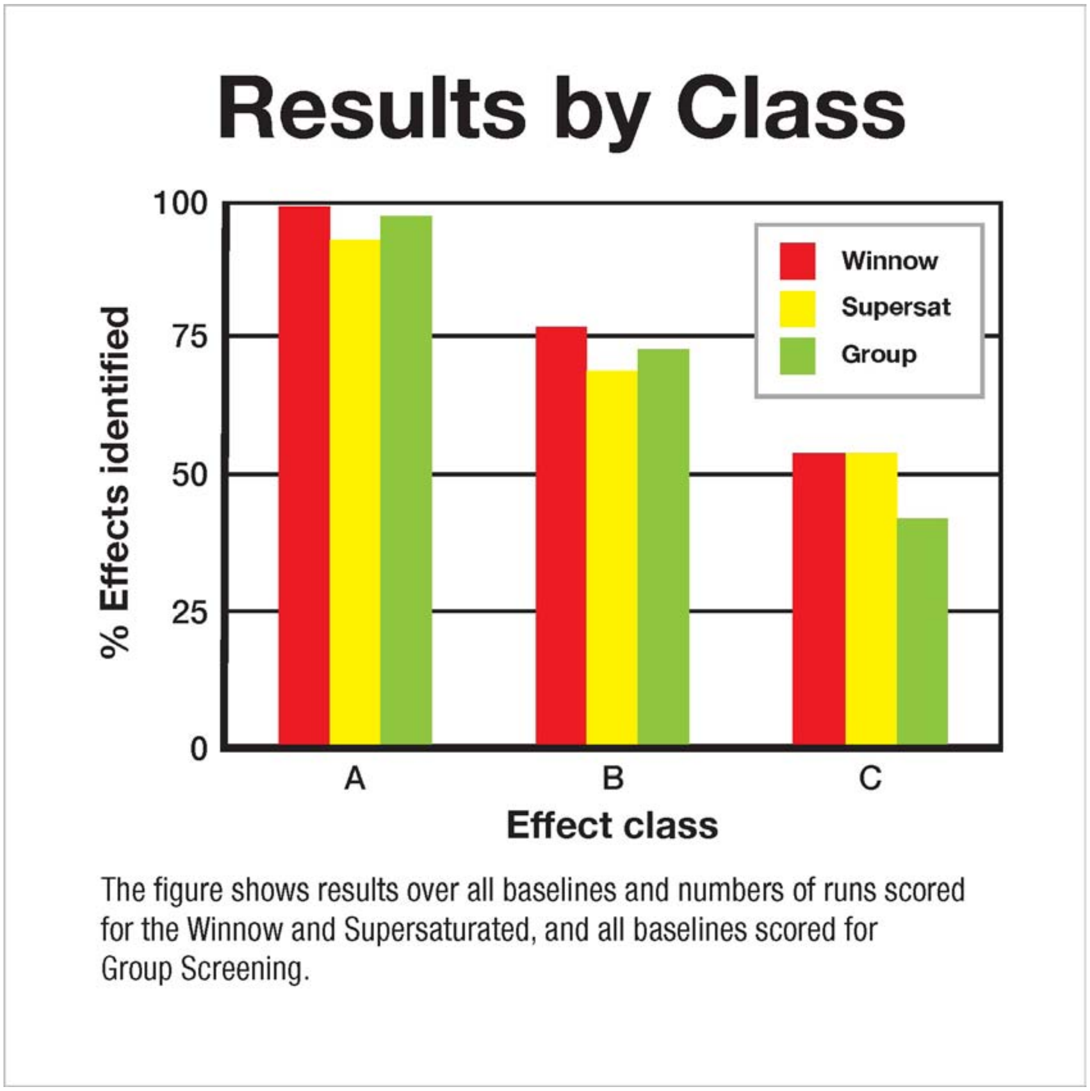
Columns 6 through 8 of Table 1 indicate for which values of N supersaturated designs were executed. Winnow results were extracted for the same values of N. So for example the results at 12 and 18 runs using Osp2 were found along the way to a final 24 runs. No supersaturated design for 78 factors was found, so no supersaturated results were obtained for Osp8.

