Ensembles of Adaptive One-Factor-at-a-time Experiments: Methods, Evaluation, and Theory

by
Nandan Sudarsanam

B.E., Mechanical Engineering, SASTRA University
M.S., Industrial Engineering and Management, Oklahoma State University

Submitted to the Engineering Systems Division
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Engineering Systems
at the
Massachusetts Institute of Technology

August 2008

Author.....
Engineering Systems Division, July 25th, 2008

Certified by.....
Daniel D. Frey
Robert N. Noyce Career Development Professor
Associate Professor of Mechanical Engineering and Engineering Systems
Thesis Supervisor

Certified by.....
Richard Larson
R Mitsui Professor of Engineering Systems and Civil and Environmental Engineering
Director, Center for Engineering Systems Fundamentals

Certified by.....
Christopher L. Magee
Professor of the Practice of Mechanical Engineering and Engineering Systems
Co-Lead, Engineering Design and Advanced Manufacturing (EDAM) Focus Area, MIT Portugal Program

Certified by.....
Roy Welsch
Professor of Statistics and Management Science and Engineering Systems

Accepted by.....
Nancy Leveson
Professor of Engineering Systems and Aeronautics & Astronautics
Chair, ESD Education Committee

Ensembles of Adaptive One-Factor-at-a-time Experiments: Methods, Evaluation, and Theory

by
Nandan Sudarsanam

Submitted to the Engineering Systems Division
on July 25th, 2008, in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Engineering Systems

This thesis recommends an experimentation methodology which can be used to improve systems, processes and products. The proposed technique borrows insights from statistical prediction practices referred to as *Ensemble Methods*, to extend *Adaptive One-Factor-at-a-Time* (*a*OFAT) experimentation. The algorithm is developed for an input space where each variable assumes two or more discrete levels.

Ensemble methods are common data mining procedures in which a set of similar predictors is created and the overall prediction is achieved through the aggregation of these units. In a methodologically similar way this study proposes to plan and execute multiple *a*OFAT experiments on the same system with minor differences in experimental setup, such as starting points, or order of variable changes. Experimental conclusions are arrived at by aggregating the multiple, individual *a*OFATs. Different strategies for selecting starting points, order of variable changes, and aggregation techniques are explored. The proposed algorithm is compared to the performance of a traditional form of experimentation, namely a single orthogonal array (full and fractional factorial designs), which is equally resource intensive. Comparisons between the two experimental algorithms are conducted using a hierarchical probability meta-model (HPM) and an illustrative case study. The case is a wet clutch system with the goal of minimizing drag torque. Across both studies (HPM and case study), it is found that the proposed procedure is superior in performance to the traditional method. This is consistent across various levels of experimental error, comparisons at different resource intensities, and through a wide array of response surfaces generated by the meta-model. At best, the proposed algorithm provides an expected value of improvement that is 15% higher than the traditional approach, at worst, the two methods are equally effective, and on average the improvement is about 10% higher. These findings suggest that the ensemble of *a*OFATs can be an effective and often preferred alternative to the use of orthogonal arrays for experimentation.

This thesis also shows that it more effective to apply ensemble procedures to *a*OFAT versus applying ensemble techniques on multiple, highly-fractioned orthogonal designs (each being as resource intensive as a single *a*OFAT). A theoretical discussion explaining the reasons for the superior performance of the proposed algorithm supports the empirical findings.

Acknowledgements

I would like to express my deepest gratitude to my thesis advisor, Dr. Dan Frey. He has been an excellent source of support and inspiration. Dr Frey always encouraged me to work on problems of my own choosing and interest, while taking it entirely upon himself to relate my work with that of the projects that ultimately funded my education. This freedom is something that I have really enjoyed. It has, in retrospect, allowed me to say that I carved out a research program that I was truly interested in. I would also like to thank Dr. Frey for encouraging me to participate in various conferences, and research ventures. In particular, I took great pleasure in research collaborations that he initiated with the Max Planck Institute of Human development, ultimately leading to a research trip that I made to Berlin, Germany.

Committee members Roy Welsch, Chris Magee and Richard Larson have helped me a lot with my research. Their insights during the committee meetings are greatly appreciated. During my first few committee meetings I received a lot of honest criticism. This was instrumental to changing various aspects of my research and I am thankful for this.

I really enjoyed the company and support of my colleagues at the Engineering Systems Division. I had the opportunity of living with two of them, Sidharth Rupani and Dan Livengood, over the last two years. They are excellent friends who have been very supportive and helpful during this period. My conversations with Sid have been a great event to look forward to in the evenings at the apartment. These discussions have been intellectually stimulating, and have grounded various concepts in statistics and engineering systems.

I thank my Robust Lab colleagues. They have made it a pleasurable experience to share the workplace. Lab interactions, both formal and informal, have been great (this includes conflicts over the lab's thermostat). I have especially enjoyed discussing research ideas with Chad Foster. His feedback on my research plans over the last couple of years has helped a lot.

There have been other friends and family members who have been of great assistance to me through this process. They have done this by expressing confidence in me, promising support when needed, and empathizing through hard times. This form of emotional support is invaluable. My mother is in the top of this list. She has made such a genuine effort to understand my research, my department, and the key role of my committee members, all of this without having the slightest background in this topic. This makes her effort really special. The extent to which she takes pride in my accomplishments is just amazing. I would also like to thank my close friend from India, Bhakti. She has listened to my rants on the workload patiently. She remained supportive even through my highly undisciplined working practices. With each task, she gently encouraged me to start working well ahead of deadlines, while I evaded it. As the deadlines closed, she tolerated me through my panic moments without ever mentioning that she told me so.

Table of Contents

1 Introduction 11

1.1 Basics of Designed Experiments 11

1.2 Motivation 14

1.2.1 Motivation: DOE literature 14

1.2.2 Motivation: Algorithmic Properties 16

1.3 Literature Review 21

1.3.1 Background on Adaptive one-factor-at-a-time (*a*OFAT) Experimentation 21

1.3.2 Background on Ensemble methods 25

1.4 Research Summary 28

1.4.1 Research hypotheses 28

1.4.2 Research Approach 28

1.4.3 Main Contribution 29

1.5 Structure of the Thesis 31

2 Evaluating Experimental Algorithms using the Hierarchical Probability Model 32

2.1 Experimental Algorithms 32

2.1.1 Ensemble of *a*OFATs 33

2.1.2 Fractional Factorial Design and Analysis 35

2.2 Hierarchical Probability Model (HPM) 37

2.3 Example with a Single Response Equation 42

2.3.1 Ensembles of *a*OFAT 43

2.3.2 Fractional Factorial Design 45

2.4 Results and Discussion 47

3 Case Study: Improving the Performance of a Wet Clutch 49

3.1 Background 49

3.2 Model Construction 51

3.3 Results and Discussion 53

4 Comparing an Ensemble of *a*OFATs with an Ensemble of Factorial Designs 56

4.1 HPM Comparison and Results 56
4.2 Theoretical Discussion: Why Ensemble Methods to *a*OFAT? 60
5 Extending Ensemble Techniques and Factorial Design Analyses... 66
5.1 Methodological Extensions of the Proposed Algorithm 66
5.1.1 Starting locations of *a*OFATs 66
5.1.2 Order of Variable Changes 67
5.1.3 Aggregation Strategies 68
5.1.4 Evaluating the various experimental algorithms 69
5.2 Analyzing factorial designs and comparing algorithms 73
6 Conclusions and Future Work 77
6.1 Future Work 79
References 81
Appendix A 87
Appendix B 90

List of Figures

Figure 1.1 Comparison of resource consumption in *a*OFAT and Resolution III arrays using two-level factors 15

Figure 1.2 Graphical representation of the adaptive one-factor-at-a-time experimentation for a system with three factors at two levels per factor (adapted from Frey and Jugulum (2006)) 22

Figure 2.1 Comparing an ensemble of 4 *a*OFATs with a 2^{7-2} fractional factorial array using the HPM 47

Figure 2.2 Comparing an ensemble of 8 *a*OFATs with a 2^{7-1} fractional factorial array using the HPM 48

Figure 3.1 A wet clutch pack (adapted from Lloyd (1974)) 50

Figure 3.2 Comparing an ensemble of 4 *a*OFATs with a 2^{7-2} fractional factorial array for wet clutch case study 54

Figure 3.3 Comparing an ensemble of 8 *a*OFATs with a 2^{7-1} fractional factorial array for wet clutch case study 54

Figure 4.1 Comparing an Ensemble of four *a*OFATs with an ensemble of four 2^{7-4} Fractional Factorial arrays using classical analysis and take-the-best, and a single 2^{7-2} Fractional Factorial array 59

Figure 4.2 Bias and Variance Decomposition (adapted from Hastie, Tibshirani and Freidman 2001) 61

Figure 4.3 Flight times against increasing standard deviation of experimental error for the electric powered aircraft case when an *a*OFAT was applied (on the left), and a 2^{7-4} resolution III design was applied (on the right); with 10^{th} to 90^{th} percentiles (adapted from Frey and Wang, 2006) 64

Figure 5.1 Comparing an Ensemble of four *a*OFATs with nine different strategies, using the HPM; with 25^{th} and 75^{th} percentile error bars (Average) 71

Figure 5.2 Comparing multiple methods of analyzing a 2^{7-2} Fractional Factorial array, using the HPM 75

Figure 5.3 Comparing the various strategies of analyzing Ensemble *a*OFATs and Fractional Factorial arrays 76

List of Tables

Table 2.1 Select starting locations for four *a*OFATs, with 2 levels each, which maximize the mean Euclidean distance between each other 34
Table 2.2 Quantifying the HPM - Results from Li et al., (2006) 39
Table 2.3 *a*OFAT-1: an adaptive one-factor-at-a-time analysis on data from equation 11 43
Table 2.4 *a*OFAT-2: an adaptive one-factor-at-a-time analysis on data from equation 11 44
Table 2.5 *a*OFAT-3: an adaptive one-factor-at-a-time analysis on data from equation 11 44
Table 2.6 *a*OFAT-4: an adaptive one-factor-at-a-time analysis on data from equation 11 44
Table 2.7 Ensemble *a*OFAT aggregation process 450
Table 3.1 Input variables and their level settings from the wet clutch case 50
Table 3.2 Main effects for the wet clutch case 51
Table 3.3 Significant Two-way interaction effects for the wet clutch case 51
Table 3.4 Significant Three-way interaction effects for the wet clutch case 51
Table 4.1 Conceptual analogues between data mining terminology and experimentation 63
Table 5.1 Nine Ensemble *a*OFAT strategies tested in this thesis 70
Table 6.1 Summary of the performance of algorithms in this thesis 78

1 Introduction

The operation of an engineering system can be conceptualized as a combination of materials, methods, people, environment and policies (inputs), which when used together perform a service, produce a product, or complete a task (outputs) (Schmidt and Launsby, 1994). Design of Experiments (DOE) is a formalized procedure of systematically and purposefully making changes to controllable variables (a subset of the inputs), in an effort to observe their influence on the outputs, and thereby improve them. This thesis involves the advancement of experimentation methods and practices.

The applications of designed experiments span both social and technical systems. Evidence of its use can be seen in engineering, social sciences, natural sciences and management[1] (For examples see Schmidt and Launsby, 1994; Wu and Hamada 2000; Montgomery, 2001). Its wide array of applications also makes this an apt tool to deal with complex systems that are an amalgamation of the above.

1.1 Basics of Designed Experiments

This section provides an introduction to designed experiments. It presents an overview of the basic principle, and is intended as the bare minimum required to understand terminology used later in this thesis. A more detailed, yet mathematically simple discussion of DOE can be found in Schmidt and Launsby (1994). A technically advanced treatment can be gained from Wu and Hamada (2000) or Dean and Voss (1991). The section can be skipped if the reader is familiar with this topic.

[1] While there are numerous examples of formal experimentation in engineering, social science and natural science, it is rarely seen as a process improvement procedure in management. A PhD thesis that discusses the role of DOE in corporate decision-making can be seen in Tang (2006).

Design of experiments is an offline quality improvement process. Formalized sequences of steps have been presented by various authors to help plan and organize the experiment. A list of the main steps, adapted from Barker (1985), Schmidt and Launsby (1994) and, Wu and Hamada(2000) is presented here:

1. Experimental Objective: The objective of the experiment needs to be clearly stated by all the stakeholders involved with the process. The reason for any designed experiment is to provide data that can be used to answer well thought out questions. Failure to allocate sufficient time and thought in formulating these questions often result in wasting resources throughout the reminder of the experimental process. It is important that all of the key players (management, engineering, experts in area, operators, and analysts) be in involved in the planning phase

2. Choice of response: The response/s that the experiment seeks to improve needs to be selected and quantified. This process typically involves identifying the key process output variables (KPOV's) that serve as indicator to any problems or concerns that arise in the first phase (experimental objective). The chosen response could be discrete or continuous. The latter, however, is preferred. The response also needs to be classified according to stated objectives. Such as, larger-the-better or smaller-the-better or nominal-the-best.

3. Choose Factors and Levels: This step concerns identifying the key-process-input-variables (KPIV's). Inputs are selected based on an understanding of their likely relation to the response. Once selected, the variable is referred to as a 'factor'. In order to study the influence of an input variable on the output, two or more values (or stages) that the input variable can assume in the context of the experiment is selected. These are referred to as 'levels' of the factor. A 'treatment' or 'treatment

combination' is a combination of the inputs at different levels (input vector). It can be readily seen the that number of possible treatments is equal to $\prod_{n=1}^{F} L_n$ Where, F is the number of factors and L is the number of levels that the n^{th} factor can assume.

4. Choose experimental plan: The experimental plan indicates the experimental algorithm. In some cases, such as the fractional factorial arrays, it can be characterized a subset of all possible treatments that exist (sometimes each treatment could be tested multiple times). The experimental plan should be picked such that it is in accordance with the available resources, or number of factor level changes possible. There are a few different criteria for selection of the experimental plan. Some seek to maximize the knowledge that could be gained from running the plan, and thereby model the system. Other plans use expected value of improvement from running the experiment as the metric of importance. Motivated by a line of reasoning in Frey and Wang (2006), in this thesis, we use the latter criterion behind selection of experimental plans. At this stage, one is ready to conduct the experiment.

5. Analyze data: An analysis appropriate for the design follows the actual experiment. In some cases, an analysis of the data requires the construction of a model following statistical analysis to identify the important effects and quantify their exact influence. In others, it requires using direct techniques that will identify treatment that provided the best output.

6. Draw conclusions and make recommendations: Following the data analysis it is possible to identify those input settings that are ideally suited for maximizing the performance of the system. The conclusion should refer back to the stated objectives of the experiment

1.2 Motivation

The motivation for this thesis is two-fold. First, is based on past literature, and industry practice concerning designed experiments. The impetus here comes from realizing the scope for improvement in the current methods in terms of their performance, as well as their applicability to engineering systems in the field.

The second motivation is more conceptual. It discusses the nature of algorithms, across different domains, which are used to solve large and complex problems. Insights from this are used to motivate a new and improved breed of experimental algorithms.

1.2.1 Motivation: DOE literature

Experiments require studying system performance (output) at different input factor settings. In certain cases this could mean that a different resource is consumed for each input configuration that is studied. In other cases, the experiment may require changes to be made in the control variables to attain a new input vector (reusing the same resource). In the former scenario, the resources required for experimentation might be limited. In the latter, making the changes to input variables could be expensive and time consuming. It is therefore very important to study the effectiveness of experimental algorithms in the context of the number of experiments (trials) run.

In this thesis we look at two algorithms: Adaptive one-factor-at-a-time (aOFAT) and orthogonal arrays. The aOFAT is an experimental procedure where a single factor is changed in each experiment. Changes to variables that improve the response are retained and ones that are detrimental are reversed[2]. This practice is repeated sequentially until all factors are exhausted, thus always taking $1+\sum_{i=1}^{m}(n_i-1)$ experiments when m variables are being tested, at n_i levels each. Orthogonal arrays are pre-built matrices which are subsets of the full factorial design (which is an enumeration of all possible factor combinations at all possible levels). Examples of such matrices are, fractional factorial designs, composite

[2] A more detailed description of the experimental algorithms will be discussed in section 1.2.1

designs, Latin squares/hypercube, etc,. A more in depth background on the classical methods can be found in Schmidt and Launsby (1994) or Montgomery (2001). A common technique discussed (and compared) in this research is the fractional factorial design. Here, fractions equaling $\frac{1}{n^m}$ (where there are m variables each at n levels) of the full factorial are selected for the experiment. The selection of the appropriate fraction is based on mathematical criteria such as maximum resolution (Box and Hunter, 1961) or minimum aberration (Fries and Hunter, 1980). A detailed set of studies on picking the most suited fraction with theoretical reasoning can be found in the literature (Chen and Wu, 1991; Chen, 1992; Chen et al., 1993; Chen, 1998; Tang and Wu, 1996).

Given the resource requirements for the two algorithms, a reasonable comparison can be made between the *a*OFAT and Resolution III fractional factorial designs. The Resolution III designs are one of the most frugal fractional factorial designs. They are usually more, and at least equal, in resource requirements as the *a*OFAT. Figure 1.1 illustrates the resource requirements of these two algorithms for different numbers of controllable variables when each variable has two-levels.

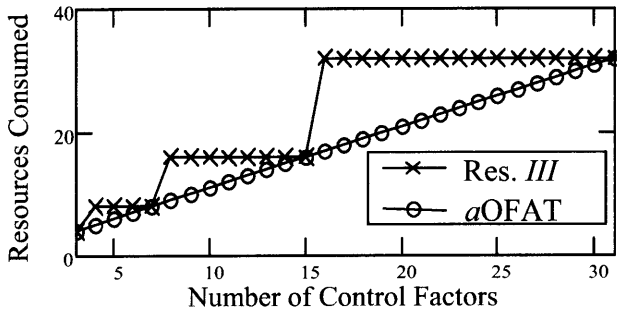

Figure 1.1 Comparison of resource consumption in *a*OFAT and Resolution III arrays using two-level factors

Under certain, standard, experimental conditions the *a*OFAT outperforms the resolution III fractional factorial arrays (Frey, Engelhardt, and Greitzer., 2003; Frey and Wang, 2006; Frey and Sudarsanam 2008). Section 1.3.1 discusses this in greater detail.

