THE SEARCH FOR EXPERIMENTAL DESIGN WITH DOZENS OF VARIABLES

by

Yaileen M. Méndez-Vázquez

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE
in
INDUSTRIAL ENGINEERING

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2015

Approved by:

_____	_____
Mayra Méndez Piñero, PhD Member, Graduate Committee	Date
_____	_____
Luis F. Gordillo, PhD Member, Graduate Committee	Date
_____	_____
Oscar Marcelo Suárez, PhD Member, Graduate Committee	Date
_____	_____
Mauricio Cabrera-Ríos, PhD President, Graduate Committee	Date
_____	_____
Edgar Acuña, PhD Representative of Graduate Studies	Date
_____	_____
Viviana Cesaní, PhD Chairperson of the Department	Date

ABSTRACT

Simulation models have importantly expanded the analysis capabilities in engineering designs. With larger computing power, more variables can be modeled to estimate their effect in ever-larger number of performance measures. Statistical experimental designs, however, are still somewhat focused on the variation of less than about a dozen variables. In this thesis, an effort to identify strategies to deal with dozens of variables is undertaken. The aim is to be able to generate designs capable to estimate full-quadratic models. Several strategies are contrasted: (1) generate designs with random numbers, (2) use designs already available in the literature, (3) generate designs under a clustering strategy, and (4) generate designs using random walk methods. The most significant area of opportunity is the manipulation of 50 or more variables, where the state-of-art seems to be at this point according to literature review, trials with different software packages and reviewers' feedback in referred journals.

RESUMEN

Los modelos de simulación han expandido de manera significativa la capacidad de análisis en diseños de ingeniería. Con un gran poder computacional, una mayor cantidad de variables pueden ser modeladas para estimar los efectos de cada vez más medidas de desempeño. Los diseños experimentales están enfocados en la variación de menos de una docena de variables. Este trabajo, se enfoca en la identificación de estrategias para trabajar con decenas de variables simultáneamente. El objetivo es la generación de un diseño experimental capaz de estimar modelos de regresión cuadráticos. Se realizará una comparación entre varias estrategias: (1) generación de diseños con números aleatorios, (2) generación de diseños mediante métodos ya existentes en la literatura o mediante programas existentes, (3) generación de diseños mediante la estrategia de "clustering", y (4) generación de diseños mediante métodos pseudo-aleatorios. El área de oportunidad de mayor significancia es la manipulación de 50 variables o más, donde el estado de arte parece ser en este punto, de acuerdo con la revisión de literatura, ensayos con diferentes paquetes de software y sugerencias por parte de revisores de revistas especializadas.

DEDICATION

To my little sister Salome, because you have been a gift from God to our family. Remember to always follow your dreams and conquer your goals with courage and determination. Every day I will continue to work hard in order to be a better role model for you and make you proud.

ACKNOLEDGMENTS

Firstly, I want to thank God for giving me the opportunity to complete my Masters degree and for allowing me to meet very special people along the road. Also, I want to thank to my family for their support in every moment. Thanks to my mom and dad, Mariluz Vázquez and Wilfredo Méndez for allowing me the opportunity to study a professional career and for their support in each of my goals and dreams. Thanks to you I have overcome and have achieved goals that I never imagined.

One of the most important people during this process of my Masters degree is my advisor, Prof. Mauricio Cabrera Ríos. The development of this thesis was possible thanks to him. Thanks for your guidance, support and your friendship. Thanks for making me believe in myself and for helping me grow personally and professionally. Thanks also to Dr. Clara Isaza for her support, counsel, friendship, and for letting me share with her family.

I want to especially thank Kasandra Ramírez Rojas and Hecny Candelario for their collaboration in this research. Also Juan Rosas, Enery Lorenzo, and Jaileene Pérez for their help in the editing process in the writing style and language. Additionally I want to thank the Applied Optimization Group, specifically Esmeralda Niño, Nitza García, Katia Camacho, Diana Sánchez, Yaritza Santiago, Yazeli Cruz, and Mary Carmen Acosta. As well IEGA members Samuel Bonet, Cesar Salazar and Isis Narváez for their support, motivation, and unconditional friendship.