The final column gives the average number of runs obtained using the group screening procedure. That procedure was done one step at a time and was rather tedious to accomplish. The cases that were run contained more runs than the other procedures and did not perform particularly well, so only four of the 8 baselines were used for this procedure.



The cumulative distribution of the magnitudes of the effects found to be greater than 1/3 the standard deviation

Figure 2: Cumulative Distribution of 94 Effects Found



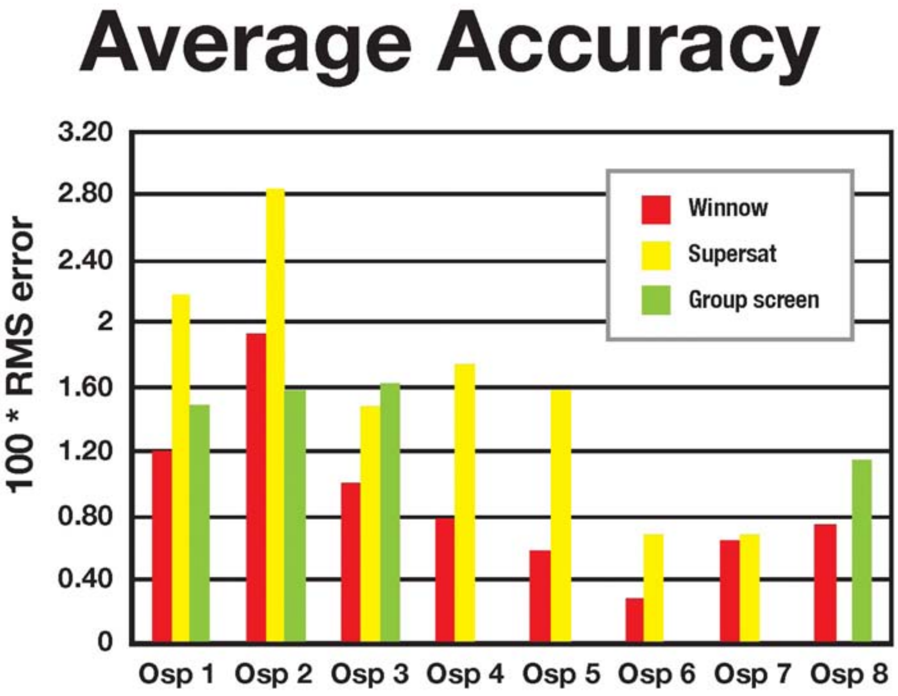
The figure shows results over all baselines and numbers of runs scored for the Winnow and Supersaturated, and all baselines scored for Group Screening.

Figure 3: Percentage of active effects identified by effect size class

5. Evaluation Results

The primary evaluation criterion for a screening design is how well it identifies active factors. Figure 3 summarizes how well the three procedures do. Results over all cases run are pooled for this figure. It is seen that performance depends on the size of the effect, with class A being effects larger than σ, B between σ/2 and σ, and C between σ/3 and σ/2. The procedures are roughly comparable, with Winnow slightly better.

Given that a procedure identifies active factors, one would like for the estimated effects of these factors to be accurate. Since the methodology used provides very accurate estimates of the true values of effects based on large factorials, accuracies of the procedures can be obtained by comparison with these "true" values. The root mean squared error was used as the measure of accuracy. Figure 4 summarizes accuracy for the eight baselines. Note that short lines are good, and that Winnow has the shortest lines except for the one case noted in the caption.