However, in many situations the number of resources available for experimentation could be significantly more than that required by *a*OFAT or resolution III arrays. It is also possible that we would like to extend our experimental plan upon finding the results from

the earlier experiments. In such situations the fractional factorial plans seem relatively more adaptable to the designers' needs. It is possible to find fractions of different degrees, for example, when seven factors are involved, plans ranging from 2^{7-4} to 2^7 are considered reasonable options depending on the resource availability (Wu and Hamada; 2000). Also, foldover techniques (Box and Hunter; 1961) and semi-folding methods (Daniel; 1962, Daniel;1976, and Barnet et al.,; 1997) allow us to make sensible extensions of even the most highly fractioned array such that there is no penalty incurred from initially planning a frugal design. No such extensions are currently available with *a*OFAT. This specific shortcoming, in light of its superior performance per unit resource consumption (as shown by Frey et al., 2003; Frey and Wang, 2006; Frey and Sudarsanam 2008), makes it important to investigate possible extensions of *a*OFAT experimentation. A solution would retain the performance benefits of the *a*OFAT while simultaneously providing designs suited for various amounts of resource availability. This thesis seeks to provide one such solution.

1.2.2 Motivation: Algorithmic Properties

The motivation to use an ensemble of adaptive one-factor-at-a-time experiments is grounded in the algorithmic properties of such a strategy. The various aspects in the nature of such an algorithm, and how each serves as a motivation, are discussed in the following sections

(i) Why Many Are Better Than One

The idea that large groups of naïve individuals, when brought together in the right way, are better at solving large-scale, complex problems than an exceptional few is the central premise of James Surowiecki's book "Wisdom of Crowds" (Surowiecki, 2004). A claim supported by an eclectic set of examples ranging from how ant colonies function in the wild, to investors in a stock market, scientists collaborating on a research project, or people guessing an ox's weight at a local fair. In fact, understanding the value of such 'swarm intelligence' is at the root of many problem solving initiatives adopted by

businesses. Examples include a company that solved a routing and truck allocation problem for pick up and delivery, by using ant colony inspired rules. A large airline conglomerate solved the complex problem of gate and runway allocation for airplanes in airports. Delphi methods and prediction markets are used in management for various forms of business forecasting. The common strand among all these initiatives is that the building block in each case (truck driver, pilot, etc.,) always act on simple rules, localized information, decentralized decisions and are unaware of the other agents (Miller, 2007).

This principle, seen in various industrial and social practices, goes even further. We can find an array of examples in formal academic literature. Jordan and Jacobs (1994) write that 'Divide and conquer' algorithms enjoy a wide range of applicability in applied mathematics. Optimization techniques like Genetic algorithms, Simulated Annealing, and the Ant Crawl algorithm use these concepts in part or whole. Agent-based modeling is used to represent complex systems. In Machine learning a formalized set of methods called 'ensemble techniques' are used to create superior predictors.

Yet, such a divide-and-conquer, decentralized, agent-based approach has not been seen in formal design-of-experiments. Standard procedures recommend a single full or fractional design array, depending on resource availability. This motivates us to examine the effectiveness of aggregating multiple highly frugal experiments, and comparing its performance to a single standard array.

(ii) Why *a*OFAT: Greedy Algorithms

The principle described so far motivates the use of multiple small experiments as a replacement for one large experiment, the motivation to use *a*OFATs in this context as the building blocks is also manifold.

Most importantly, it can be argued that the algorithmic structure of *a*OFAT is well suited to the exploit gains that are accrued from aggregation. The *a*OFAT in many ways is

similar to a breed of algorithms called 'greedy algorithms'. "*A greedy algorithm always makes the choice that looks best at the moment. That is, it makes a locally optimal choice in the hope that this choice will lead to a globally optimal solution*" (Cormen et al., 1990). In a stochastic environment, this makes greedy methods assume 'low bias and high variance', when compared to their more conservative, non-greedy counterparts. Similarly, when compared to a fraction factorial design, *a*OFAT has properties reflective of a lower bias and higher variance. This is explained in section 4.2. The specific fact that aggregation techniques target the variance component and not the bias is what makes *a*OFAT suited for ensemble techniques. This theoretical argument for this is also made in section 4.2.

(iii) Why *a*OFAT: Adaptive Algorithms

Adaptive one-factor-at-a-time experimentation, as the name suggests, is an adaptive procedure. Academic work from prominent researchers in experimentation, such as Box and Liu (1999), Friedman and Savage (1947), Daniel (1973), Sasena et al., (2005) has been in support of using an adaptive, sequential approach.

George Box and Patrick Liu emphasize the importance of using an adaptive approach to experimentation in the following quote:
An industrial innovation of major importance, such as the development of a new drug or the design of a new engineering system, comes about as the result of investigation requiring a sequence of experiments. Such research and development is a process of learning: dynamic, not stationary; adaptive, not one-shot. (Box and Liu, 1999)

While Box and Wilson (1951) suggest Response Surface Methodology as one way of "*practicing the process of investigation*" suggested in the above quote (Box and Liu, 1999), *a*OFAT is a conceptually simpler yet adaptive approach to experimentation.

Friedman and Savage call attention to the trend where non-adaptive procedures lead to spending unnecessary resources at points far from the target of interest. Cuthbert Daniel highlights the benefit of being able to react to data in a sequential process, in the following quote:

Some scientists do their experimental work in single steps. They hope to learn something from each run or trial. They see and react to data more rapidly. The statistician who tells such an experimenter that he can secure greater precision or validity by doing 16 or more runs in a balanced set, may be listened to with courtesy, but rarely with enthusiasm (Daniel, 1973)

The works of some of these authors have directly contributed to one-factor-at-a-time plans and will be discussed in greater detail when reviewing the literature of this experimentation technique in section 1.3.1. Insights from such academic work serve as a key motivation to look at an experimentation technique that is adaptive.

(iv) Why *a*OFAT: Algorithmic Modeling Culture

In the data mining literature Leo Breiman describes two cultures in statistical modeling: The data modeling culture and algorithmic modeling culture (Breiman, 2001b). The data modeling culture is one where a fixed, presupposed mathematical structure (usually stochastic) is assumed in a prediction problem. The field data is used to estimate the parameters in the mathematical structure. Examples of such a procedure include various forms of regression, discriminant analysis, neural nets, support vector machines, etc. Alternatively, there is another culture which uses no presupposed mathematical structure or statistical model, but only algorithmic instructions in order to establish a relationship between predictor (input) and response (output) variables. Examples of this kind include nearest neighbor methods, CART (Classification And Regression Trees) (Breiman, 1984), etc. Breiman argues that the focus of the statistical community on data models has:

- *Led to irrelevant theory and questionable scientific conclusions*

- *Kept statisticians from using more suitable algorithmic models*
- *Prevented statisticians from working on exciting new problems*

(Breiman, 2001b)

A similar criticism can be made of current practice, and literature in design of experiments. Orthogonal arrays are fixed matrices, while *a*OFAT is best described by a set of instructions (there is no one matrix that represents *a*OFAT for a set number of resources). By incorporating algorithmic ideas in experimentation one can create an adaptive approach, which is useful for planning experiments, and also for selecting and analyzing them. This in turn will be more effective in improving the system.

1.3 Literature Review

This research lies in the intersection of two topics: Adaptive one-factor-at-a-time experimentation from the DOE literature, and Ensemble Methods from the data mining literature. A discussion of past literature on *a*OFAT experimentation is presented in section 1.3.1. This section does not seek to review work on factorial designs or design of experiments as a whole. A comprehensive treatment of the literature in design of experiments over the last hundred years has been summarized as a journal article by Atkinson and Bailey (2001).

1.3.1 Background on Adaptive one-factor-at-a-time (*a*OFAT) Experimentation

The *a*OFAT is an experimental procedure where a single factor is changed in each experiment. Changes to variables that improve the response are retained and ones that are detrimental are reversed. This practice is repeated sequentially until all factors are exhausted, thus always taking $1+\sum_{i=1}^{m}(n_i-1)$ experiments when m variables are being tested at n_i levels each. Figure 1.2 illustrates the mechanism associated with *a*OFAT as it is applied to a system with three factors, and each factor is tested at two levels.

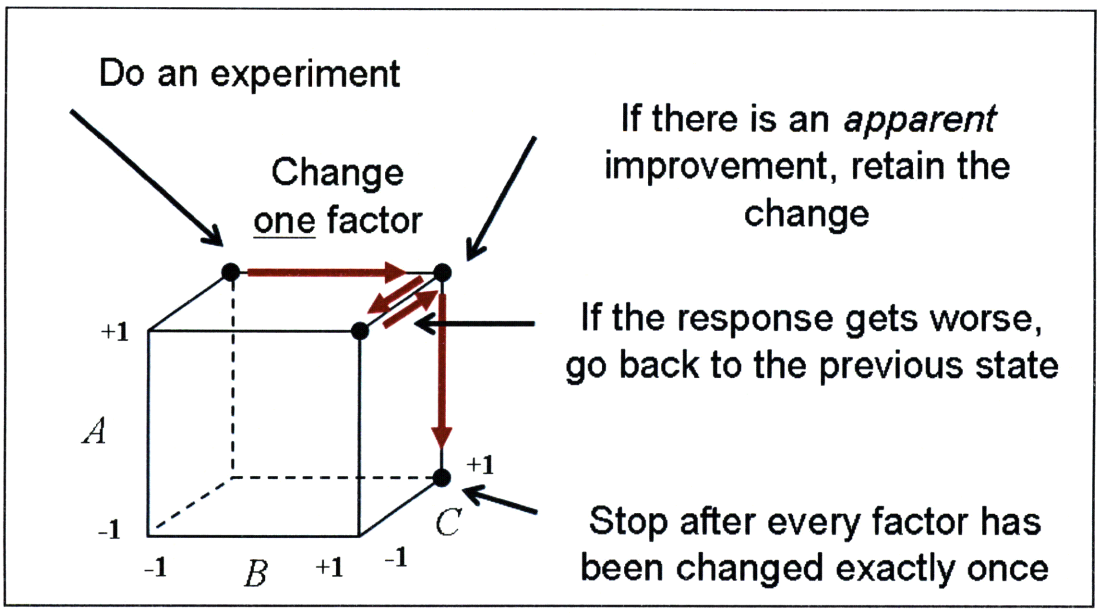

Figure 1.2 Graphical representation of the adaptive one-factor-at-a-time experimentation for a system with three factors at two levels per factor (adapted from Frey and Jugulum (2006))

One-factor-at-a-time (OFAT) experimentation is generally discouraged in the literature on experimental design and quality improvement (Box, Hunter, and Hunter, 1978, Logothetis and Wynn, 1994, Czitrom, 1999, Wu and Hamada, 2000). Reasons mentioned include:

- They require more runs for the same precision in effect estimation
- Interactions between variables cannot be captured,
- Can miss optimal settings,
- It is an ad-hoc procedure which does not allow us to form a mathematical model of the system.
- Sensitive to biases in time trend since randomization is not possible.

Citing the above reasons, researchers have preferred, and made theoretical arguments for an alternate form of experimentation, Orthogonal Arrays. The use of orthogonal arrays was formalized by R.A. Fisher in the 1920's, where this approach was used to study crop variation (Fisher, 1926). Unlike *a*OFAT, the orthogonal arrays result in a 'balanced' design (all the levels of each factor are tested an equal number of times in the

experimental plan). It is claimed that orthogonal arrays can provide factor estimates with greater precision and our less susceptible to the above mentioned weaknesses of *a*OFAT.

While this might be true, *a*OFAT has one important advantage over orthogonal arrays, in that it is adaptive in nature (The treatment combination for each subsequent experimental trial in an *a*OFAT is dictated by the findings from the previous trials). Milton Friedman and Leonard Savage have favored iterative one-at-a-time plans by arguing that their adaptive quality leads to more trials being conducted in regions in the design space that are closer to the maximum (or closer to the desired target). Orthogonal arrays, in contrast, due to their fixed one-shot approach, set aside many experimental trials that are far from the maximum. Friedman and Savage state that such an approach could be preferred in an environment where interactions are likely to exist. The plan proposed here uses multiple rounds of *a*OFAT, where each new round starts with the combination of values of the independent variables reached at the end of the previous round (Friedman and Savage, 1947). Cuthbert Daniel also acknowledged the value of *a*OFAT experimentation by pointing out that such a procedure allows the experimenter to react to the data. Daniel claimed that when random error is small in comparison to the main effects, such experiments are economical and effective. Daniel suggests minimal augmentations to the one-at-a-time plan to make it more effective (Daniel, 1973). Although not specifically addressing a one-at-a-time approach, George Box and Patrick Liu have emphasized the importance of adaptive experimentation by stating that a dynamic approach is needed over a one shot approach (Box and Liu, 1999).

Other examples include, Qu and Wu (2004) who present two classes of one-at-a-time plans that are of resolution V. They state that such plans facilitate sequential learning while consuming fewer resources than a tradition resolution V design. McDaniel and Ankenman (2000) state that for 'small change problems', a one-factor-at-a-time in conjunction with Box-Behnken designs outperforms a resource equivalent strategy using fractional factorial designs when there is no experimental error. Koita(1994) and Koita and Staelin (1995) present a one-at-a-time strategy as a part of an overall experimentation

approach where fractional factorial arrays are used to identifying selected interactions and other optimal experiments are carried out.

Some of the insights from the above mentioned literature motivated Frey, Engelhardt, and Gretizer (2003) to look at a two-level, discrete version of *a*OFAT in an empirical study that used data from 66 response variables in 27 full factorial experiments. This meta analysis showed that when experimental error was low in comparison to main factor effects and when interaction effects was more than one quarter of all main factor effects, an *a*OFAT strategy achieves greater improvement than an orthogonal design that is equally resource intensive.

Frey and Wang (2006) provided the theoretical basis for the empirical findings. This was achieved by constructing a mathematical model to calculate the expected value of improvement in using an *a*OFAT and orthogonal designs. Based on this model, a comparison can be made between the exact improvement that can be afforded by *a*OFAT and resolution III arrays (A set of orthogonal arrays that are typically more, and at least equal, in their resource requirements as the *a*OFAT). The findings in this study complement that of Frey et al., (2003), in that when experimental error is reasonably low and the strength of interaction effects is reasonably high (both quantities measured in relation to the strength of the main effects), then *a*OFAT outperforms the Resolution III arrays. To determine what values are reasonable, a study by Li, Sudarsanam, and Frey (2006) looked at 113 published full factorial experiments. They determined that the ratio of interaction strength to main variable effects is typically 1/3.6.

Frey and Jugulum (2006) explicate these results by discussing the mechanism by which *a*OFAT delivers the improvements that it does. Four specific mechanisms are identified and explained. Frey and Sudarsanam (2008) corroborated the findings from the above mentioned studies by adopting a case study approach in the context of robust design. They present a methodological simple alternate to the traditional crossed array by replacing the inner array with an *a*OFAT design.

A portion of the literature in one-factor-at-a-time design also concerns itself with a continuous input design space. This is beyond the scope of this research. Many algorithms and ideas have been suggested under such a framework. It is important that the reader make this distinction (between continuous and discrete design space) in order to fully place the contribution of this thesis with respect to the general body of available literature There are many other cases of adaptive algorithms that use optimization on a continuous space to select, sequentially, the next point in the design space to conduct the experiment. Some examples include EGO by Jones et al. (1998), and SuperEGO by Sasena et al., (2005), Hardwick and Stout (1998), and Wang (2007). Even, some of discrete algorithms assume more than a predetermined number of levels for each factor. For example the algorithm by Friedman and Savage (1947) suggest convergence in each of their iterative rounds of *a*OFAT, suggesting that a continuous range of possible values is available for each input factor. In this thesis the scope is different. It is constrained to solving the problem where there are a fixed, finite number of levels for each factor and their settings or values are known before conducting even a single experiment.

1.3.2 Background on Ensemble methods

Statistical learning[3] is the art and science of extracting information from data in order to establish a logical relationship between the inputs (independent variables) and outputs (dependent variables), of any given process. This approach is especially valuable in situations where the exact association between inputs and outputs cannot be derived from first principles, or an in depth analysis of the underlying physics is not possible. In these cases, statistical learning can be used to fill in gaps left over from such an analysis, or build the relationships from scratch (Hand et al., 2001).

Building this logical relationship allows us to make predictions. In supervised learning, this is achieved through a learning analysis of sample data that contain instances where both input and output variables are known. This is referred to as 'training set' of data (Hastie et al., 2001). The learning process uses the training data to construct a model or

[3] The terms Statistical learning, Data mining, and Machine learning are used interchangeably in this thesis. The difference in terminology can be mostly attributed to the different fields of origin, and do not hold any significant, meaningful difference in the context of our study.

algorithmic structure that is capable of predicting the dependent variable given any vector of independent variables. In order to achieve this, a basic skeletal structure for predictor is assumed prior to any data analysis. The effectiveness of a given predictor can be judged by how accurately it achieves the intended task of prediction. It is good practice to set aside a pre-specified amount of data is separated for running model validations and tests. Most training algorithms seek to optimize a metric of performance. This metric depends on whether the output is a continuous variable (regression problems) or categorical one (classification problems). Metrics such as squared deviation, mean absolute deviation are seen in regression, while misclassification errors, cross entropy are used for classification. (Hastie et al., 2001)

As mentioned in section 1.2.2, when the predictor links the dependent and independent variable through some mathematical function, this broad genre of data mining procedures are referred to as "Data modeling culture" by Leo Breiman (Breiman; 2001b). Examples include multiple regression, logistic regression, discriminant analysis, neural nets, support vector machines, etc. An alternative mechanism, referred to as "Algorithmic modeling culture" involves the use of algorithmic instructions to predict the dependent variable from the independent variables. Examples here are nearest neighbor methods, random forests, etc. Ensemble techniques belong to this class of procedures (Berk; 2006).

Ensemble methods were primarily developed as a classification technique in machine learning. They are learning algorithms that construct a set of classifiers and then classify by taking some form of voting or aggregation of their predictions (Dietterich; 2000a). Originally the term was coined in the context of creating and aggregating multiple neural networks (Hansen and Salamon; 1990). However, since then, research work on ensemble techniques has primarily used 'Tree predictors'. In fact, classification and regression Trees are often used as a didactic tool to explain ensemble techniques (Berk; 2006, Breiman; 2001a, Friedman and Hall; 2007). A detailed description on Trees as predictors and the underlying algorithms, such as CART or C4.5, can be found in Breiman et al. (1984) and Quinlan (1993), respectively. As we will see later in this thesis (section 4.2), the reasons for suggesting a match between *a*OFAT and ensemble methods, is the same

reason that the Tree algorithms mentioned above (also 'greedy algorithms') are used as the building block for ensemble technique exercises.