A very special group of people are the personal of General Engineering Department, Dr. Jaime Ramírez Vick, Edda Rosado, Crimilda Pagán, Lucia Balaguer, and Norma Gómez. Thanks for their support, advice, and friendship. You are an essential part in the development of my career.

Also I want to thank the members of my graduate committee for their guidance and their advice. As well as the Industrial Engineering Department, the Crest Program (Grant HRD 0833112, NSF), and the NIH MARC Grant 5T36GM095335-02 'Bioinformatics Programs at Minority Institutions' for their financial support during the process of my Masters degree.

All of you were essential in the development of this thesis and my master's degree. I am very grateful to all of you.

Table of Contents

Table List

Figure List

1 Introduction

Systems in engineering and the sciences are affected by multiple variables simultaneously. Understanding how these variables affect key performance indicators is important for design, control and optimization purposes. Moreover, achieving an appropriate understanding level must commonly be carried out while being mindful of resource consumption. Assessing the effects of multiple variables on multiple performance measures has been made a lot more convenient by the development of computer simulation, where the resources are mostly computing time and power.

A somewhat standard approach to characterize and model variation through experimental means is the use of a regression model. Of special interest to this work is the situation in which curvature is suspected in the experimental response of interest, thus, a full quadratic regression model is sought. There seems to be an imbalance, however, between the increasing capability of computer simulation models to relate large numbers of variables to similar numbers in performance measures and the restricted focus of statistical experimental designs in dealing with a low number of variables.

This thesis attempts to bring attention to this imbalance and foster the generation of designs to investigate dozens of variables at a time. A more effective use of simulation models is possible with developments in this area including a more powerful use of simulation-based optimization.

1.1 Motivation

Experiments are key to characterize, model and optimize engineering systems. Planning, executing and analyzing experiments are activities that belong to the field of statistical design of experiments. Because conducting experiments require consumption of resources of many sorts, including time, energy, materials and money, special care has been devoted to keep a manageable number of experimental variables, runs and replicates.

The use of computer models and hence of computer simulation, has allowed engineering to predict the effect of dozen and sometimes hundreds of variables at a time in a particular system. Such capability, however, has been hampered by the combinational explosion that results from using classical techniques to generate experimental designs to analyze engineering systems. For example, even at the low number of dozen variables, the well-known full factorial design at three levels would result in 3^{10}= 59,049 experimental runs. Thus number goes up to 3^{20} = 3,486,284,401 runs for 20 variables at three levels each.

Computer-generated designs have been successfully coded in several software packages, but in the experience of this researcher it is still computationally difficult to generate designs for 40 or 50 variables. Such experience includes the use of Minitab, R, and JMP. Computer simulation can be greatly enhanced by the possibility to effectively explore dozen of variables in an efficient manner. This thesis intends to provide such capability.

1.2 Objective

The objective of this thesis is to devise strategies to generate experimental designs to explore the variation of dozen of variables simultaneously. These designs will, in turn, allow estimating full quadratic regression models with a minimum number of runs. An initial proof-of-concept aim has been set at 50 variables. It is further envisioned that the resulting strategy be implementable in a personal computer.

1.3 Thesis Organization

This thesis is organized as follows: The second chapter present an analysis of the most relevant recent literature. The third chapter introduces the relevant existing experimental designs and current design-generation algorithms used in commercially-available software. The original strategies to this thesis are presented in chapter four. The fifth chapter provides a comparison of the strategies/designs introduced previously in terms of statistical measures, as well as computational cost. The sixth chapter demonstrates the capability of designs with tens of variables in a series of simulation-optimization tasks. Finally, the seventh chapter discusses the general conclusions of the thesis and suggests directions for future research in this line of work.

2 Literature Review

Experiments play an important role in technology and manufacturing because they provide the basis to establish the cause-effect relationships between the controllable variables in the systems and the performance measures of interest **[25]**.

As technology and computational capacity increase, the possibility to analyze and simulate systems that are affected by multiple variables simultaneously is increasingly more attractive and feasible. Design of experiments (DOE), makes possible the study of these systems in an efficient manner, by providing a framework for planning, executing and analyzing experiments to arrive to objective conclusions that can be verified and replicated **[25]**.