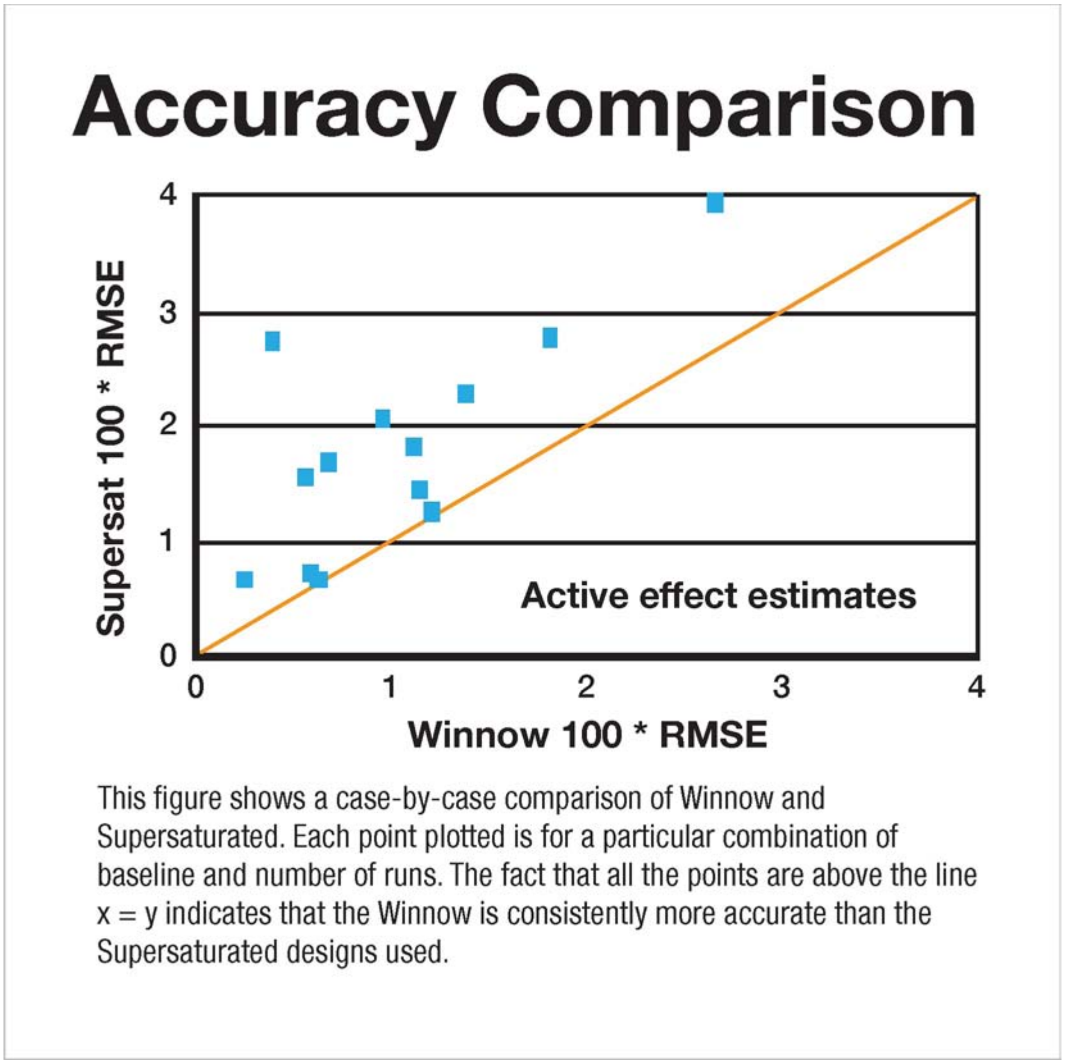


The figure shows that the accuracy of the Winnow is consistently better than Supersaturated or Group Screening. The only exception is for the baseline Osp 2, where Group Screening is more accurate than Winnow. However, the Winnow is based on 12 and 18 observations, while Group Screening averages over 27 observations.