The common aspect in all ensemble methods is that they seek to apply, on what is usually a deterministic algorithm, elements of stochasticity (whether it is on the input data, or algorithmic methodology). The end results are multiple, modified versions of a structurally similar set of predictors. For example, one of the most common, formalized prediction techniques that uses such an approach on Tree algorithms is called 'Random Forests' (Breiman, 2001a). Random Forests achieve the different training data sets through 'Bagging' or '**B**ootstrap **agg**regat**ing**' (Breiman, 1996). Here, the training data sets are formed by using bootstrap (sampling with replacement) replicates, where each replicate serves as a separate learning set. The differences in trees are also brought about through various other techniques. One common technique used in Random Forests is called 'Random Subspace method' (Ho; 1998). Here, a subset of the input variables is selected for decision-making at each node. That is, if there are M input variables, a number $m<<M$ is specified. Such that at each node, m variables are selected at random out of the M and the best split on this subset is used to split the node. The value of m is held constant during the forest growth. Another similar technique is 'Random Split selection' (Dietterich; 2000b). Here, at each node the split is selected randomly (as opposed to picking the optimal split) among K best splits while using the entire set of input variables for decision-making at each node. Other alternatives or modifications have been suggested by Amit and Geman (1997) and Robnik-Sikonja(2004). Additionally, all these methods have to deal with an aggregation technique which is central to the procedures effectiveness.

In this study we attempt to utilize the methodology seen in such procedures by applying the same experimental algorithm (aOFAT) with procedural modifications such as starting locations and order in which variables are changed. This thesis also looks at different aggregation techniques.

1.4 Research Summary

1.4.1 Research hypotheses

The broad hypothesis in this research is that an approach to experimentation which is more adaptive, algorithmic, decentralized and agent-based will lead to greater system improvement. We seek to test this through a specific experimentation procedure characterized by an ensemble of *a*OFATs. The detailed hypotheses in the context of this technique are:

- The application of ensemble techniques to create and aggregate multiple *a*OFAT procedures will outperform the use of a single, full/fractional factorial, orthogonal array when both methods are equally resource intensive.

- The use of ensemble techniques on *a*OFAT will outperform the use of ensemble techniques on multiple, highly-fractioned orthogonal designs (each being as resource intensive as a single *a*OFAT).

These empirical findings will be corroborated by a theoretical argument for why *a*OFAT methods are better suited for Ensemble techniques.

1.4.2 Research Approach

The research approach, which is discussed in detail in Chapter 2, is briefly introduced in this section as a research summary. The following steps summarize the approach adopted in testing the two main hypotheses discussed in section 1.4.1.

I. The Meta Modeling approach

1. Create a meta model that represents actual engineering systems which experiments can be carried out upon

(i) Use a mathematical structure/formulation that accounts for various regularities seen in engineering systems

(ii) Use real world data to estimate various parameters in the math formulation.

(iii) Create the meta model by coding the math model as a computer program

2. Instantiate multiple response surfaces using the meta model that has been constructed in step 1.
3. Simulate each experimental algorithm (ensemble *a*OFATs and orthogonal arrays) on the responses
4. Evaluate the primary variable indicating the overall performance of the algorithm
5. Analyze and visualize data

II. Case study approach

Apply the experimental algorithms on a case study of an actual engineering system and thereby validate the findings from the meta-analysis.

1.4.3 Main Contribution

The main contributions from this research can be summarized as follows:

1. Encourage the use of industrial experimentation as a means to improve system performance, by:
    a. Providing an algorithmically superior and more effective form of experimentation: This is quantified by providing an experimental strategy that provides a greater expected value of improvement in system performance, when compared to the current methods.
    b. Providing an experimentation strategy that is more responsive to the designers needs: The proposed method can cater to wider range of resource availabilities. By using multiple *a*OFATs, any resource

availability can be closely matched by using the nearest multiple of a single *a*OFAT's resource consumption.

c. Encouraging concurrent engineering practices: Since this procedure requires multiple *a*OFATs that operate independently of each other, different teams within the same company can be assigned the job of running an individual *a*OFAT simultaneously. Since each *a*OFAT is a complete experimental investigation on its own, many of the benefits that come from performing a goal oriented exercise (as opposed to different teams just running unrelated treatment combinations) can be benefited from.

2. Provide insights into the general properties of algorithms used to solve large complex problems. Specifically, provide a context for the use of decentralized, agent-based approaches to solving problems. Offer insights on what types of agents are better suited for such a decentralized approach. Look at the relative advantages of using greedy algorithms as agents. Also, show the value of algorithmic and adaptive approaches as opposed to structured ones. Much of this contribution can be generalized to areas beyond experimentation.

1.5 Structure of the Thesis

The subsequent chapters in this thesis are structured as follows: **Chapter 2** starts with an introduction of the experimental algorithms that are compared in this study. It includes the proposed method of Ensemble *a*OFATs, and the traditional approach of factorial designs. The chapter then introduces the procedure used for making comparisons. This refers to the concept of creating and using a meta-model in conjunction with simulations to evaluate the two algorithms. A randomly selected equation from the meta-model is used as an illustration to guide the reader through the actual application of both algorithms. The chapter is completed with the results from the analysis performed and a discussion with the interpretation of the results. **Chapter 3** covers a case study. The case pertains to a wet clutch system, where the goal of the experimental exercise is to minimize drag torque. A background on the case is provided, followed by test methodology, and finally the results with a discussion. In **Chapter 4** the meta-modeling approach is again used to test the secondary hypothesis of this thesis. A comparison between the proposed methodology and an ensemble of equally frugal factorial designs is conducted. The chapter concludes with a theoretical discussion, which explains the empirical findings. **Chapter 5** provides a set of methodological extensions to the proposed algorithm. These include using different strategies for starting points, order of variable changes, and aggregation of *a*OFATs. Alternate ways of analyzing the fractional factorial arrays is also explored. **Chapter 6** presents a summary of the findings and potential areas for future work.

2 Evaluating Experimental Algorithms using the Hierarchical Probability Model

In this chapter, the two experimental algorithms (ensemble of *a*OFATs and Fractional factorial arrays) are tested using a hierarchical probability meta-model. Section 2.1 introduces the algorithms in the context of the specific experimental plans that are tested. Section 2.2 introduces the meta-model and the procedure used to make the comparisons. Section 2.3 presents a single example of a response equation with both experimental algorithms applied to improving the system. And, section 2.4 presents the results with the discussion on findings.

2.1 Experimental Algorithms

The first phase of comparisons in this study will look at variables that have exactly two levels each. In order to reasonably compare two experimental strategies, it is important to adopt plans that are equally resource intensive. To achieve this, we take seven design variables. This will lead to a single *a*OFAT plan requiring eight resources ($7.(2-1)+1$), and ensembles of *a*OFAT requiring resources that are multiples of eight. Fractional factorial designs, when two-level variables are used, consume 2^{7-k} resources (where k is the degree of fractioning). Such a setup will allow us to make comparisons between these two strategies at multiple degrees of resource availabilities. Specifically, in this study we will compare a low-resource and a high-resource version. The low-resource version will consist of an ensemble of four *a*OFAT designs and are compared to a 2^{7-2} fractional factorial design. With seven explanatory variables at two levels, both experimental plans will need thirty two resources. The second pair-wise comparison, which is a high-resource intensive assessment, is between eight *a*OFATs and a 2^{7-1} factorial array. Both of these experiments require sixty four resources.

2.1.1 Ensemble of *a*OFATs

In our study we apply a divide-and-conquer, decentralized approach of using multiple *a*OFATs , each seeking to find the optimal design setting. The findings of every *a*OFAT are aggregated to provide a single recommendation. One key strategic aspect that influences the performance of this algorithm is the starting locations of the different *a*OFATs.

This refers to the default initial level settings for the variables. There are two reasons why this aspect is thought to have a large impact on the algorithms performance. First, past literature, when referring to the wisdom of crowd effect in such algorithms, acknowledge the importance of diversity among the individual agents (Surowiecki, 2004 and Miller, 2007). Also, in the academic literature of ensemble techniques Amit and Gelman (1997), and Breiman (2001a) note the importance of low correlation between the individual predictors. Having different starting locations for the *a*OFAT experiments is seen as a way of achieving this dissimilarity or low correlation. The second motivation to look at starting points of *a*OFATs comes from previous findings that the performance of this technique is greatly influenced by changing the starting location. Frey and Sudarsanam (2008) in an investigation of *a*OFAT in the context of robust design, present an algorithm where the each factor for *a*OFAT is started by setting it at the correct stage with a 75% probability, when the factor has two levels. This modification to the *a*OFAT led to a considerably improved performance when compared to the standard algorithm (Frey and Sudarsanam, 2008). These findings suggest that by testing *a*OFATs at different starting locations, there is a greater probability of finding a starting location that is closer to the optimum.

Three plans are tested to look at the influence of starting locations of *a*OFAT. The first is a plan that maximizes the dissimilarity of starting points. Here, *a*OFAT's are started at locations which minimize the average correlation between all pairs of starting points (maximize the negative correlation). This is synonymous to maximizing the mean Euclidean distance between each pair of starting locations in the input variable design space, when the variable levels are coded as +1 and -1. An alternative strategy looks at

the effectiveness of a method where the initial locations are independent of each other (zero correlation as opposed to negative correlation). This is achieved by selecting each variable's initial setting randomly for each *a*OFAT. And lastly, a plan where the same location is selected for all *a*OFATs is looked at. The first plan performed considerably better in terms of maximizing the performance metric for our study. A more detailed comparison and demonstration of results between these plans is shown in Chapter 5. However, for the experimental comparison, both low-resource and high-resource, the plan of maximizing the negative correlation is the criterion used to select starting locations. Note that while the aim here is to maximize the *average* Euclidean distance between any pair of points, it is not necessary that all points are equally distanced from each other, and this is often the case. In the four *a*OFAT ensemble case, the starting vectors shown in table 2.1 achieved this.

	A	B	C	D	E	F	G
*a*OFAT 1	-1	-1	-1	-1	-1	-1	-1
*a*OFAT 2	-1	+1	-1	+1	-1	+1	-1
*a*OFAT 3	+1	-1	+1	-1	+1	-1	+1
*a*OFAT 4	+1	+1	+1	+1	+1	+1	+1

Table 2.1 Select starting locations for four *a*OFATs, with 2 levels each, which maximize the mean Euclidean distance between each other

In addition to the location of the starting point, there are other strategic aspects that influence the performance of ensemble *a*OFATs. They are the order in which the variables are changed, and the aggregation technique used. In this thesis various different plans are tested for each of these strategic aspects and their effect is discussed in Chapter 5. In this section we present a more elemental version of the algorithm. The order in which variables are changed is fixed and the same for each *a*OFAT. The aggregation technique uses a simple voting strategy where each *a*OFAT votes on each variable's level settings based on its experimental findings. Each vote is weighted by the performance of the response which is measured for the respective *a*OFAT, during its proposed best treatment combination run. To ensure robustness, the weighting is not directly based on the response magnitude, instead, through a ranking mechanism similar to the Mann-Whitney *U* test, also known as, Wilcoxon rank-sum test (Wilcoxon; 1945). The rank sum

for the different levels of each variable is computed at each level and final design decisions are made based on the aggregation of these ranked recommendations. Since the aim here is to maximize the response, *a*OFATs that have higher responses have higher ranks and therefore contribute more in terms of the rank sum. This method, probabilistically, provides due advantage for *a*OFATs that found truly optimal solutions and penalizes the voting power for ones that just found inferior solutions, or local optimums. Also, this approach is seen as an intermediate between a strategy that simply votes (equal weights) and one which just adopts the recommendation of the *a*OFAT which had the best response (take-the-best). It is important to see that using the final response magnitude only (as in the case of take-the-best) from the final recommendations exposes the method to erroneous conclusions due to experimental error seen during the final run. Ties resulting in this algorithm, which are usually infrequent, were resolved through the assumption of squared loss function. This placed more benefit to treatments that were consistent.

2.1.2 Fractional Factorial Design and Analysis

Ensemble *a*OFATs are compared with fractional factorial designs. Specifically, four *a*OFATs will be compared to a 2^{7-2} (32 runs) fractional factorial design and eight *a*OFATs with a 2^{7-1} (64 runs) factorial array. The fractional factorial design is selected in accordance to maximum resolution (Box and Hunter; 1961) and minimum aberration (Fries and Hunter; 1980) criteria. The arrays that meet these criteria are summarized in Wu and Hamada (2000). It is seen that four arrays for the size 2^{7-2} and two arrays for size 2^{7-1} meet these requirements of optimality. Our simulations randomly select between the four arrays for 2^{7-2} design and two arrays for 2^{7-1} design.

Statistical analysis of the fractional factorial arrays is carried out using a slightly modified version of stepwise regression as discussed in Draper and Smith (1981) and summarized in Hastie et al., (2001). This is a procedure which is a hybrid version of forward stepwise selection and backward stepwise selection. In forward stepwise

selection, the model starts with only the intercept. Subsequent addition of variables is done on a step-by-step basis based on a criterion that evaluates goodness of fit, such as the p-value from the F-Statistic. The stopping criterion is met when the addition of a term does not improve the p-value or when the percentage increase is below a tolerance level. Similarly, backward selection starts with all explanatory variables in the model, and sequentially removes them one-by-one. The hybrid method, which is the one used in our study, starts with an empty model (only constant term), but at each stage considers both options of adding and dropping terms. The criterion for addition and removal is set at different levels so that infinite loops are avoided. In many cases using a fixed, pre-determined α (p-value of 0.05 for entry of terms and 0.15 for removal of terms) provided very few significant lower order terms, which was insufficient to make design vector recommendation. For example, if only two three-way interaction terms are found significant, this hardly helps us narrow down to a single treatment combination that can be proposed as the best. And yet, the identification of the three-way interactions is useful information that cannot be ignored. In such cases α (critical p-value for adding terms) and the critical p-value for removing terms is adjusted to allow for concrete design recommendations. This modification greatly improved the performance of the fractional factorial designs over using a standard stepwise regression with fixed p-values.

2.2 Hierarchical Probability Model (HPM)

There are certain structural regularities often seen in an experimental environment. The Hierarchical Probability Model is a mathematical form of expressing an engineering system which can account for these regularities. A discussion of three prominent structures discussed in literature can be found in Wu and Hamada (2000). A brief summary is provided here:

- *Sparsity of effects*: In the context of experimentation, this principle states that a small percentage of the tested effects (main effects and higher order interactions) will influence the response in a statistically significant way. One of the earliest explicit statements of this concept is seen in Box and Meyers (1986).
- *Hierarchy*: The hierarchical ordering regularity states that lower order effects are more likely to be important (statistically significant and have larger coefficients) than higher order ones. This means main effects are likely to be more important than two-way interactions, and two-way interactions are likely to be more important than three-way, and so on. Chipman(1996) discusses this concept, including choice of priors for the higher order interactions.
- *Heredity/ Inheritance*: This principle suggests that an interaction is more likely to be significant if its parent main effects are significant. In the case of a two-factor interaction, the presence of both parent main effects makes it more likely that the interaction is active, more than if either one of the parent effects is active, which in turn holds a higher probability than if neither of the effects is significant. This concept was formalized by Hamada and Wu (1992).

The HPM structure uses a General Linear Model (GLM) to express the system as a linear combination of input variables, including their interactions and higher order effects. An interpretation of the HPM by Chipman et al. (1997) is expressed in the equations 1 through 9.

$$y(x_1, x_2, ..., x_n) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \beta_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \sum_{\substack{k=1 \\ k>j}}^{n} \beta_{ij} x_i x_j x_k + \varepsilon \quad (1)$$

$$x_i \in \{+1, -1\} \; i \in 1...n \quad (2)$$

$$\varepsilon \sim NID(0, \sigma_\varepsilon^2) \quad (3)$$

$$\Pr(\delta_i = 1) = p \quad (4)$$

$$\Pr(\delta_{ij} = 1 \mid \delta_i, \delta_j) = \begin{cases} P_{00} & \text{if } \delta_i + \delta_j = 0 \\ P_{01} & \text{if } \delta_i + \delta_j = 1 \\ P_{11} & \text{if } \delta_i + \delta_j = 2 \end{cases} \quad (5)$$

$$\Pr(\delta_{ij} = 1 \mid \delta_i, \delta_j) = \begin{cases} P_{000} & \text{if } \delta_i + \delta_j + \delta_k = 0 \\ P_{001} & \text{if } \delta_i + \delta_j + \delta_k = 1 \\ P_{011} & \text{if } \delta_i + \delta_j + \delta_k = 2 \\ P_{111} & \text{if } \delta_i + \delta_j + \delta_k = 3 \end{cases} \quad (6)$$

$$f(\beta_i \mid \delta_i) = \begin{cases} N(0,1) & \text{if } \delta_i = 0 \\ N(0, C^2) & \text{if } \delta_i = 1 \end{cases} \quad (7)$$

$$f(\beta_{ij} \mid \delta_{ij}) = \frac{1}{s_1} \begin{cases} N(0,1) & \text{if } \delta_{ij} = 0 \\ N(0, C^2) & \text{if } \delta_{ij} = 1 \end{cases} \quad (8)$$

$$f(\beta_{ijk} \mid \delta_{ijk}) = \frac{1}{s_2} \begin{cases} N(0,1) & \text{if } \delta_{ijk} = 0 \\ N(0, C^2) & \text{if } \delta_{ijk} = 1 \end{cases} \quad (9)$$

The formulation above provides us with a structure that is capable of expressing an engineering system, with its expected regularities, in a mathematical form. However, we still need to quantify the various parameters (such as s_1, s_2, p, p_{00}, p_{111}, etc). A study by Li, Sudarsanam, and Frey (2006) executes a meta-analysis of 113 data sets from published factorial experiments. The study which focused on quantifying the above

mentioned regularities provides us with actual values for the parameters of interest. The values shown in table 2.2 summarize the results that will be used in this study.

Parameters	P	P11	P01	P00	P111	P011	P001	P000	σ_ε^2	s1	s2	C
	41%	33%	4.5%	0.48%	15%	6.7%	3.5%	1.2%	0-10	3.6	7.3	10

Table 2.2 Quantifying the HPM - Results from Li et al., (2006)

The use of the HPM to evaluate experimentation techniques was proposed by Frey and Li (2008). In that study, multiple Robust parameter design methods were compared. Mainly, the performance of two planning methods (crossed arrays versus combined arrays), and two methods of analysis (classical analysis versus interaction modeling) were evaluated. In this paper we adopt a similar process of using the HPM to evaluate our experimentation algorithms. The proposed methodology has four key steps:

(i) Instantiate multiple response surfaces

(ii) Simulate each Experimental algorithm on the responses

(iii) Evaluate the primary variable indicating the overall performance of the algorithm

(iv) Analyze and visualize the data

Each of these steps is briefly explained below (a more in depth discussion is available in Frey and Li (2008)):

(i) Instantiate Multiple Response Surfaces:

In this step a response surface is instantiated by creating a polynomial equation described by the GLM, shown in equation 1. Using a pseudo-random number generator we can create random variates in accordance to the model described in equations 7, 8 and 9 and therefore indirectly 4, 5 and 6. This will provide us with the coefficients (the $\beta's$) in the GLM required to create the response surface. The observations are subject to different degrees of experimental error, which is also subject to pseudo-random number generator in accordance to equation 3. In this study we propose to analyze our experimental algorithms at different degrees of experimental error, and hence we systematically change σ_ε, and perform a separate analysis for each scenario.