Many experimental strategys exist, including the best-guess-approach and one-factor-at-a-time approach. The best-guess-approach is based in the knowledge of the experimenter and has many limitations, one of these is that using this method the user does not have the guarantee of neither repeatability nor optimality**[25]**. In the one-factor-at-a-time strategy the experimenter varies one variable at diferent levels while holding others constant. The limitations of this strategy is that it does not consider interactions and it can be very inefficient. As a popular statistical experimental design, the factorial design is already superior to both strategies just discussed. The factorial design allows to vary several factors simultaneously, and thus, to detect potential interactions. Two types of factorial designs are known: (i) full factorial and (ii) fractional factorial.

The full factorial design tries all possible combinations of the levels of the variables of interest. For example, two variables at three and four levels will results in a full factorial of 3 x 4 = 12 runs in one replicate. Needless to say, this strategy becomes impractical rapidly with a small

number of variables. For instance, for 10 variables at three levels, the full factorial requires 3^{10}= 59,049 runs. For this reason this design is commonly used for the experiments of systems that are affected only by a small number of variables **[4, 18, 38-39]**. Due to this combinatorial explosion, an alternative strategy is to run a fraction of the full factorial design; this design is known as a fractional factorial design.

The primary goal of the fractional factorial design is the selection of a subset of a full factorial experiment that bring the maximum amount of information of a system with a limited number of runs **[16]**. This experimental strategy is often used in process or product design, process improvement and industrial experimentation, where the principal objective is to perform a screening experiment to identify the variables that significantly affect a performance measure of interest **[9, 25]**. The two level fractional factorial design is widely used for this purpose. Three-level fractional factorial designs have not been favored in the literature due to the difficulty on separating important variable effects **[34]**.

For many experiments in the literature, the number of variables being investigated is less than a dozen; in fact, in most cases it is only three or four factors **[1-2, 5, 12-15, 18-19, 21, 24, 27, 29, 33-34, 38, 40-41]**. For many cases, the intention is to characterize the system with a second order model **[1-3, 9-10, 15, 17, 19, 26-29, 36-37, 40, 43, 46, 48-49]**. These models are often used in optimization experiments **[25]**. The experimental designs used for these systems are usually the fractional factorial design, the central composite design or the Box-Behnken design **[1-2, 12-13, 19, 25, 26, 36, 40, 43, 46, 48-49].** In the literature, it is mostly screening experiments that involve more than five or six variables. When the objective is to build a regression model, it is usually less

than that number **[5, 8, 41]**. This is true especially with full quadratic models for which a minimum of three levels per variable is necessary to estimate purely quadratic effects **[4-5, 41]**.

One of the best-known practices when fitting a full quadratic model is to employ a Central Composite Design, which entails the use of either a two-level full factorial or a fractional factorial design, plus 2 axial runs per variable involved, plus a predefined number of center runs **[1, 25, 28, 36-37, 40, 43, 47-49]**. This design often capitalizes on the use of a fractional factorial to keep the number of runs low while providing a stable and minimal variance in the coefficients **[11, 14]**.

Often the full factorial and the central composite design are not practically convenient as the number of runs grows exponentially **[25]**. They are also limited when exploring experimental regions that are irregular in shape. In such cases designs based in an optimality criteria are feasible alternatives **[25]**. These experimental designs are known as optimal designs, and are constructed depending on the criteria to be optimized. These designs are often generated using exchanges algorithms like the point exchange algorithm or the coordinate exchange method. In general, exchange algorithms begin with n-points in an initial design **[21**]. Then new points are added and other deleted aiming to improve a selected criterion **[21]**.

For the point exchange algorithm a grid of points is introduced as an input by the experimenter. From this grid, an experimental design X is selected initially. Basically, the algorithm exchange points that are in the grid, with points that are currently in the design X with the aim to improve the selected optimality criterion **[25]**.

Another method to generate optimal designs is the coordinate exchange algorithm. This is very useful in the development of optimal experimental designs for large number of variables **[21**]. It has been shown to have a reduction of two orders of magnitude in computational time in the generation of experimental designs with many variables **[21].** "This method searches over each

coordinate of every point in the initial design recursively until no improvement in the optimality criterion is found **[25]**." The coordinate exchange algorithm is more efficient than the point exchange, and is the method that is most often used in commercially-available software **[25]**. Both algorithms are heuristic in nature, and therefore, do not guarantee global optimality**[25]**.