Figure 4: Accuracy of estimates of active effects for 8 baselines

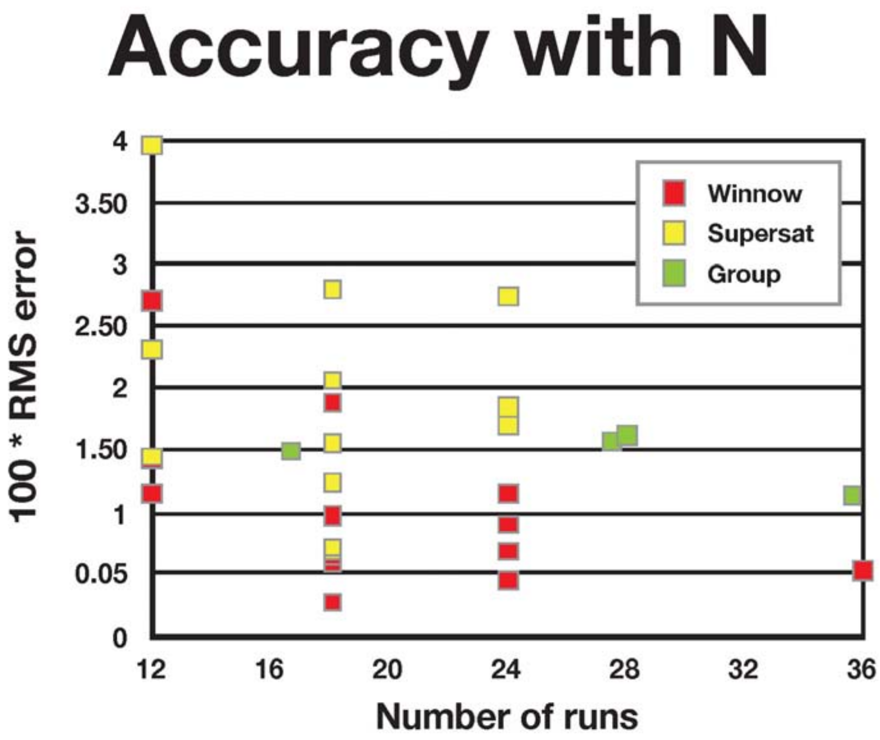
Another comparison of accuracy for Winnow and Supersaturated was made on a case-by-case basis. That is, a direct comparison was made for each baseline and number of runs; these are represented by the X in columns 6 through 8 of Table 1 for the first seven baselines. There are 13 such X's, so 13 direct comparisons can be made. The comparison is shown in Figure 5. For all 13 cases the accuracy of Winnow was better than Supersaturated.

A final look at accuracy in Figure 6 considers the number of runs. One data point has been added that is not given in Table 1 or included in the previous analyses – that for Winnow with 36 runs for baseline Osp8. This was added to provide a comparison with group screening for this baseline. In general the accuracy of Winnow improves with the number of runs because larger N results build upon those for smaller N. Accuracy of Supersaturated results is quite variable. Results for group screening are more or less constant because each effect estimate is obtained from four observations.



This figure shows a case-by-case comparison of Winnow and Supersaturated. Each point plotted is for a particular combination of baseline and number of runs. The fact that all the points are above the line x = y indicates that the Winnow is consistently more accurate than the Supersaturated designs used.

Figure 5: Direct comparison of Winnow and Supersaturated accuracies



For Winnow the accuracy of the estimated effects increases as the number of samples increases. For Supersaturated the accuracy varies a lot with different designs. For Group Screening the accuracy is more or less constant.

Figure 6: Shown is the change in accuracy as the number of runs increases

6. Conclusions

The study described evaluated the new Winnow procedure for factor screening. The evaluation was limited to one simulation model run at 8 different baselines. The simulation used was a large legacy Monte Carlo military simulation of the performance of a generic antisatellite missile system. The evaluation was done by comparing the new procedure with use of supersaturated designs and group screening. Within each of the baselines common random numbers were used in the simulation to decrease the variance of the effect estimates.

The effects to be found with the screening procedures were grouped into three classes based on true values obtained from large fractional factorials. Class A contained effects larger than σ up to about 2σ. The majority, falling between σ/2 and σ, were called class B effects. Those between σ/3 and σ/2 were assigned to class C. It was found that class A effects were found with high probability, those in class B were found about 75% of the time, and those in class C about 50% of the time. This performance is a function of the model used and in particular its tight assignment of random number strings to entities simulated. Nevertheless it represents a benchmark for screening procedures in general.

The Winnow outperformed the other procedures used in the study: it identified active factors at slightly higher rates and the accuracy of the effect estimates is consistently better. If the experiments cannot be run sequentially, or if the number of runs does not allow as many as four iterations of the Winnow augmentation procedure, then supersaturated designs are a good alternative. The version of group screening used required more runs than the other procedures, so was not competitive for the baselines used and the numbers of effects judged to be active.

The Winnow procedure was found to be flexible and easy to use. Its performance scales well with the total number of factors. Any number of factors and any number of runs may be used. It does not require input of a design matrix as does the use of supersaturated designs. Interaction terms may be considered and allowed in the model, although there were virtually no interaction terms in classes A to C for the baselines evaluated. The accuracy improves as the number of runs increases. Finally, the user can examine the results sequentially and decide to stop or continue, although this was not done in this evaluation.