(ii) Simulate each Experimental algorithm on the responses:

The experimental strategies analyzed in this paper (as explained in the previous section) are coded as an algorithm that is executed by a computer. We describe this process as a simulation since it is carried out multiple times. Each simulation run provides a different response surface and therefore the performance of each experimental algorithm will change from run-to-run. In addition, we need to understand that even for a fixed response surface it is important to execute multiple simulation runs. This is because the experimental error is modeled as a random variable, and it assumes different values for each experiment. Hence the same algorithm, on the same response surface, at the same *mean* error level can lead to different results.

(iii) Evaluate the primary variable indicating the overall performance of the algorithm:

The goal of the experimental methods under scrutiny is to find the ideal design setting (treatment combination) that produces the best output (highest error-free response in a larger-the-better case). Hence, we report the algorithms performance as a ratio between the increase in output value achieved from the algorithms recommendation over the maximum attainable increase in output. This is expressed as a percentage. Equation 10 characterizes this for a larger-the-better scenario:

$$\text{Performance} = \frac{\text{Response from algorithm's recommendation - mean response}}{\text{maximum possible response value - mean response}}.100 \qquad (10)$$

It is important to note that while the experimental algorithms are tested on data with varying levels of experimental error, the evaluation of the algorithm's performance is carried out at its theoretical, error-free value. This will not be possible during actual field testing. It is possible only because we are running a computer experiment. Since we are simulating the data and adding stochasticity later on through experimental error, we can have an objective conception of performance.

(iv) Analyze and visualize the data:

The results from the simulations are gathered, analyzed and graphically plotted. To report the performance of the two experimental strategies we look at the mean performance across all simulations of that strategy. This is reported at different levels of experimental error. The main reason for doing this is to show that an experimental strategy's performance is greatly influenced by the experimental error, and it is important to define the range within which our findings are valid. Since the computer simulations are run a large number of times (10,000), even small differences in performance are statistically significant.

2.3 Example with a Single Response Equation

In this section we take a random, sample equation that was generated from the HPM and apply both algorithms to the response. It is hoped that this will concretely demonstrate the procedure described in this chapter. Since this is conducted strictly for illustrative purposes, only a low-resource comparison will be conducted. And, while both methods will be carried out on the response, only the proposed method will be explained and outlined in detail.

The response equation with only its significant terms is shown in equation 11.

$$y(x_A, x_B, x_C, x_D, x_E, x_F, x_G) = 11.2x_C - 1.5x_E + 19.5x_G + 2.4x_C x_E + 0.4x_C x_G - 0.4x_A x_C + 0.4x_A x_B x_G + \varepsilon \quad (11)$$

$$x_i \in \{+1, -1\}, i \in A...G$$
$$\varepsilon \sim NID(0, 10^2)$$

This response can be summarized as being typical representation of the HPM. With seven factors each having a 41% chance of being significant, the expected number of significant main effects is approximately 3. Similarly using the expected value calculations we find that there are no extreme anomalies with respect to the number of two-way and three-way interactions. It is important to note however that as the number of input variables in the experiment increase (main effects), the corresponding number of two-way and three-way interactions increase in a combinatorial fashion. Other regularities such as sparsity, hierarchy and heredity seem to be captured to a reasonable extent with most deviations from the expected value within the bounds of the expected variability. The standard deviation of the experimental error is set at 10 units. This value is seen as a medium value of error and is equal to size of the main effects as indicated by the value of 'C' from equations 7 through 9 and table 2.2.

In order to apply both experimental algorithms a full factorial design was generated using the response equation from 11. The full array is presented in Appendix A. The highest response, of 32.8 units is obtained with the following treatment: A = -1, B = -1, C = +1, E = +1, G = +1. Since this equation is not a function of x_D or x_F, only the remaining

variables contribute to the response. Since there is no constant term, and this is a balanced design, the mean response value is zero. A random simulation of the response with experimental error is computed and this will be used as the output data for the two experimental algorithms. The output with experimental error is also presented in Appendix A.

2.3.1 Ensembles of *a*OFAT

Four *a*OFAT algorithms are executed on the output data presented in Appendix A. The starting points for *a*OFATs are different and in accordance to the criteria of maximizing the Euclidean distance (also shown in table 2.1). The result of performing such an analysis is captured in tables 2.3 through 2.6. The best solution reached by each *a*OFAT is highlighted.

A	B	C	D	E	F	G	Output
-1	-1	-1	-1	-1	-1	-1	-15.91
+1	-1	-1	-1	-1	-1	-1	-29.13
-1	+1	-1	-1	-1	-1	-1	-15.31
-1	+1	+1	-1	-1	-1	-1	-25.19
-1	+1	-1	+1	-1	-1	-1	-33.73
-1	+1	-1	-1	+1	-1	-1	-47.12
-1	+1	-1	-1	-1	+1	-1	-0.66
-1	**+1**	**-1**	**-1**	**-1**	**+1**	**+1**	**6.87**

Table 2.3 *a*OFAT-1: an adaptive one-factor-at-a-time analysis on data from equation 11

A	B	C	D	E	F	G	Output
+1	+1	+1	+1	+1	+1	+1	19.95
-1	+1	+1	+1	+1	+1	+1	19.08
+1	**-1**	**+1**	**+1**	**+1**	**+1**	**+1**	**40.66**
+1	-1	-1	+1	+1	+1	+1	4.30
+1	-1	+1	-1	+1	+1	+1	32.80
+1	-1	+1	+1	-1	+1	+1	32.93
+1	-1	+1	+1	+1	-1	+1	23.98
+1	-1	+1	+1	+1	+1	-1	-6.08

Table 2.4 *a*OFAT-2: an adaptive one-factor-at-a-time analysis on data from equation 11

A	B	C	D	E	F	G	Output
-1	+1	-1	+1	-1	+1	-1	-28.05
+1	+1	-1	+1	-1	+1	-1	-43.38
-1	-1	-1	+1	-1	+1	-1	-48.34
-1	+1	+1	+1	-1	+1	-1	-19.92
-1	+1	+1	-1	-1	+1	-1	2.28
-1	+1	+1	-1	+1	+1	-1	-7.70
-1	+1	+1	-1	-1	-1	-1	-25.19
-1	**+1**	**+1**	**-1**	**-1**	**+1**	**+1**	**39.41**

Table 2.5 *a*OFAT-3: an adaptive one-factor-at-a-time analysis on data from equation 11

A	B	C	D	E	F	G	Output
+1	-1	+1	-1	+1	-1	+1	22.38
-1	-1	+1	-1	+1	-1	+1	39.60
-1	**+1**	**+1**	**-1**	**+1**	**-1**	**+1**	**49.67**
-1	+1	-1	-1	+1	-1	+1	9.45
-1	+1	+1	+1	+1	-1	+1	21.48
-1	+1	+1	-1	-1	-1	+1	32.74
-1	+1	+1	-1	+1	+1	+1	34.85
-1	+1	+1	-1	+1	-1	-1	-5.54

Table 2.6 *a*OFAT-4: An adaptive one-factor-at-a-time analysis on data from equation 11

It is noteworthy that none of the *a*OFATs above discovered the treatments that could provide the highest response. The findings from each *a*OFAT are now aggregated as using weighted voting method as described in section 2.1.1, and shown in table 2.7.

	A	B	C	D	E	F	G	Response	Rank	Voting Weights
*a*OFAT 1	-1	+1	-1	-1	-1	+1	+1	6.87	4	1
*a*OFAT 2	+1	-1	+1	+1	+1	+1	+1	40.66	2	3
*a*OFAT 3	-1	+1	+1	-1	-1	+1	+1	39.41	3	2
*a*OFAT 4	-1	+1	+1	-1	+1	-1	+1	49.67	1	4
Aggregated	-1	+1	+1	-1	+1	+1	+1			

Table 2.7 Ensemble *a*OFAT aggregation process

The treatment combination recommended by the final aggregated process after a weighted voting is shown in table 2.7. This treatment results in a response (without error) of 32.0. This is not the best treatment, and therefore not the highest response. The percentage improvement, as characterized by equation 10, is 97.6%

2.3.2 Fractional Factorial Design

A Resolution IV, 2^{7-2} design which consumes 32 resources is applied on the data. Four main fractions of the full factorial design that meet the criteria of maximum resolution (Box and Hunter, 1961) can be selected. The selection of the fraction is done randomly. A fraction with the aliasing pattern of F = ABCD, and G = ABDE, was selected. This gives us the following defining words I = CEFG = ABCDF = ABDEG. The entire aliasing structure is presented in Appendix B. This fractional factorial array allows us to estimate all main effects, most of the two-way interactions (18 out of the 21), and a few three-way interactions (6 out of 35). Owing to this aliasing pattern three two-way interactions cannot be estimated and the designer is free to select these. However, the designer might be unaware of which ones are truly significant, or not. In our response equation three two-way interactions, are significant and we would ideally like to have them in the model. Similarly, the same problem applies in estimating three-way interactions. In our model we randomly selected 18 of the 21 two-way interactions and 6 of the 35 three-way's. All the two-ways in the response (equation 11) were selected, but

the three-way was not. It is noteworthy that in a simulation of many random trials there is a random selection of fractions as well as interaction terms selected within the possible set of aliased terms. This would lead to the selection of interaction terms that are in the system by chance sometimes and fail to select them at others.

A stepwise regression analysis (as described in section 2.1.2) is performed on the data with experimental error (as shown in Appendix A). The resulting regression equation is shown below:

$$y = 13.1x_C + 19x_G - 0.2x_B + 1.1x_Cx_E - 0.2x_Dx_E + 0.1x_Ax_C \quad (12)$$

An optimization of this equation results in selecting the following settings for the variables A = +1, B = -1, C = +1, D = -1, E = +1, G = +1. The factor F is not in this equation or the original response surface and therefore does not influence the outcome. This treatment combination resulted in a response of 31.2 units. This is 95.1% of the maximum possible response of 32.8 units.

2.4 Results and Discussion

The results from applying the experimental algorithms on the HPM will be discussed in this section. The goal of the experimental methods under scrutiny is to find the ideal design setting (treatment combination). In the HPM simulations we assume a larger-the-better scenario and seek to find the treatment that provides the largest response. The response surfaces in these simulations vary with each run and are generally characterized by equations 1 through 9. Experimental error is varied from 1 to 20 units, and the expected value of the largest main effect is approximately 16 units (since each simulation creates different, independent values for the coefficients). The standard step sizes for the experimental error is 2.5 units. A total number of 10,000 simulation runs are executed at each error level. The aim of the experimental algorithms is to find a treatment combination that maximizes the response surface. The metric that measures the performance of these procedures is shown in equation 10.

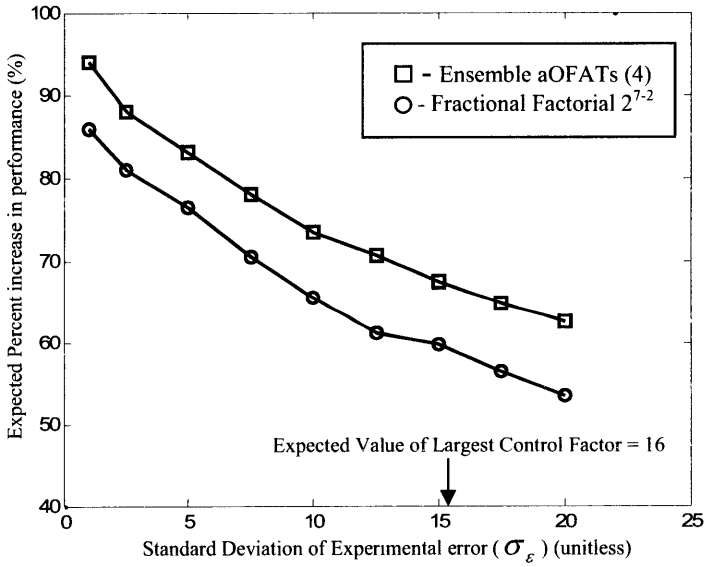

Figure 2.1 Comparing an ensemble of 4 *a*OFATs with a 2^{7-2} fractional factorial array using the HPM

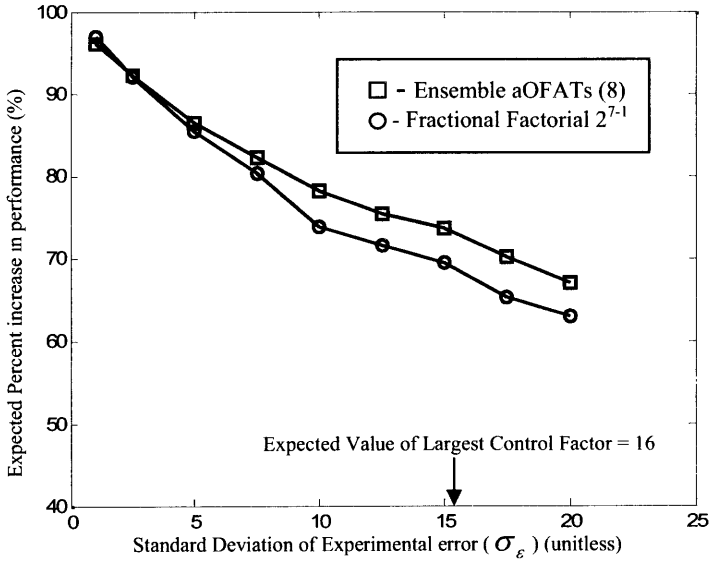

Figure 2.2 Comparing an ensemble of 8 *a*OFATs with a 2^{7-1} fractional factorial array using the HPM

The results from the simulations are depicted in figures 2.1 and 2.2. These findings suggest that at best the proposed algorithm is in general superior to the fractional factorial designs and at worst they provide equal improvements.

In the low-resource intensive comparison (Ensemble of 4 *a*OFATs versus a 2^{7-2} Fractional Factorial design) the ensemble *a*OFAT algorithm provides superior or identical design recommendations in 66% of the cases at low error levels (2.5 units), 83% at medium error levels (10 units), and 85% at high error levels(17.5 units). In the high-resource intensive comparison (Ensemble of 8 *a*OFATs versus a 2^{7-2} Fractional Factorial design) the ensemble *a*OFAT algorithm provides superior or identical design recommendations in 59% of the cases at low error levels (2.5 units), 74% at medium error levels (10 units), and 81% at high error levels(17.5 units).

3 Case Study: Improving the Performance of a Wet Clutch

In this chapter we introduce an illustrative case study which shows the application of our experimental techniques to a specific engineering system. This case is adapted from a study performed by Fredrick A. Lloyd (Lloyd, 1974) which looks at the use of design of experiments to identify the parameters contributing to power loss in a disengaged wet clutch. This is a typical case which has been used in other studies in experimental design as an illustrative example. Both Frey and Li (2008), and Frey and Jugulum (2006) utilize this experiment as a sample test bed to illustrate the point of their respective papers. Li et al. (2006) use this experiment as one of their 113 cases to study regularities in experimental environments.

3.1 Background

A clutch, such as the one shown in figure 3.1, is a device used to transmit torque from an input shaft, which is normally connected to a motor or an engine, to an output. This is accomplished through the use of two or more friction discs that are pressed together. The clutch can be engaged or disengaged. In the disengaged position, the friction discs are separated by a very small clearance, and expected to spin at different speeds. In the wet clutch, the entire setup is immersed in a cooling lubricating fluid. This provides multiple advantages over a dry clutch, such as smoother performance, longer life and surfaces are usually cleaner. However, there are some disadvantages, such as the energy lost to the fluid, primarily through the viscous shear forces seen during operation. In this study, the power loss is measured by studying the drag torque resulting from these shear forces transmitted through the fluid.

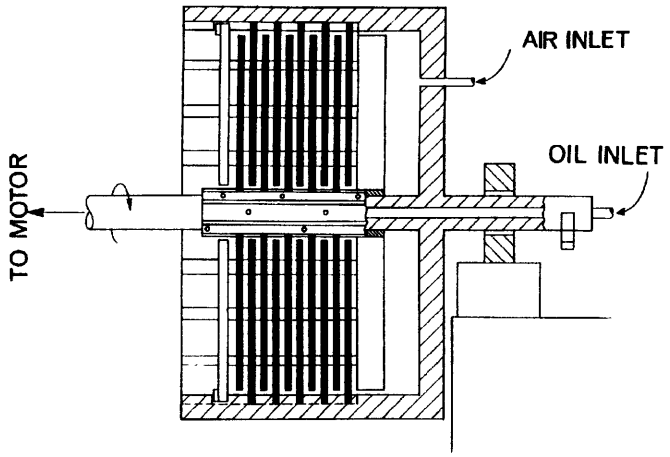

Figure 3.1 A wet clutch pack (adapted from Lloyd (1974))

The friction assemblies were preburnished with the reaction plates on an inertia dynamometer in order to reproduce a more realistic use of a clutch pack that has seen some service. Drag torque, the response of interest, was measured from the inertia dynamometer. Seven input variables were selected for this experiment. Each variable is tested at two levels. The seven variables with their actual settings are shown in table 3.1

Variable	Factor Name	Low Level (-1)	High Level (+1)
A	Oil Flow, gal/min	0.75	2.00
B	Pack Clearance, in	0.050	0.150
C	Spacer Flatness	Flat	Waved
D	Grooving Patterns	No Groove	RM No. 17 Waffle
E	Oil Viscosity	SAE10W	SAE40W
F	Friction Material	R-4893-1	W-1349
G	Rotational Velocity	1000	3000

Table 3.1 Input variables and their level settings from the wet clutch case

A full factorial 2^7 experimental plan was carried out and the results analyzed by Lloyd (1974).

3.2 Model Construction

For our purposes, we analyze the full factorial data using a model and procedure similar to that performed in the original paper. The raw data is available in Lloyd (1974). The analysis of the experiment is carried out using stepwise regression. This allows us to quantifying all the main effects, two-way and three-way interactions. This procedure also allows us to assess which effects are statistically significant (at the 99% level). In our problem, we have included all main effects, and only the significant two-way and three way effects in the GLM. This is in accordance to Chipman's 'Strong Heredity' principle (Chipman; 1996), which states that all parent main effects be compulsorily included if an interaction term which contains this effect is found to be statistically significant. Table 3.2 shows the values of the various main effects and Table 3.3 shows the significant two way interactions, and Table 3.4 shows the significant three-way interactions. Our findings, in terms of effect significance and magnitude, is very similar to that of Lloyd (1974) and previous studies by our group on the same dataset (Frey and Jugulum; 2006).