There are several optimality criteria used to generate optimal designs. One of the most popular in the literature is the D-optimal criterion. The D-optimal design can be used for first and second order models **[25]**. In this design one can decide upon the number of experimental runs **[6]**. The objective of this strategy is to generate designs that result in the model parameter estimation with the lowest variance on the regression coefficients, as explained later in this document.

This strategy, as coded in some commercial and open-source software packages, uses an initial enumeration in which the predefined number of runs is chosen with the objective to meet the optimality criterion **[25]**. The experimental runs are chosen using in the most cases the coordinate exchange algorithms described previously. Optimal designs are useful when the sampling is expensive and when taking no more samples than absolutely necessary is encouraged.

As another option, if simplicity is important, a naïve way to generate a design is by using a probability mass function to prescribe a desired number of experimental combinations. This strategy is considered here due to its feasibility to explore several dozen of variables simultaneously, although no control over variance or any other statistical properties can be exercised in this instance.

Finally, for the selection of a design it is necessary to consider many issues for the particular system to be investigated. Currently, the majority of DOE techniques are somewhat focused in the analysis of the effect of less than a dozen of variables. Sanchez (2012), explains, however, that due to the magnitude of many complex and expensive simulation models used in the

Department of Defense, dozens of variables at a time must be studied concurrently **[30]**. She explained that many of these simulations often have hundreds or thousands of variables at hand **[30]**. Sanchez (2012) also sets forth the hypothetical situation where it is attractive to build a simulation model of 100 variables at two levels. She explains that for a petaflop computer, one with a capacity of thousand trillion operations per second, the evaluation of such model would take 40 millions of years **[30]**. It is recognized that efficient experimental design offers a great amount of information and benefits at a lower cost, offering also the capacity of dealing with higher dimensionality **[30]**. For this reason it is important to identify strategies to deal with large-scale systems in an efficient way, considering resources, time, and capacity consumption, but without compromising the quality of the results **[30, 44]**.

3 Background

3.1 Design of Experiment

Design of experiments (DOE) refers to the process of planning and conducting an experiment to obtain information about the effect of the variation of controllable variables in a given performance measure. The data obtained will be analyzed by statistical methods with the aim of obtaining valid and objective results in terms of a hypothesis about the system. Through experimentation techniques it is possible to prove cause-effect relationships. There are many types of experimental designs. The most relevant to this work are briefly explained next.

3.1.1 Full Factorial Design

The full factorial design contains all possible combinations of a set of factors in the analysis of a system. The full factorial design gives us a large amount of information, although at a high cost in the number of experimental runs **[25]**.

Indeed, the size of the design grows exponentially when increasing the number of variables. The number of experimental runs required to perform a full factorial design is n^k, where k represents the number of variables, and n represent the number of levels that the variables take. For this reason this design becomes impractical for an analysis of even a small number of variables, especially when the experimental runs are expensive.

3.1.2 Central Composite Design

The central composite design is particularly popular in the analysis of systems with the aim to characterize them with a second order model. This design is composed of a 2^k factorial

with n_F factorial runs, 2k axial runs, and n_c center runs **[25]**. Alternatively, a 2^{k-p} fractional factorial can replace the 2^k design for economy **[10]**.

There are three types of central composite designs. These are the circumscribed, the faced centered, and the inscribed design. The circumscribed design has the axial points extending beyond the cube position. In the faced design, the axial points are on the cube face. In the inscribed design the cube is inscribed within the axial points **[7]**.

Most of the statistical software packages have the capability to build a central composite design in short time, and yield to the user the opportunity to select the type of the design that is more useful according with their properties. Comparing with the full factorial, the central composite design reduces the size of the experimental design in a meaningful way. The Box-Behnken design is very similar to the central composite design, but it differs in the values of the parameters which are at the midpoints of edges of the design space and the center **[7]**.