Appendix – Description of the Osprey Model

China recently tested an antisatellite system by destroying an obsolete Chinese weather satellite. This is a specific example of the use of ground-based missiles against satellites. Such systems have been studied and prototype hardware developed over the last several decades of the last century. A notional concept of operation for such a system follows.

A set of hostile military satellites has been identified and an order given to destroy them. They are of several different types, such as photoreconnaisence at high or low resolution, electronic intelligence, radar trackers, navigation aids, etc. Each type has constellation structure and orbital characteristics suited to its mission. The satellites are tracked extensively by ground-based space surveillance radars and optical sensors to obtain precise orbital ephemerides. Ground based missiles are available for this mission that are used to boost a small kill vehicle into an intercept trajectory. The kill vehicle is equipped with an electro-optical sensor that acquires the target. The kill vehicle maneuvers in such a way as to allow the satellite to impact it, destroying the satellite and kill vehicle. Engagement plans are generated that specify the launch time, direction, and trajectory shape for the missile. Constraints include lighting of the target and favorable locations of the sun so that the sensor will acquire the target reliably. Alternate engagement plans are evaluated and the best selected. A timeline is then established for preparing the missile for launch, performing checkout, loading mission data, and launching. The missile is monitored for any indication of failure of function or to acquire the target during missile flight. During and after the scheduled intercept time, the same ground-based sensors are used to assess the results of the intercept attempt. On this basis the satellite is declared to be destroyed or is placed back on the active target list to be re-engaged. To be militarily useful, the ASAT system must be capable of destroying the targets reliably and quickly without unreasonable use of resources.

Such a system is modeled by the Osprey simulation program, originally developed by Teledyne Brown Engineering. This simulation was used extensively by several contractors and military organizations from 1977 to about 1995 to perform feasibility and trade studies of various design features and options of several potential antisatellite systems. For the current study, the simulation was tailored to represent a notional unclassified antisatellite system and scenario The target set was 15 military satellites in low earth orbits. The antisatellite missiles are located at a base with limited launch pads and checkout and support equipment. One hundred replications are made at any combination of values of the input variables, and the logarithm of the time in hours required to shoot down all the targets used as the response of interest. Osprey contains over 11,000 lines of Fortran code. It is structured as an event-oriented model with 15 event types such as launch, failure, or repair. Separate random number strings are used for each entity (eg satellite, missile, operations center, launch equipment). A few of the key factors that are active in the studies described are in Table 2.

Table 2: Description of Typical Variables Used in the Studies Reported

1.acqrng	The maximum range for target acquisition by the kill vehicle
7.azlim	Launch azimuth limits from the missile base
8.bmtbf	Mean time to failure for the base electronics systems
10.bvbo	Missile burnout velocity
13.check	Time for checkout and launch preparations for a missile
20.dmoc	Time required for decision at mission operations center
28.eanglim	Minimum viewing angle between the target and edge of earth
32.errvel1	Error in burnout velocity for the missile in the direction of flight
71.pp1	Divert velocity available on the kill vehicle for endgame
75.pp5	Field of view of the kill vehicle sensor
76.prb	Probability of success of the missile and kill vehicle
84.sanglim	Minimum viewing angle between target and the sun
100.vclmax	Maximum closing velocity at time of intercept

Acknowledgment

The author thanks Richard Escarcega, principal of Pacific Illustrations, Irvine, CA for his expert assistance in preparing the figures and the layout of the poster.

References

Booth, K. H. V. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics* 4:489-495.

Dean, A. and Lewis, S. (2006). *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics.* New York: Springer.

Kleijnen, J. (1987). *Statistical Tools for Simulation Practitioners.* Marcel Dekker.

Lin, Dennis K. J. (1995). Generating systematic supersaturated designs. *Technometrics* 37: 213-225.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika* 33:303-325.

Webb, Steve (2011). Screening experiments with many factors, *Communications in Statistics - Theory and Methods*, 40:10, 1879-1892.

Xu, Hongquan (2009). Algorithmic construction of efficient fractional factorial designs with large run sizes. *Technometrics* 51:262-277.