Effect	Drag torque [ft lbs]
A	1.33
B	-1.55
C	-1.81
D	0.07
E	2.81
F	-0.09
G	3.01

Table 3.2 Main effects for the wet clutch case

Effect	Drag torque [ft lbs]
AD	0.53
AG	0.96
BD	-0.52
BG	-0.83
CD	0.68
CG	-0.49
DE	0.51
DG	0.83
EG	0.46

Table 3.3 Significant two-way interaction effects for the wet clutch case

Effect	Drag torque [ft lbs]
BCD	0.43
CDE	-0.49
CDF	0.51
CDG	-0.83
CEF	0.46

Table 3.4 Significant three-way interaction effects for the wet clutch case

This analysis allows us to create the following response equation which is seen as a predictive fit for the experimental data:

Drag torque =

$$\begin{aligned} &4.92+0.66x_A-0.77x_B-0.90x_C+0.03x_D+1.40x_E-0.05x_F-1.50x_G \\ &+0.26x_{AD}-0.48x_{AG}-0.26x_{BD}+0.41x_{BG}+0.34x_{CD}+0.35x_{CG}+0.32x_{DE}+0.46x_{DG}-0.66x_{EG} \\ &+0.21x_{BCD}-0.25x_{CDE}+0.25x_{CDF}-0.42x_{CDG}+0.23x_{CEF} \end{aligned} \quad (13)$$

Where, drag torque is measured in ft-lbs and x_A through x_G are the input factors in coded Hi – Lo settings of +1 and -1. Equation 13 is a response surface description of the physical phenomenon associated with a wet clutches' workings. We will use this description to study the performance of our experimentation strategies, where the goal of both strategies is to minimize the drag torque.

Similar to the analysis seen in chapter 2 with the HPM, simulations of the experimental algorithms on the response surface (equation 13) is conducted. The response in this case is a single fixed surface, as opposed to the HPM where the response is characterized by a stochastic formulation and each simulation run presents a different response equation. Nevertheless, we conduct simulations on this single surface because the experimental error is a source of randomness which leads to different findings in each simulation run.

3.3 Results and Discussion

The response surface used for this case is shown in equation (13). The coded settings of the input variables that correspond to a minimization of drag torque are, A = -1, B = -1, C = +1, D = +1, E = +1, F = -1 and G = +1. These settings lead to an expected drag torque of 0.304 lb-ft. The average drag torque is 4.93 lb-ft (this is the same as β_0), and maximum drag torque is 12.52 lb-ft.

Similar to equation 10 that was used for the HPM simulations, our metric is the decrease in drag torque, expressed as a percentage of the maximum possible decrease. This is defined through equation 14.

$$\text{Percent decrease in drag torque} = \frac{\text{mean drag torque - drag torque of method}}{\text{mean drag torque - minimum possible drag torque}}.100 \quad (14)$$

The 'drag torque of method' is computed by looking at the error-free, theoretical drag torque that will be achieved when a method's recommended treatment combination is applied to the response surface. The mean and minimum possible drag torques are constants.

We analyze the performance of the two strategies across a reasonable range of experimental errors. The value of σ_ε is varied from 0.5 lb-ft to 5 lb-ft, with step sizes of 0.5 lb-ft. It is important to look at this range in the context of the magnitude of the largest main effect which is 3.01 lb-ft and the square root of the average residual sum of squares (RSS) is 0.877 lb-ft, which is found by comparing the full-factorial experimental data reported in the original study with the fitted GLM model. The performance of the different algorithms is determined by simulating 10,000 trials at each level of σ_ε, and averaging the results to find the expected percent decrease in drag torque.

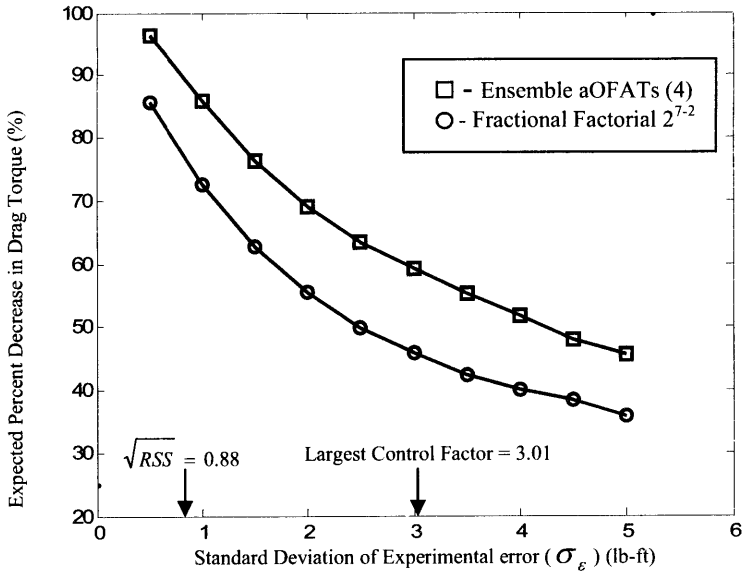

Figure 3.2 Comparing an ensemble of 4 *a*OFATs with a 2^{7-2} fractional factorial array for wet clutch case study

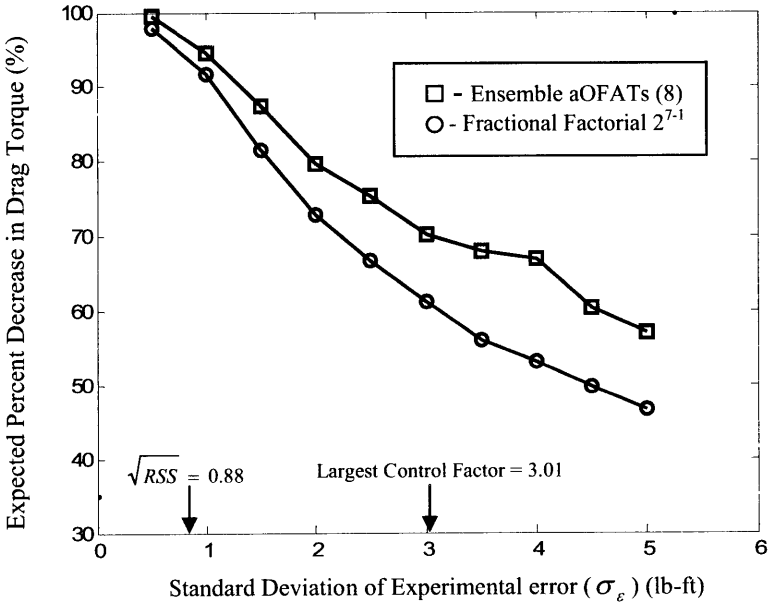

Figure 3.3 Comparing an ensemble of 8 *a*OFATs with a 2^{7-1} fractional factorial array for wet clutch case study

The results are graphically illustrated in figures 3.2 and 3.3. In both cases the performance of the ensemble *a*OFATs are superior to that of the fractional factorial arrays, consistently for the various error levels.

At low error levels (error = 1lb-ft), in the low-resource intensive comparison (Ensemble of 4 *a*OFATs versus a 2^{7-2} Fractional Factorial design), the ensemble *a*OFAT algorithm identifies the best treatment combination (the treatment combination that minimizes drag torque) in 99% of the simulation runs. The fractional factorial design identifies this treatment combination only in 97% of the cases. At this error level, the proposed method provides an 99% expected decrease in drag torque, while fractional factorial provides a 98% decrease. At medium error levels (error = 3 lb-ft), in this same comparison, the proposed algorithm identifies the best treatment combination in 48% of the cases, while factorial arrays identify them 40% of the time. The expected decreases in drag torques are 59% and 46%, respectively. At high error levels (error = 5 lb-ft), in this same comparison, the proposed algorithm identifies the best treatment combination in 31% of the cases, while factorial arrays identify them 30% of the time. The expected decreases in drag torques are 46% and 36%, respectively.

At low error levels (error = 1lb-ft), in the high-resource intensive comparison (Ensemble of 8 *a*OFATs versus a 2^{7-1} Fractional Factorial design), the ensemble *a*OFAT algorithm identifies the best treatment combination (the treatment combination that minimizes drag torque) in 83% of the simulation runs. The fractional factorial design identifies this treatment combination only in 61% of the cases. At this error level, the proposed method provides an 86% expected decrease in drag torque, while fractional factorial provides a 72% decrease. At medium error levels (error = 3 lb-ft), in this same comparison, the proposed algorithm identifies the best treatment combination in 65% of the cases, while factorial arrays identify them 55% of the times. The expected decreases in drag torques are 70% and 61%, respectively. At high error levels (error = 5 lb-ft), in this same comparison, both algorithms identify the best treatment combination in 41% of the cases. The expected decreases in drag torques are 57% and 47%, respectively.

4 Comparing an Ensemble of *a*OFATs with an Ensemble of Factorial Designs

In this chapter we test the secondary hypothesis in this thesis that, *the use of ensemble techniques on aOFAT will outperform the use of ensemble techniques on multiple, highly-fractioned orthogonal designs (each being as resource intensive as a single aOFAT).*

4.1 HPM Comparison and Results

Similar to the methodological approach adopted in Chapter 2, we use the HPM to make a comparison between an ensemble of *a*OFATs and an ensemble of Resolution III fractional factorial designs. Again, an experimental setup with seven factors is considered where each factor has two levels each. With this setup, both, the *a*OFAT and 2^{7-4} fractional factorial plan consumes 8 resources. The low-resource version of using four *a*OFATs will be compared with four of the factorial designs. Both sets of plans consume 32 resources each.

The 2^{7-4}, resolution III array is a saturated design where the main effects are aliased with all two-way interactions, thus making it impossible to estimate the two-way interaction effects. The selection of the four arrays for each experimental plan is performed as a two step process. Initially, one of the four 2^{7-2} resolution IV designs that meet the maximum resolution criteria (Box and Hunter, 1961) is randomly selected. Each of these 32 run designs is split into four 2^{7-4} resolution III designs. Essentially, there are 16 arrays of size 2^{7-4} that meet the requirements of maximum resolution, and four of them are subsets of any of the selected 2^{7-2} design. These four arrays will be used to analyze the design space and subsequently aggregated. By performing this two-step process we can guarantee that each of the four fractions follow the guidelines for statistical propriety, but also that they are relevant in the context of each other.

Stepwise regression is not used to analyze each Resolution III design. This is mostly because there are overall 64 effects to be estimated (1 constant + 7 main effects + 21 two-

way interactions + 35 three-way interactions) from 8 experimental runs. In most simulation runs, this is not a sufficient amount of data to construct a reasonable model. Especially, one that can be optimized it to provide a single best treatment. An approach that aggregates coefficients and significance levels from the different arrays in some statistically sound manner would be akin to analyzing the larger fraction (2^{7-2}) in the first place. In this thesis we use two simpler strategies to analyze the factorial arrays.

(i) Classical Analysis: This procedure is a conceptually simple way of analyzing factorial designs. Here, each factor is analyzed separately. The average response at the different levels of the factor is compared and the best one is selected. This is repeated for all factors, sequentially. This method is popularly used in the context of robust design when crossed arrays are used with highly fractioned inner arrays. It was popularized by Genichi Taguchi (Phadke, 1989). This technique is likely to work well in environments with high experimental error. The reason being that, the use of multiple experimental values to compute the average effect is beneficial. The variance of the estimate is lower than individual values since it is the average of many independent random variables. The decrease in variance scales linearly with an increase in the number of trials. However, this technique is likely to perform poorly in environments with a lot of interactions. The procedure ignores interaction effects. While an interaction term in the system will indirectly contribute towards a main effect through this analysis, it is likely to be weighed down by the other experimental values of the same main effect which possess a different interaction structure.

(ii) Take-the-best: Is a fairly straightforward technique which picks the treatment combination with the highest output observed in the experiment. This method is more susceptible to experimental error, since it just takes one erroneously represented outcome to mislead the entire analysis. As a result, this analysis is likely to perform well at regions of low experimental error and poorly in regions of high error. Take-the-best is fully capable of exploiting interactions, but fails to account for treatment combinations that aren't available in the selected fraction of the full factorial space.

Due to some of the limitations mentioned above, neither of these techniques is effective enough to be recommended as a standalone procedure to analyze factorial designs. One of the main reason for this is that they do not perform consistently well across all regions of experimental error. However, the take-the-best is an effective strategy when experimental error is really low and classical analysis is effective when the experimental error is really high. By focusing the comparison of our algorithm to take-the-best at low errors, and classical analysis at high errors, we would be providing the ensemble of factorial designs a substantial advantage. This is because, in the real world experimental environment, the value of pure error is seldom known and can be hard to estimate.

The rank sum voting technique is used to aggregate the factorial designs. This is the same technique that is used to aggregate the *a*OFATs (introduced in section 2.1.1). Here, each agent, in this case 2^{7-4} fraction factorial array, votes on the levels of each factor. Their votes are weighted by ranking system. The factorial designs that find higher responses get higher ranks.

Besides comparing the ensemble of four *a*OFATs, with two sets of four 2^{7-4} factorial designs (one using classical analysis and one using take-the-best), we also compare this with the original 2^{7-2} design with stepwise regression. All the designs use 32 resources. The methods compared using HPM simulations are shown in figure 4.1.

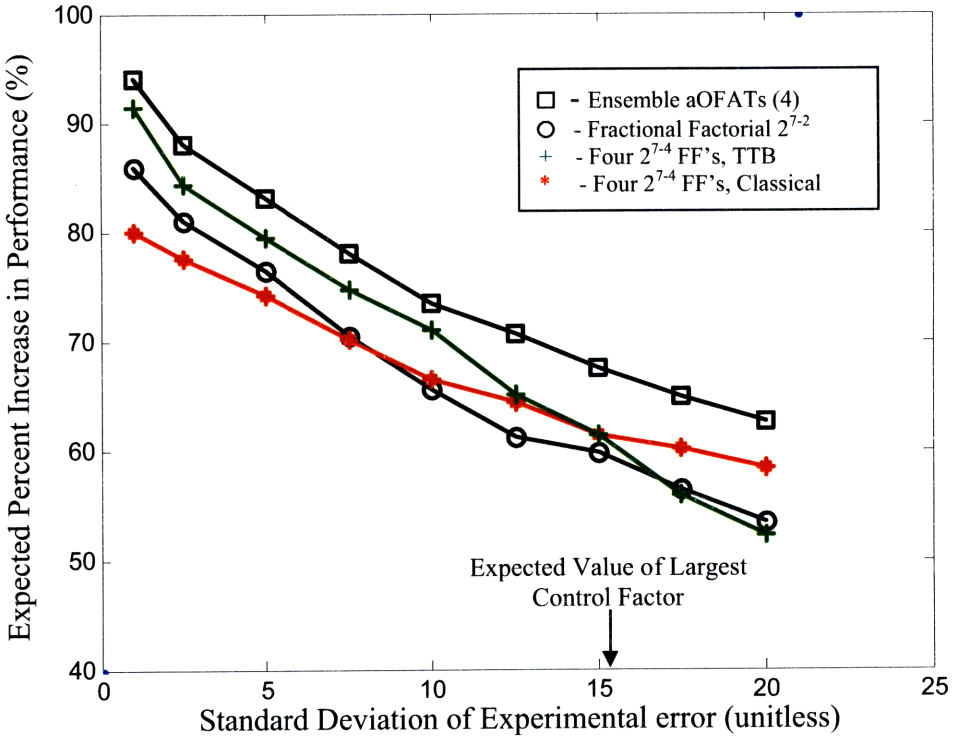

Figure 4.1 Comparing an Ensemble of four aOFATs with an ensemble of four 2^{7-4} Fractional Factorial arrays using classical analysis and take-the-best, and a single 2^{7-2} Fractional Factorial array

The key finding in this study is that, the ensemble of *a*OFATs performed consistently better than an ensemble of fractional factorial designs, through all ranges of experimental error, and irrespective of the strategy used to analyze the factorial design. Of specific interest is the fact that an ensemble of *a*OFATs outperformed an ensemble of fractional factorials when they were analyzed using a take-the-best strategy at regions of low experimental error. And, an ensemble of *a*OFATs outperformed an ensemble of fractional factorials when they were analyzed using a classical analysis at regions of high experimental error. Unrelated to the hypothesis of this study, it was interesting to observe that an ensemble of frugal fractions (2^{7-4}'s) using heuristic analysis strategies (like classical analysis, take-the-best, and voted aggregations) performed comparably to a single 2^{7-2} factorial array which used stepwise regression with an adaptive α value.

4.2 Theoretical Discussion: Why Ensemble Methods to aOFAT?

The findings in section 4.1 suggest that the superior performance of the proposed algorithm is not entirely due to the effect of ensembles (since factorial designs that are aggregated do not perform as well). The performance is also not entirely due to *a*OFATs design. This is because, previous studies, such as Frey et al.,(2003), Frey and Wang (2006) and Frey and Sudarsanam (2008), show that *a*OFATs when compared to factorial designs do not perform so well in regions of high experimental error (something not seen with the algorithm proposed in this thesis). This section argues the interactive effect *a*OFATs and ensemble techniques that make this algorithm work well.

The aggregation of many individual classifiers, or predictors, results in an ensemble classifier. Hence, the strength of the ensemble classifier depends on the predictive capabilities of the individual predictors and the interrelations between them (Amit and Gelman; (1997), and Breiman (2001a)). The Predictive strength of an algorithm is often represented as a percentage or ratio of perfect predictability, the complement of which is prediction error. This error can be split into two components - 'variance' and 'bias' (Bishop; 1995, Hastie, et al.,; 2001). Such a dichotomy is very useful in understanding exactly how the predictive strength of the building blocks scale up to that of the ensemble classifier's.

Using the following derivation from Hastie et al., (2001):

If we take an environment, $Y = f(X) + \varepsilon$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$,

We can derive an expression for the expected prediction error of a fitted model $\hat{f}(X)$ at a point $X = x_0$ using squared error loss:

$$ERR(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0] \qquad (15)$$

$$ERR(x_0) = \sigma_\varepsilon^2 + [E(\hat{f}(x_0)) - f(x_0)]^2 + E[(\hat{f}(x_0) - E(\hat{f}(x_0)))^2] \quad (16)$$

$$ERR(x_0) = \text{Irreducible error} + \text{Bias}^2 + \text{Variance} \quad (17)$$

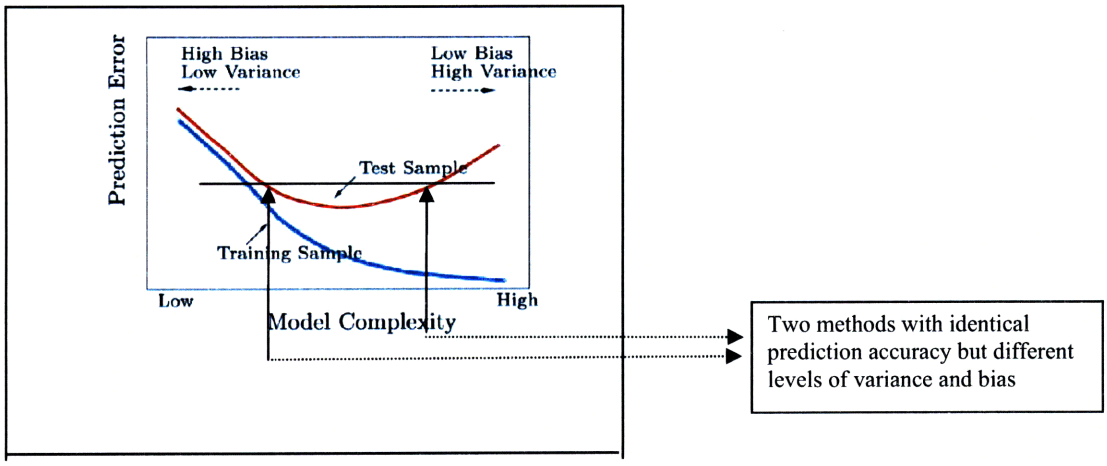

Figure 4.2 Bias and Variance Decomposition (adapted from Hastie, Tibshirani and Freidman 2001)

From equation 17 it can be inferred that it is possible for two individual prediction methods to have the same predictive strength but one could have high bias and low variance, while the other could have a low bias and high variance. Using a third variable called 'model complexity', Figure 4.2 (adapted from Hastie et al,; 2001) illustrates this. In such cases, is one prediction algorithm more suited for ensemble methods than the other?

This study, and many others in the literature, (for example, Keijzer and Babovic (2000), Buhlmann and Yu (2002), Friedman and Hall (2007)) assert that when two methods with the similar predictive accuracies are compared, the method with relatively lower bias and higher variance is better suited for ensemble techniques. The reasoning is that ensemble methods work by utilizing the law of large numbers to achieve variance reduction that results from the aggregation of multiple different predictors. In cases where the predictor used has low bias and high variance the use of ensemble methods prevents overfitting of the model to data. Conversely, if the individual predictor has a high bias and low

variance, then ensemble techniques don't contribute much to the predictability, because the bias component is not improved through the aggregation.

A conception similar to prediction error, bias, and variance can be applied to experimentation. A parallel notion of prediction error in experimentation is the loss incurred in selecting a treatment combination that is suboptimal. For a given system there exists one treatment combination (at zero error) that results in the best response value. An algorithm that identifies this treatment combination is similar to a data mining algorithm that makes a perfect prediction ($\hat{f}(x) = f(x)$).A lower response (in a larger-the-better case) resulting from an alternate treatment can be expressed as a fraction of the maximum possible value. We define this loss as 'Percent experimental loss' and quantify it as:

$$\text{Percent experimental loss} = (1 - \frac{response_{experiment}}{response_{optimal}}).100 \qquad (18)$$

Similar to the bias-variance dichotomy in data mining, the experimental loss can be split into two sources, error-dependent loss and error-independent loss. The experimental loss occurring due to experimental error (also referred to as pure error) is similar to the variance component in a prediction algorithm. We call this 'error dependent loss', since this usually increases as a function of increasing experimental error. And this increase is dependent on the algorithm used. The variance in a predictor also increases as a function of error. Analogous to the bias of a predictor is the 'error independent loss' of an experimental strategy. In an effort to save resources and militate against this experimental error various algorithms utilize statistical or heuristic strategies which prevent them from identifying the best treatment combination. This is the loss incurred even when there is no experimental error and hence can be quantified as the percent experimental loss when experimental error is zero. The bias in a prediction technique can similarly be ascertained by calculating the prediction error in $f(X)$ with zero pure error ($\sigma_\varepsilon = 0$).

Data Mining	Experimentation
Prediction Error	Percent Experimental Loss
Bias	Error independent Loss
Variance+ Irreducible error	Error dependent loss

Table 4.1 Conceptual analogues between data mining terminology and experimentation

The *a*OFAT algorithm when compared to resolution III fractional factorial arrays is a method with lower 'error independent loss' and higher 'error dependent loss'. Theoretical proof that *a*OFAT performs better than Resolution III arrays in environments with low pure error (including zero error) has been shown, while the opposite is true in environments with high experimental error (Frey and Wang; 2006). These findings have been corroborated through case studies by Frey and Sudarsanam (2008) in the context of robust design.

Figure 4.3 is an illustrative example from Frey and Wang (2006) that shows that at low levels of experimental error *a*OFAT outperforms resolution III factional factorial arrays (while the opposite is true at high error levels). This particular case looks at maximum flight time of an electric powered model aircraft. An *a*OFAT with seven design variables, each at two levels, is compared with a 2^{7-4} resolution III fractional factorial design.

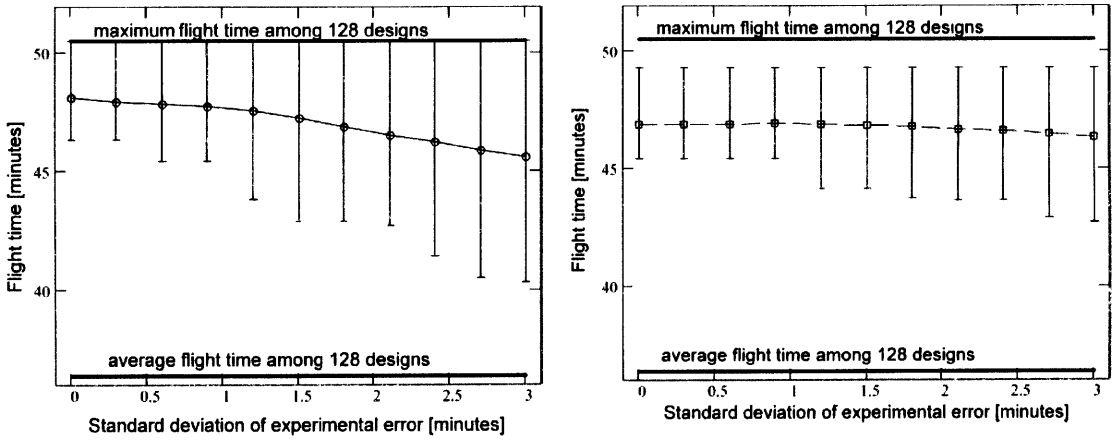

Figure 4.3 Flight times against increasing standard deviation of experimental error for the electric powered aircraft case when an *a*OFAT was applied (on the left), and a 2^{7-4} resolution III design was applied (on the right); with 10^{th} to 90^{th} percentiles (adapted from Frey and Wang, 2006)

A secondary line of reasoning to show that *a*OFAT when compared to resolution III fractional factorial arrays is a method with lower 'error independent loss' and higher 'error dependent loss', follows from looking at environmental complexity. This reasoning also uses the conceptual analogues that Bias corresponds to 'error independent loss' and that variance corresponds to 'error dependent loss'. Low bias and high variance is synonymous with model complexity (Hastie, et al.,; 2001). This is because such models are capable of taking up complex structures in order to fit data (which can be an advantageous or detrimental property). It therefore follows that predictors with low bias and high variance perform better than their counterparts in environments that are complex. Environmental complexity can be defined by complexity of the function $f(X)$ in $Y = f(X) + \varepsilon$. In the context of designed experiments where each input variable often only takes two levels, complexity is analogous to the average strength and presence of interaction terms (as higher order, logarithmic and trigonometric terms are not possible). Again, theoretical proof that *a*OFAT performs better than Resolution III arrays in environments where the ratio between strength of interaction terms to main effect terms is high has been demonstrated by Frey and Wang (2004) which has been corroborated through case studies by Frey and Sudarsanam (2008).

Similar to reasoning that the aggregation process in ensemble techniques targets the variance component in a predictor, the aggregation of experiments targets the error dependent loss in an experimental strategy. Hence, a technique with high error dependent loss and low error independent loss is better suited for ensemble experiments. This notion in conjunction with the findings of Frey and Wang (2006) and Frey and Sudarsanam (2008) motivates us to use an *a*OFAT strategy over Resolution III fractional factorial arrays in ensemble experiments.

5 Extending Ensemble Techniques and Factorial Design Analyses

This chapter looks at the effect of various algorithmic modifications to the proposed method. It also looks at the comparative performance of fractional factorial designs when they are analyzed using different strategies.

5.1 Methodological Extensions of the Proposed Algorithm

In an effort to seek greater improvement in performance, changes to certain aspects of the proposed algorithm are tested. Three key areas that impact the performance of an ensemble of *a*OFATs are: (i) Starting points of *a*OFATs, (ii) Order in which variables are changed, and (iii) Aggregation techniques used. In this section we will address each of these aspects and provide multiple strategic plans for them.

It is important to note that these three methodological areas cannot be looked at separately, but in the context of each other. A strong interaction exists between these three domains. For example, an aggregation strategy which places a lot of weight on the *a*OFATs with superior performances is better combined with a plan for having dissimilar starting points. On the contrary, an aggregation strategy that is more egalitarian (the findings of each *a*OFAT are weighted equally, irrespective of its outputs) works better when the starting points are closer to each other. Acknowledging this interactive nature between the three areas leads us to study them in combinations. The tests are conducted in accordingly.

5.1.1 Starting locations of *a*OFATs

The starting point of *a*OFATs, refers to the levels at which the factors are originally set, when the *a*OFAT is started. An *a*OFAT algorithm which starts closer to the peak of the response (more levels of the starting vector match the levels of the best treatment), has a

higher probability of finding the peak or getting close to it, than one which starts far away (Frey and Sudarsanam, 2008). In a scenario where there is no prior knowledge on the response peak, this insight does not help guide us in picking a starting location for a single *a*OFAT plan. However, when using an ensemble of *a*OFATs, in conjunction with an aggregation strategy, or variable change order, this information can be utilized in a beneficial way. A strategy which spreads starting locations is likely to have placed an *a*OFAT closer to the peak than one which starts them all at the same point. A strategy that maximizes the Euclidean distance between starting locations achieves this. Motivated by this line of reasoning, and a comparative analysis conducted earlier in this thesis (section 2.1.1), the experimental algorithm studied thus far in this thesis uses this strategy. However, earlier studies looked at the various possible starting location strategies with a fixed strategy for variable change order and aggregation. We revisit the different possibilities for starting points in light of looking at its combined effect with the other strategies which will also be changed. The three main ideas for starting locations are summarized below:

(i) Maximizing Euclidean distance: In this plan, the aim is to maximize the average Euclidean distance between pairs of points. It is important to note that our design space with discrete levels is characterized a hyper-cube. An algorithm that maximizes Euclidean distance will not place points that are equidistant from each other. Almost always this algorithm encourages points to find the corners of the cube, and therefore some pairs will be closer or further away than others.

(ii) Random Starting Locations: Inspired by random forest algorithms (Breiman, 2001a), and wisdom-of-crowd recommendations that emphasize the importance independence between agents (Surowiecki, 2004), this plan randomly picks starting locations for *a*OFATs.

(iii) Fixed starting locations: This plan picks a randomly selected starting location that is fixed for all the *a*OFATs in the ensemble.

5.1.2 Order of Variable Changes

The performance of an *a*OFAT design, being sequential in nature, is influenced by the order in which the variables are tested. Typically, the performance can be improved if the factors with the largest impact (on the response) are toggled first, followed by others in a descending order. The detailed explanation of this result is beyond the scope of this thesis. However, the basic idea contained in this effect concerns exploiting two-way interaction terms. When a variable with a large impact is changed towards the end of the experimental plan, it impacts the contribution of all the two-way interactions in which it is involved. Having different variable change orders is also seen as a way of creating dissimilarity between the *a*OFAT experiments. The main plans for variable change orders are summarized here:

(i) Maximize Negative Correlation: A plan of maximizing negative correlation between variable change orders is tested. Similar to the plan of maximizing Euclidean distance between starting points, this plan seeks to maximize the average negative correlation between pairs of *a*OFATs.

(ii) Random Change order: The second plan changes variables in a randomly different order for each *a*OFAT. It is motivated by the importance of presenting stochastic properties to a deterministic algorithm, as seen in ensemble techniques (Berk, 2006) and providing independence between experimental agents.

(iii) Fixed Change order: The third plan provides a randomly selected order, which is then fixed, and is held constant across all *a*OFATs. The experimental comparison conducted up to this point in this thesis only used this strategy.

5.1.3 Aggregation Strategies

The aggregation strategy used is critical to the performance of an Ensemble technique, and is similarly very important in this experimental algorithm. At a very simple level, aggregation techniques seen in Ensemble methods range from a very equal participation strategy as seen in 'Bagging', to a more a performance weighted approach such as 'Boosting' (Freund and Schapire, 1997). The aggregation strategies presented here attempt to capture this range. However, this is a simplification. More complex conceptions are possible in Ensemble techniques such as that seen with Hierarchical

Mixture of Experts and other Bayesian strategies (Jordan and Jacobs, 1994). Similarly, the possibilities for creating complex and effective aggregation strategies in experimentation are endless and would be a reasonable area of future work for this research. Four main algorithms are presented here for aggregation:

(i) Rank Sum point system: This plan has been used on the proposed algorithm in the previous chapters. This plan is weighted voting plan, where each *a*OFAT votes on a level for each factor, independently. The weights are determined according to the rank of the *a*OFATs in relation to each other and not a complex function of the response. The *a*OFAT with the worst performance is given a rank point of 1, the next 2, and so on. This is seen as a strategy that lends robustness to the experiment and is similar to Wilcoxon rank-sum test(Wilcoxon; 1945).

(ii) Response Weighted: Similar to the rank sum point system, each *aOFAT* votes on each factor's level settings based on its experimental findings. The weight to each vote however is a function of the best response discovered by each *a*OFAT. A square function is used in this case.

(iii) Equal Voting: Equal voting uses a similar strategy as the above two plans, but with the exception of using equal weights for all *a*OFATs irrespective of their performance. In this strategy and the Rank Sum point system, ties are broken by assuming squared error loss function. This usually means the more consistent *a*OFAT performers are given more preference in the case of a tie.

(iv) Take-the-best: This strategy does not really aggregate, it selects the recommendation of the *a*OFAT that discovered the best treatment (characterized by highest experimental response), and disregards the others. This exposes the method to erroneous conclusions due to experimental error seen during the best run. It is expected that this strategy will work well at regions of low experimental error.

5.1.4 Evaluating the various experimental algorithms

There are three plans proposed for starting points, three for variable change order, and four for aggregation techniques. The combinations of these three areas could therefore provide 36 experimental algorithms. However, not all 36 algorithms are presented here. A subset of nine algorithms with the best performers from preliminary tests, along with some representative samples of the others is presented in this thesis. The combinations of tests that will be evaluated are shown in table 5.1.

S.No.	Aggregation Technique	Starting Point Strategy	Variable Change order
1	Rank Sum	Fixed Location	Minimize Correlation
2	Response Weighted	Max. Euclidean Distance	Minimize Correlation
3	Equal Voting	Random	Random
4	Take-The-Best	Fixed Location	Same Order
5	Rank Sum	Max. Euclidean Distance	Same Order
6	Response Weighted	Random	Minimize Correlation
7	Equal Voting	Max. Euclidean Distance	Same Order
8	Take-The-Best	Max. Euclidean Distance	Minimize Correlation
9	Rank Sum	Max. Euclidean Distance	Random

Table 5.1Nine Ensemble *a*OFAT strategies tested in this thesis

Technique with serial number 5 from table 5.1 has been the default algorithm discussed in the previous chapters of the thesis. These nine algorithms are tested in using the HPM, on the low-resource version of aggregating 4 *a*OFATs. The results are graphically shown in figure 5.1.

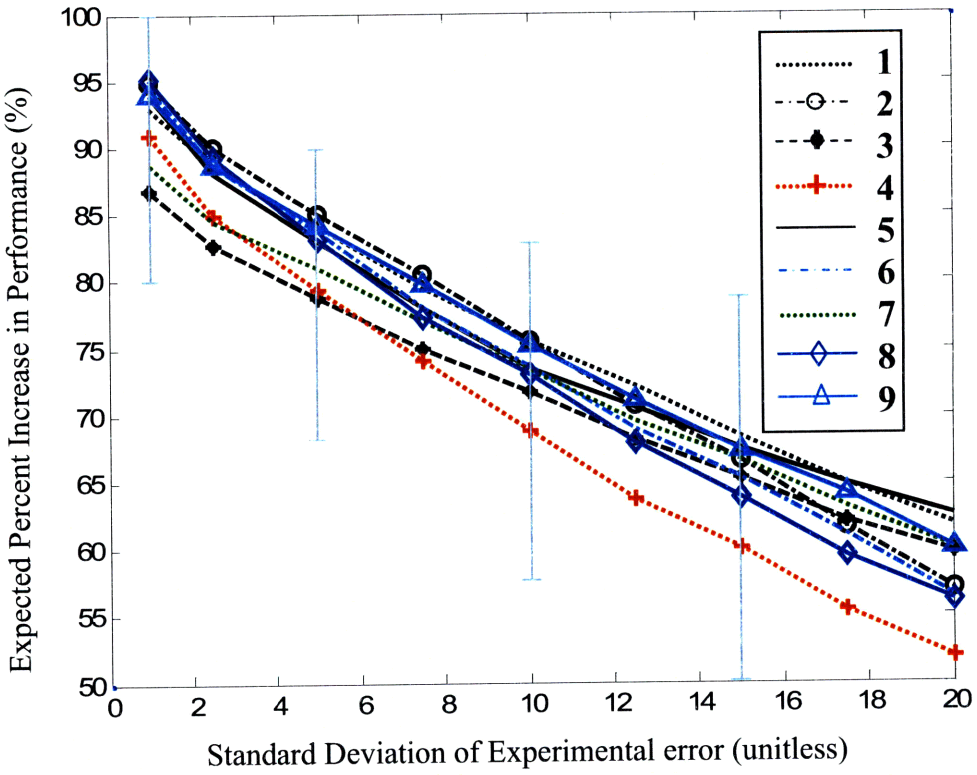

Figure 5.1Comparing an Ensemble of four *a*OFATs with nine different strategies, using the HPM; with 25th and 75th Percentile error bars (average)

Figure 5.1 shows the performance of the algorithms labeled 1 through 9 in accordance to the serial numbers from table 5.1. The main conclusions drawn from these simulations are:

(i) The lack of large differences between the nine methods.

(ii) In a relative comparison of algorithms, the ones that work best at low error, tend to perform poorly at high error, and vice versa. Exception to this are algorithms 1,5, and 9.

(iii) Algorithm 5 works reasonably well at high and low errors but performs poorly at medium error.

(iv) In terms of the average expected value of improvement across the entire error range, algorithms 9 and 1 performed the best. Algorithm 9 corresponds to a rank sum aggregation, with starting points maximally separated, with variables changed in random order. Algorithm 1 corresponds to a rank sum aggregation with fixed starting points using a random order in which variables were changed

(v) The performance of algorithms 3 and 4 was poor. Algorithm 3 aggregated using equal voting, with random starting locations, and a random variable change order. Algorithm 4 used a take-the-best strategy to aggregate, with a fixed starting location, and used a fixed order for variable changes.

(vi) The strategy of maximizing the starting locations, and maximizing negative correlation of variable change orders, were two strategies that consistently performed well. However, their contribution when placed together was not synergistic.

(vii) The variation in performance was similar across all the algorithms. The error bars shown in figure 5.1 are obtained by taking the average lower and upper quartiles across all nine algorithms for the specified error level. This describes the variability we can expect from running a given algorithm at a certain error level.

5.2 Analyzing factorial designs and comparing algorithms

In the previous sections, fractional factorial designs have been analyzed using stepwise regression with a variable α (acceptance p-value). This algorithm is described in section 2.1.2. In this portion of the thesis, other strategies of analyzing the factorial designs are investigated. The performances of these analysis techniques are compared to each other, and with ensemble *a*OFAT results. These methods are applied on a 2^{7-2}, resolution IV fractional factorial design (with randomly selected aliasing patterns in each simulation). Four fractional factorial analysis techniques are tested:

(iii) Stepwise Regression with a changing α: This is the method of analysis that is employed in the thesis thus far. Its performance is superior to a stepwise regression with fixed value for accepting and excluding terms (0.05 and 0.15). The procedure followed for stepwise regression is adopted from Draper and Smith (1981).

(iv) Lenth Method: The Lenth Method of analysis (Lenth, 1989) is a computationally simple method for estimating significant effects in factorial designs. It is applicably on designs which are unreplicated. In this technique a plot is made of all the effects, and a threshold value is created for separating significant effects from those that are deemed not significant. In the first step, a parameter s_0 is created, such that $s_0 = 1.5 \times meidan|\beta|$, where β includes all estimated effects and interactions including main effects and interactions. Using s_0, the pseudo standard error (PSE) is computed. This is defined by $PSE = 1.5 \times median|\beta|$, for $\forall \beta < 2.5 \times s_0$. Subsequently, the margin of error of the effects is defined as $MOE = t_{0.025,df} \times PSE$. The margin of error effects serves as the threshold demarcating between significant and insignificant effects. Lenth proposed using one third the total number of effects for degrees of freedom for the *t*-distribution (Lenth, 1989). In the Lenth method there is a more conservative measure for effect significance, termed as simultaneous margin of error. This was not investigated since the past experience with effect significance (when using

stepwise regression with fixed α of 0.05) has been that there are often insufficient terms to pick the best treatment.

(v) Classical Analysis: This conceptually simple procedure was introduced in section 4.1. Here, each factor is analyzed separately. The average response at the different levels of the factor is compared and best one is selected. This is repeated for all factors, sequentially. It is popularly used in the context of robust design when crossed arrays are used in conjunction with signal-to-noise ratios (Phadke, 1989). This technique is likely to work well in environments with high experimental error and perform poorly in environments with a lot of interactions.

(vi) Take-the-best: This straightforward technique was also introduced in section 4.1. Here the algorithm picks the treatment combination with the highest output observed in the experiment. This method is more susceptible to experimental error, since it just takes one erroneously represented outcome to mislead the entire analysis. As a result, this analysis is likely to perform well at regions of low experimental error and poorly in regions of high error. Also, in this technique the recommended treatment can only be selected from the set of treatments that were tested in the experimental fraction. This is not true of the other methods. In fact, quite often the treatment recommended by the other methods will be a vector that was not a part of the experimental plan, and the experimenter might need a confirmatory run to ensure the effectiveness of this combination.

The performance of these four analysis methods is shown in the figure 5.2.

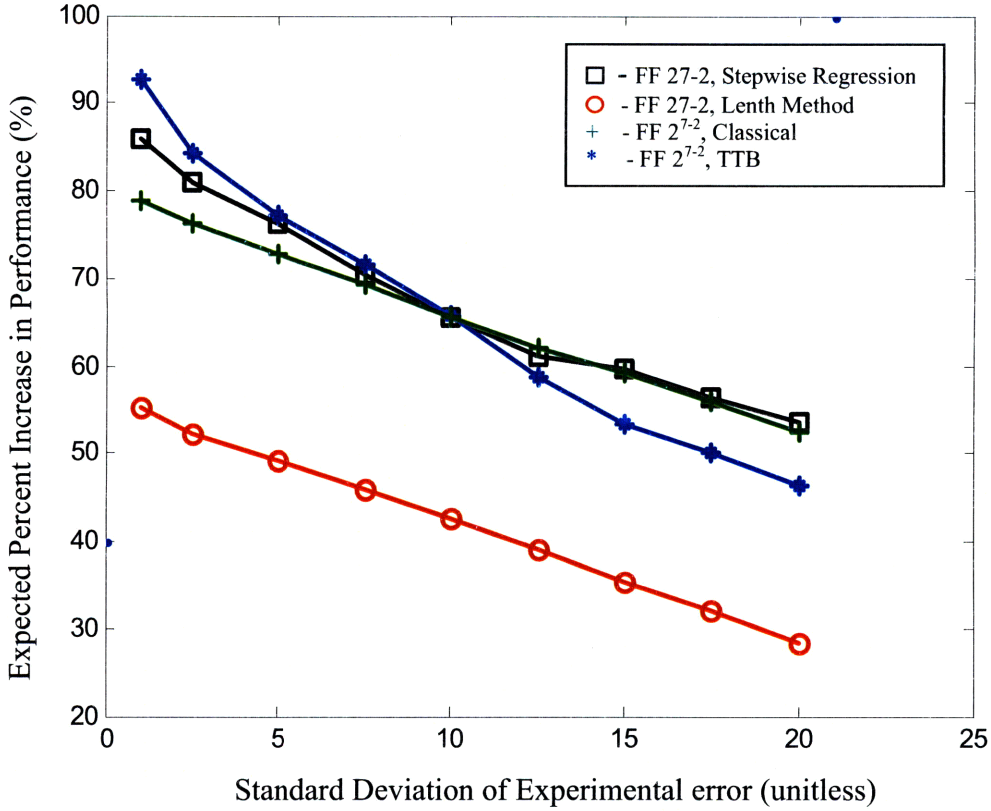

Figure 5.2 Comparing multiple methods of analyzing a 2^{7-2} Fractional Factorial array, using the HPM

As expected, at regions of low experimental error take-the-best strategy worked well and classical analysis performed poorly. And, at regions of high experimental error classical analysis performed well while take-the-best performed poorly. Stepwise regression performed well through out most regions of experimental error. Surprisingly, Lenth method performed poorly through out the entire range of experimental errors. In most cases, this was mostly because of the methods failure to identify sufficient number of significant terms. On average this analysis detected only 33% of the significant terms, although almost never falsely labeling a term as being significant when it was not. This method seemed to work better when the number control factors was reduced.

It is of interest to compare the range of performances in Ensemble *a*OFATs (figure 5.1), with the range of performance of analyzing fractional factorial arrays (figure 5.2). This is shown in figure 5.3.

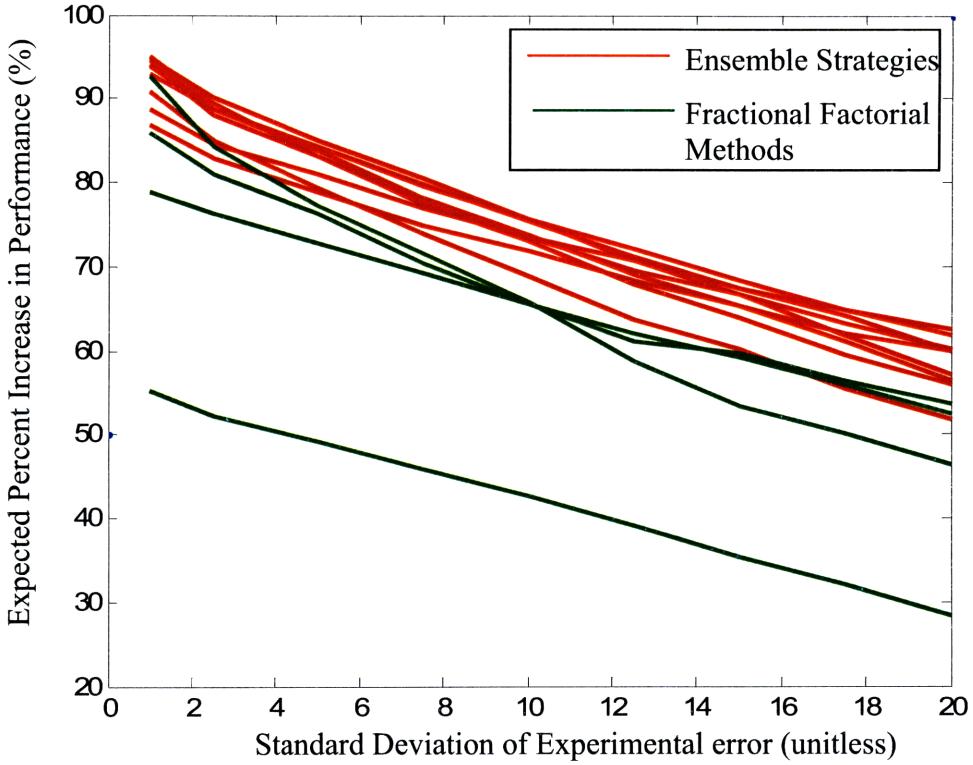

Figure 5.3 Comparing the various strategies of analyzing Ensemble *a*OFATs and Fractional Factorial arrays

The conclusion that was drawn from this comparison is that on average the Ensemble *a*OFAT procedures performed better than the differently analyzed factorial arrays. The best performance of the ensemble of *a*OFATs outperformed the best performances of the factorial design.

6 Conclusions and Future Work

Looking at adaptive, decentralized, algorithmic approaches to experimentation provides improved benefits to the system. This study proposes one such experimental algorithm. The conclusion reached in this study is that the proposed method outperforms a traditional factorial design and analysis method in improving the performance of the systems under experimentation. This is achieved by identifying superior design settings with a higher probability. The algorithms are tested on a hierarchical probability meta-model (HPM) and an illustrative case study that aims reduce the drag torque of a wet clutch. Two versions of each of the two algorithms (Fractional factorials and Ensemble *a*OFATs) were compared. The low-resource intensive version which requires 32 resources (Fractional Factorial 2^{7-2} versus an ensemble of 4 *a*OFAT's), and a high-resource intensive version which requires 64 resources (Fractional Factorial 2^{7-1} versus an ensemble of 8 *a*OFAT's) is tested. The superior performance of the proposed method is consistent across both studies (HPM and Wet clutch case study), both versions (high resource and low resource), as well as the range of experimental errors at which the methods were tested. At best the proposed algorithm is significantly better than fractional factorial arrays and at worst the performances are indistinguishable. This study also showed that the ensemble of *a*OFATs outperformed an ensemble of fractional factorial designs. The discussion following these empirical results provides a theoretical explanation for the observations.

A summary of the performances of the these algorithms from chapters 2, 3 and 4 are summarized in table 6.1

Experimental Environment	Experimental Error	Ensemble of *a*OFATs	Fractional Factorial	Ensemble of Fractional Factorials	
				TTB	Classical
HPM Low-Resource Version (32 Resources)	Low	94.0%	85.9%	92.3%	79.9%
	Medium	73.5%	65.5%	71.0%	66.3%
	High	62.5%	53.4%	52.1%	58.4%
HPM High-Resource Version (64 Resources)	Low	96.0%	96.9%	N/A	
	Medium	78.2%	73.7%		
	High	66.9%	62.9%		
Case Study Low-Resource Version (32 Resources)	Low	96.4%	85.5%		
	Medium	63.46%	50.0%		
	High	45.73%	35.9%		
Case Study High-Resource Version (64 Resources)	Low	99.5%	97.8%		
	Medium	75.3%	66.6%		
	High	57.1%	46.6%		

Table 6.1Summary of the performance of algorithms in this thesis

Finally, a set of methodological extensions to the ensemble *a*OFAT algorithm is studied, along with alternate ways of analyzing factorial designs. These extensions and modifications to previously studied algorithms further strengthen the main findings that, in general, the proposed algorithm outperforms factorial designs.

By outperforming standard experimental plans using a non-traditional experimentation approach, this thesis encourages further advances in a relatively unexplored line of experimental study and practice. While this study shows a specific experimental algorithm to be effective, a more general conclusion is that an adaptive and algorithmic approach to experimentation could lead to benefits previously unachieved.

6.1 Future Work

This idea associated with aggregating multiple *a*OFATs is relatively unexplored and the findings from this study open up a promising set of extensions. A potential list is presented here:

1. This thesis estimates the expected value of improvement through meta-models and case studies. A theoretical formulation quantifying the expected value of improvement for the proposed methodology is seen as the most important next step in formalizing such a procedure. A comparison to the expected value of improvement of traditional approaches will also serve as a validation to the results found in this thesis.
2. The range and extent of algorithmic modifications that can be made are endless. Specifically, this author feels that scope for improvement in aggregation techniques is promising. In this thesis all the weighted voting techniques, gather their weights based as some function of the response of the best treatment (the recommendation of the *a*OFAT). Other alternatives could look at where in the *a*OFAT scheme the best location was discovered, or the average performance of the other treatments tested by the *a*OFAT, or response improvements in individual level changes of variables. Also, Bayesian techniques could be used to aggregate the *a*OFATs, if a theoretical formulation quantifying expected value of improvement is constructed.
3. Having different starting points, and changing order in which variables are tested are ways of creating structurally similar *a*OFAT experiments that explore the design space in different ways. Such diversity can be achieved through other means. An example of which is to incorporate stochasticity in the *a*OFAT algorithm. In a standard *a*OFAT, once all the levels of a certain factor is tested, the best one is set before moving on to the next factor. This can be changed to an algorithm which provides probabilities for the levels that each factor could be set at. This could be a function of the response that resulted in each level, thereby giving levels that discovered high responses a high probability. This is seen as

being especially valuable when the problem at hand provides factors with many levels (3 or more)

4. The current algorithm creates multiple *a*OFATs which function independently and in a decentralized manner. The findings from each experiment are aggregated in the end. Alternatives to the decentralized nature of the technique by allowing *a*OFATs to communicate their findings with each other at intermediate points in the experiment could lead to advances. Such a quasi-decentralized algorithm could be more effective than the method currently proposed.
5. Alternate methods of analyzing Fractional Factorial arrays could be explored. In this study stepwise regression, Lenth method, and two more basic concepts are tested. The use of best subsets regression was explored but found to be computationally infeasible. However, other approaches, some with Bayesian reasoning have emerged in the literature (Joseph and Delaney, 2008). The impact such techniques have on the efficacy of factorial designs is of interest.
6. The use of multiple, independent experiments allows for them to be run simultaneously on different resources. The implication of this in terms of concurrent engineering practices and time taken to complete experiments has not been discussed in this paper. An on-site study exploring this aspect could provide more insights on the effectiveness of such methods in the industry.
7. The value in using adaptive, decentralized, agent-based, algorithmic approaches to experiments has been shown in this thesis. The value of such approaches can be extended to other formal tools and techniques used to solve large, complex problems. Some examples of domains where these insights can be applied include, optimization, multi-attribute decision-making, network analysis, etc.

References

Amit, Y., and Geman, D., (1997), "Shape Quantization and Recognition with Randomized Trees", *Neural Computation*, vol. 9, pp. 1545-1588.

Atkinson, A. C., and Bailey, R. A., (2001), "One Hundred Years of the Design of Experiments On and Off the Pages of Biometrika", *Biometrika*, vol.88, pp. 53-97

Barker, T. B., (1985), *Quality by Experimental Design*, Marcel Dekker, Inc. New York.

Barnett, J., Czitrom, V., John, P. W. M., and Leon, R.V., (1997), "Using Fewer Wafers to Resolve Confounding in Screening Experiments", *Statistical Case Studies for Industrial Process Improvement,* eds. V. Czitrom and P.D. Spagon, Philadelphia: SIAM, pp. 235-250

Berk, R. A., (2006), "An Introduction to Ensemble Methods for Data Analysis", *Sociological Methods & Research,* vol.34, p 263.

Bishop C.M., (1995), *Neural Networks for Pattern Recognition*, Clarandon Press, Oxford

Box, G.E.P., and Draper, N.R., (1987), *Empirical Model-Building and Response Surfaces,* John Wiley & Sons, New York

Box, G.E.P., and Hunter, J.S., (1961), "The 2^{k-p} Fractional Factorial Design". *Technometrics*, vol.3, pp 311-351; 449-458.

Box, G.E.P., Hunter, J.S., and Hunter, W.G., (1978) *Statistics for Experimenters*, Wiley, New York.

Box, G. E. P., and Liu, P.Y.T., (1999). "Statistics as a Catalyst to Learning by Scientific Method Part I - An Example", *Journal of Quality Technology*, vol.31, pp. 1-15.

Box, G.E.P., and Meyer, R.D., (1986), "An Analysis for Unreplicated Fractional Factorials", *Technometrics*, vol.28, no.1, pp 11-18.

Box, G.E.P., and Wilson, K.B., (1951), "On the Experimental Attainment of Optimum Conditions", *Journal of Royal Statistical Society*, vol. B 13, pp. 1-38.

Breiman L. (1996), "Bagging Predictors", *Machine Learning* 26:123-40

Breiman L. (2001a), "Random Forests", *Machine Learning* 45:5-32

Breiman L. (2001b), "Statistical Modeling: Two Cultures (with discussion)", *Statistical Science* 16:199-231

Breiman, L., Friedman, J., Olshen, R., and Stone, C., (1984), *Classification and regression trees.* Wadsworth, Belmont.

Buhlmann, P., and Yu, B., (2002), "Analyzing Bagging", *The Annals of Statistics*, vol.30, no.4, pp.927-961.

Chen, J., (1992)" Some results on 2^{n-k} Fractional Factorial Designs and search for Minimum Aberration Designs," *Annals of Statistics,* vol.20, pp. 2124-2141.

Chen, J., (1998), " Intelligent search for 2^{13-6} and 2^{14-7} Minimum Aberration Designs", *Statistica Sinica*, vol.8, pp.1265-1270.

Chen, J., Sun, D.X., and Wu C.F.J., (1993), " A Catalogue of Two-Level and Three-Level Fractional Factorial Designs with Small Runs," *International Statistical Review*, vol. 61, pp.131-145.

Chen, J., and Wu, C.F.J, (1991) "Some Results of S^{n-k} Fractional factorial designs with Minimum Aberration or Optimal Moments", *Annals of Statistics,* vol.19, pp.1028- 1041.

Chipman, H., (1996), "Bayesian Variable Selection with Related Predictors", *Canadian Journal of Statistics*, vol.24, pp17-36.

Chipman, H., Hamada, M., and Wu, C.F.J, (1997), "A Bayesian Variable Selection for Analyzing Designed Experiments with Complex Aliasing", *Technometrics*, vol.39, pp.372-382

Cormen, Leiserson, and Rivest (1990), "Greedy Algorithms", *Introduction to Algorithms*, pp. 329.

Czitrom, V., (1999), "One-Factor-at-a-Time Versus Designed Experiments", *American Statistician,* 53, 2

Daniel, C., (1962), "Sequences of fractional replicates in a 2^{p-q} series", *Journal of Americal Statistical Association*, vol.63, pp. 403-429.

Daniel, C., (1976), *Application of Statistics to Industrial Experimentation*, John Wiley, New York.

Daniel, C., (1973), "One-at-a-time Plans", *Journal of American Statistical Association*, vol. 68, pp 353-360.

Daniel, C., (1976), *Application of Statistics to Industrial Experimentation*, John Wiley, New York.

Dean, A. and Voss, D., (1991), *Design and Analysis of Experiments*, Springer-Verlag, New York.

Dietterich, T. G., (2000a). "Ensemble Methods in Machine Learning" In J. Kittler and F. Roli (Ed.) *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science* (pp. 1-15). New York: Springer Verlag

Dietterich, T.G., (2000b), "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization", *Machine Learning*, 40(2), pp.139-158.

Draper, N.R., and Smith, H., (1981), *Applied Regression Analysis*, Wiley, New York

Fisher, R. A., (1926), "The Arrangement of Field Experiments" *Journal of the Ministry of Agriculture of Great Britain*, vol. 33: pp 503-513

Freund, Y., and Schapire, R., (1997), "A decision-theoretic Generalization of Online Learning and an Application to Boosting", Journal of Computer and System Sciences. Vol 55, pp. 119-139

Frey, D. D., Engelhardt, F., and Greitzer, E. M., 2003, "A Role for One Factor at a Time Experimentation in Parameter Design", *Research in Engineering Design* **14**, pp. 65-74.

Frey, D. D., and Jugulum, R., (2006), "The Mechanisms by which Adaptive One-Factor-at-a-Time Experimentation Leads to Improvement," *ASME Journal of Mechanical Design*, vol. 128, pp. 1050-60.

Frey, D.D., and Li, X., (2008), "Using Hierarchical Probability Models to Evaluate Robust Parameter Design Methods", *Journal of Quality Technology*, vol.40, no.1, pp.59-77.

Frey, D. D., and Sudarsanam, N., (2008), "An Adaptive One-factor-at-a-time Method for Robust Parameter Design: Comparison with Crossed Arrays via Case Studies," *ASME Journal of Mechanical Design*, Vol. 130, Is.2, Feb, pp.02140-14.

Frey, D. D., and Wang, H., (2006), "Adaptive One-factor-at-a-time Experimentation and Expected Value of Improvement", *Technometrics* **48**, pp. 418-31.

Friedman, J.H., and Hall, P., (2007), "On Bagging and Nonlinear Estimation", *Journal of Statistical Planning and Inference*, vol. 137, iss.3, pp. 669-683.

Friedman, M., and Savage, L.J., (1947), "Planning Experiments Seeking Maxima," in *Techniques of Statistical Analysis*, pp. 365-372.

Fries, A., and Hunter, W.G., (1980), "Minimum Abberation 2^{k-p} Designs", *Technometrics*, vol.22, pp.601-608.

Hamada, M., and Wu, C.F.J., (1992), "Analysis of Designed Experiments with Complex Aliasing", *Journal of Quality Technology*, vol.24, iss.3, pp. 130-137.

Hand, D. J., Mannila, K., and Smyth, P., (2001), *Principles of Data Mining (Adaptive Computation and Machine Learning)*, MIT Press: Cambridge.

Hansen, L.K., and Salamon, P., (1990), "Neural Network Ensembles", *IEEE Transactions in Pattern Analysis and Machine Intelligence*, vol.12, no.10, Oct., pp. 993-1001

Hardwick, J. P., and Stout, Q. F., (1998), "Flexible Algorithms for Creating and Analyzing Adaptive Sampling Procedures," *New Developments and Applications in Experimental Design: Selected Proceedings of a 1997 Joint AMS-IMSSIAM Summer Conference*, Flourny, N., Rosenberger, W. F., and Wong, W. K.,eds., Institute of Mathematical Statistics, Hayward, CA, pp. 91–105.

Hastie, T., Tibshirani, R., and Friedman, J.,(2001) *The Elements of Statistical Learning: Data Mining, Inferences, and Prediction.* (Springer; New York).

Ho, T.K., (1998), "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(8), pp 832-844

Joseph, R.V., and Delaney, J.D., (2008), "Analysis of Optimization Experiments", *Journal of Quality Technology*, to appear.

Keijzer, M., and Babovic, V., (2000), "Genetic Programming, Ensemble Methods and the Bias/Variance Tradeoff – Introductory Investigations", *Genetic Programming: Third European Conference*, pp 76-90.

Koita, R., (1994), "Strategies for Sequential Design of Experiments", S.M. Thesis, Massachusetts Institute of Technology.

Koita, R., and Staelin, D.H., (1995), "One-at-a-time Design Strategy for Two-Level Factorials Experiments", *ASA Winter Conference, Raleigh, North Carolina.*

Lenth, R.V. Quick and easy analysis of unreplicated factorials. Technometrics 1989, 31-4, 469–473.

Li, X., Sudarsanam, N., and Frey, D. D., (2006), "Regularities in Data from Factorial Experiments," *Complexity* **11**, pp. 32-45.

Lloyd, F. A., (1974), "Parameters Contributing to Power Loss in Disengaged Wet Clutches", *Soc. Automotive Engineers Trans*, 83, pp 2498-2507

Logothetis N., and Wynn, H.P., (1994), *Quality through Design*, Clarendon, Oxford

Mcdaniel, W. R., and Ankenman, B.E., (2000), "Comparing Experimental Design Strategies for Quality Improvement with Minimal Changes to Factor Levels", *Quality and Reliability Engineering International*, Vol.16, pp. 355-362.

Miller, P.,(2007) "Swarm Theory", *National Geographic*, vol.212(1), pp126-147

Phadke, M.S. (1989). *Quality Engineering Using Robust Design*, Prentice-Hall, Englewood Cliffs, NJ.

Qu, X., and Wu, C.F.J, (2005), "One-Factor-at-a-Time Designs of Resolution V", *Journal of Statistical Planning and Inference*, vol. 131(2), pp. 407-416.

Quinlan, J.R., (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Robnik-Sikonja, M., (2004), "Improving Random Forests", *The 15th European Conference on Machine Learning*, pp. 359-370

Sasena, M., Parkinson, M., Reed, M.P., Papalambros, P.Y., and Goovaerts, P., (2005), "Improving an Ergonomics Testing Procedure via Approximation based Adaptive Experimental Design", *Journal of Mechanical Design*, vol. 136, pp. 1006-1013

Schmidt , S.R., and Launsby, R.G., Understanding Industrial Designed Experiments, 4th ed., Air Academy Press: Colorado.

Surowiecki, J., (2004), *The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* New York. Doubleday.

Tang, V., (2006), "Corporate Decision Analysis: An Engineering Approach", Ph.D. Thesis , Massachusetts Institute of Technology.

Tang, B., and Wu, C.F.J, (1996) "Characterization of Minimum Aberration 2^{n-k} Design in Terms of their Complimentary Designs," *Annals of Statistics*, vol.24, pp. 2549-2559.

Wang, H., (2007), "Sequential Optimization through Adaptive Design of Experiments", Ph.D. Thesis, Massachusetts Institute of Technology

Wilcoxon, F., (1945), "Individual comparisons by ranking methods", *Biometrics*, vol.1, pp. 80-83.

Wu, C.F.J., and Hamada, M., (2000) *Experiments: Planning, Design, and Parameter optimization*. John Wiley & Sons: New York.

Appendix A

	A	B	C	D	E	F	G		Response		Response with Experimental Error
1	-1	-1	-1	-1	-1	-1	-1		-27.2		-15.91
2	-1	-1	-1	-1	-1	-1	1		11.8		-7.20
3	-1	-1	-1	-1	-1	1	-1		-27.2		-29.79
4	-1	-1	-1	-1	-1	1	1		11.8		2.74
5	-1	-1	-1	-1	1	-1	-1		-35		-50.91
6	-1	-1	-1	-1	1	-1	1		4		9.80
7	-1	-1	-1	-1	1	1	-1		-35		-26.09
8	-1	-1	-1	-1	1	1	1		4		-1.90
9	-1	-1	-1	1	-1	-1	-1		-27.2		-33.79
10	-1	-1	-1	1	-1	-1	1		11.8		7.91
11	-1	-1	-1	1	-1	1	-1		-27.2		-48.34
12	-1	-1	-1	1	-1	1	1		11.8		-1.82
13	-1	-1	-1	1	1	-1	-1		-35		-31.79
14	-1	-1	-1	1	1	-1	1		4		7.07
15	-1	-1	-1	1	1	1	-1		-35		-41.12
16	-1	-1	-1	1	1	1	1		4		8.32
17	-1	-1	1	-1	-1	-1	-1		-9.6		6.06
18	-1	-1	1	-1	-1	-1	1		31		8.14
19	-1	-1	1	-1	-1	1	-1		-9.6		-11.69
20	-1	-1	1	-1	-1	1	1		31		47.07
21	-1	-1	1	-1	1	-1	-1		-7.8		-4.30
22	-1	-1	1	-1	1	-1	1		32.8		39.60
23	-1	-1	1	-1	1	1	-1		-7.8		-6.43
24	-1	-1	1	-1	1	1	1		32.8		36.24
25	-1	-1	1	1	-1	-1	-1		-9.6		-10.46
26	-1	-1	1	1	-1	-1	1		31		42.71
27	-1	-1	1	1	-1	1	-1		-9.6		-20.68
28	-1	-1	1	1	-1	1	1		31		16.92
29	-1	-1	1	1	1	-1	-1		-7.8		-5.37
30	-1	-1	1	1	1	-1	1		32.8		39.46
31	-1	-1	1	1	1	1	-1		-7.8		-3.87
32	-1	-1	1	1	1	1	1		32.8		43.06
33	-1	1	-1	-1	-1	-1	-1		-26.4		-15.31
34	-1	1	-1	-1	-1	-1	1		11		6.78
35	-1	1	-1	-1	-1	1	-1		-26.4		-0.67
36	-1	1	-1	-1	-1	1	1		11		6.87
37	-1	1	-1	-1	1	-1	-1		-34.2		-47.12
38	-1	1	-1	-1	1	-1	1		3.2		9.45
39	-1	1	-1	-1	1	1	-1		-34.2		-49.54
40	-1	1	-1	-1	1	1	1		3.2		6.87

41	-1	1	-1	1	-1	-1	-1		-26.4		-33.73
42	-1	1	-1	1	-1	-1	1		11		4.34
43	-1	1	-1	1	-1	1	-1		-26.4		-28.05
44	-1	1	-1	1	-1	1	1		11		10.11
45	-1	1	-1	1	1	-1	-1		-34.2		-29.94
46	-1	1	-1	1	1	-1	1		3.2		4.06
47	-1	1	-1	1	1	1	-1		-34.2		-36.21
48	-1	1	-1	1	1	1	1		3.2		5.21
49	-1	1	1	-1	-1	-1	-1		-8.8		-25.19
50	-1	1	1	-1	-1	-1	1		30.2		32.74
51	-1	1	1	-1	-1	1	-1		-8.8		2.28
52	-1	1	1	-1	-1	1	1		30.2		39.41
53	-1	1	1	-1	1	-1	-1		-7		-5.54
54	-1	1	1	-1	1	-1	1		32		49.68
55	-1	1	1	-1	1	1	-1		-7		-7.70
56	-1	1	1	-1	1	1	1		32		34.85
57	-1	1	1	1	-1	-1	-1		-8.8		-6.40
58	-1	1	1	1	-1	-1	1		30.2		39.69
59	-1	1	1	1	-1	1	-1		-8.8		-19.92
60	-1	1	1	1	-1	1	1		30.2		22.56
61	-1	1	1	1	1	-1	-1		-7		-16.02
62	-1	1	1	1	1	-1	1		32		21.48
63	-1	1	1	1	1	1	-1		-7		-6.39
64	-1	1	1	1	1	1	1		32		19.08
65	1	-1	-1	-1	-1	-1	-1		-25.6		-29.13
66	1	-1	-1	-1	-1	-1	1		11.8		20.33
67	1	-1	-1	-1	-1	1	-1		-25.6		-26.98
68	1	-1	-1	-1	-1	1	1		11.8		3.01
69	1	-1	-1	-1	1	-1	-1		-33.4		-9.28
70	1	-1	-1	-1	1	-1	1		4		-7.44
71	1	-1	-1	-1	1	1	-1		-33.4		-44.62
72	1	-1	-1	-1	1	1	1		4		2.40
73	1	-1	-1	1	-1	-1	-1		-25.6		-16.48
74	1	-1	-1	1	-1	-1	1		11.8		6.20
75	1	-1	-1	1	-1	1	-1		-25.6		-33.61
76	1	-1	-1	1	-1	1	1		11.8		18.93
77	1	-1	-1	1	1	-1	-1		-33.4		-37.25
78	1	-1	-1	1	1	-1	1		4		20.62
79	1	-1	-1	1	1	1	-1		-33.4		-36.20
80	1	-1	-1	1	1	1	1		4		4.30
81	1	-1	1	-1	-1	-1	-1		-9.6		-6.70
82	1	-1	1	-1	-1	-1	1		29.4		33.63
83	1	-1	1	-1	-1	1	-1		-9.6		-4.94
84	1	-1	1	-1	-1	1	1		29.4		48.78
85	1	-1	1	-1	1	-1	-1		-7.8		-16.08

86	1	-1	1	-1	1	-1	1		31.2		22.38
87	1	-1	1	-1	1	1	-1		-7.8		-11.35
88	1	-1	1	-1	1	1	1		31.2		32.80
89	1	-1	1	1	-1	-1	-1		-9.6		-3.82
90	1	-1	1	1	-1	-1	1		29.4		50.69
91	1	-1	1	1	-1	1	-1		-9.6		-20.00
92	1	-1	1	1	-1	1	1		29.4		32.93
93	1	-1	1	1	1	-1	-1		-7.8		-6.54
94	1	-1	1	1	1	-1	1		31.2		23.98
95	1	-1	1	1	1	1	-1		-7.8		-6.08
96	1	-1	1	1	1	1	1		31.2		40.66
97	1	1	-1	-1	-1	-1	-1		-26.4		-17.97
98	1	1	-1	-1	-1	-1	1		12.6		6.21
99	1	1	-1	-1	-1	1	-1		-26.4		-24.87
100	1	1	-1	-1	-1	1	1		12.6		34.68
101	1	1	-1	-1	1	-1	-1		-34.2		-15.88
102	1	1	-1	-1	1	-1	1		4.8		6.51
103	1	1	-1	-1	1	1	-1		-34.2		-27.43
104	1	1	-1	-1	1	1	1		4.8		9.92
105	1	1	-1	1	-1	-1	-1		-26.4		-24.41
106	1	1	-1	1	-1	-1	1		12.6		11.43
107	1	1	-1	1	-1	1	-1		-26.4		-43.38
108	1	1	-1	1	-1	1	1		12.6		20.96
109	1	1	-1	1	1	-1	-1		-34.2		-33.08
110	1	1	-1	1	1	-1	1		4.8		26.15
111	1	1	-1	1	1	1	-1		-34.2		-40.94
112	1	1	-1	1	1	1	1		4.8		5.42
113	1	1	1	-1	-1	-1	-1		-10.4		-43.89
114	1	1	1	-1	-1	-1	1		30.2		23.50
115	1	1	1	-1	-1	1	-1		-10.4		-4.31
116	1	1	1	-1	-1	1	1		30.2		26.82
117	1	1	1	-1	1	-1	-1		-8.6		-16.18
118	1	1	1	-1	1	-1	1		32		34.40
119	1	1	1	-1	1	1	-1		-8.6		-5.41
120	1	1	1	-1	1	1	1		32		32.04
121	1	1	1	1	-1	-1	-1		-10.4		0.20
122	1	1	1	1	-1	-1	1		30.2		37.36
123	1	1	1	1	-1	1	-1		-10.4		-9.66
124	1	1	1	1	-1	1	1		30.2		47.02
125	1	1	1	1	1	-1	-1		-8.6		0.66
126	1	1	1	1	1	-1	1		32		33.50
127	1	1	1	1	1	1	-1		-8.6		0.49
128	1	1	1	1	1	1	1		32		19.95

Appendix B

Alias Structure for a 2^{7-4}, resolution IV fractional factorial design

I = CEFG = ABCDF = ABDEG

A = BCDF = BDEG = ACEFG

B = ACDF = ADEG = BCEFG

C = EFG = ABDF = ABCDEG

D = ABCF = ABEG = CDEFG

E = CFG = ABDG = ABCDEF

F = CEG = ABCD = ABDEFG

G = CEF = ABDE = ABCDFG

AB = CDF = DEG = ABCEFG

AC = BDF = AEFG = BCDEG

AD = BCF = BEG = ACDEFG

AE = BDG = ACFG = BCDEF

AF = BCD = ACEG = BDEFG

AG = BDE = ACEF = BCDFG

BC = ADF = BEFG = ACDEG

BD = ACF = AEG = BCDEFG

BE = ADG = BCFG = ACDEF

BF = ACD = BCEG = ADEFG

BG = ADE = BCEF = ACDFG

CD = ABF = DEFG = ABCEG

CE = FG = ABCDG = ABDEF

CF = EG = ABD = ABCDEFG

CG = EF = ABCDE = ABDFG

DE = ABG = CDFG = ABCEF

DF = ABC = CDEG = ABEFG

DG = ABE = CDEF = ABCFG

ACE = AFG = BCDG = BDEF

ACG = AEF = BCDE = BDFG

BCE = BFG = ACDG = ADEF

BCG = BEF = ACDE = ADFG

CDE = DFG = ABCG = ABEF

CDG = DEF = ABCE = ABFG