3.1.3 D-Optimal Design

In this design, the idea is to minimize the determinant of the inverse of the so-called design information matrix, $|(X'X)^{-1}|$; where X is the N x p model matrix for the design, N is the number of experimental runs and p is the number of model parameters **[21, 25]**. "A D-optimal design minimizes the volume of the joint confidence region on the vector of regression coefficients," which represents the variance associated to the parameter estimation **[25]**. This design is often used for screening purposes, where in most cases the system is characterized by a first-order-model. The D-optimal design is highly popular in several software packages **[25]**.

3.1.4 Other Optimal Designs

For optimization purposes, obtaining a second order model is often times the objective. There are many types of optimal designs. Main types are briefly explained next.

The G-optimal and the I-optimal design, are based in G and I-optimality criteria, both of which focus on prediction capability **[25]**. The G-optimal design, specifically, is based in the prediction variance criteria **[25]**. This designs basically "minimize the maximum scale prediction variance over the design region **[25]**." The maximum scale prediction variance is calculated as follow:

$$\frac{NV[\hat{Y}(x)]}{\sigma^2} \tag{1}$$

Specifically, the I-optimal design is based in the integrated variance criteria. This design is focused in the minimization of the average prediction variance of the design space that is calculated as follow:

$$I = \frac{1}{A}\iint_{-1}^{1} V[\hat{Y}(x_1, x_2)]d_{x_1}d_{x_2} \tag{2}$$

Another alternative to develop optimal designs is through the use of the A-optimality criterion. This criterion minimizes the sum of the main diagonal elements of the design information matrix, $(X'X)^{-1}$**[25]**, that is its trace**.** Finally, the V-optimal design, is focused in the minimization of the average prediction variance of a set of points that was selected of an interest design region **[25]**.

4 Proposed Methods

In this work the objective is to identify strategies to generate experimental designs for tens of variables, with the intention to obtain a full quadratic model with the least possible number of experimental runs. In this chapter different strategies are proposed to develop experimental designs using the k-means clustering algorithm and random walks methods. The idea of the proposed methods is to generate experimental designs for tens of variables, in this work specifically for 10, 20 and 50 variables in a personal computer without the need of specialized expensive software, and with the capacity to estimate a full quadratic model with the less possible number of experimental runs. The development of experimental designs to explore an experimental region efficiently for cases of tens of variables has a significant impact in the field of simulation. The traditional experimental designs that are found in the literature are focused on the experimentation of a few variables, limiting in this sense the capability of simulation to explore dozens and hundreds of variables simultaneously. The proposed methods attend this situation allowing to developed experimental designs for tens of variables in a novel and efficient way. The proposed strategies are described in this chapter. These will be evaluated in the following chapter in terms of their statistical properties and generation cost.

4.1 Clustering Design Method: Initial Version

The Cluster Design Method is currently under development in our group and it was as follows in its initial form: (i) generate a full factorial design as an initial enumeration; (ii) add a column with uniformly distributed random numbers to the full factorial design; (iii) generate k clusters with the k-means algorithm, with k being the number of necessary regression coefficients plus one; (iv)

retrieve the k-medoids associated to the k clusters; (v) delete the values associated to the column with the random numbers; and (vi) present the experimental design.

The rationale behind step (i) is to provide initially orthogonal design points. A random dummy variable is introduced as a means to add a controlled perturbation in step (ii). This is necessary because clustering equally spaced orthogonal points results in very similar clusters, and thus to very similar centroids in the next step.

The k-means algorithm is the most basic of the clustering techniques. It iteratively forms a user-defined number k of exclusive clusters with each cluster organized around its average location or centroid. As proposed here, k is set to the number of necessary regression coefficients to fit a full quadratic model plus one in step (iii). The number of coefficients for v variables of interest can be calculated as:

$$1 + 2v + \binom{v}{2} \tag{3}$$

From step (iii), then, k clusters result. In step (iv) the medoid of each cluster is obtained. The medoids, which are data points in the center of a cluster, are intended as the k runs in the resulting cluster design. In this work, an approximate medoid is computed for each cluster by using the median of each of the values of the v variables of interest within the cluster under analysis. Steps (v) and (vi) of the method are self-explanatory.

Equation (3) is useful also to show the growth of the intended method when increasing the number of variables, as shown in **Figure 1**, where this growth is contrasted with that of the full factorial design.

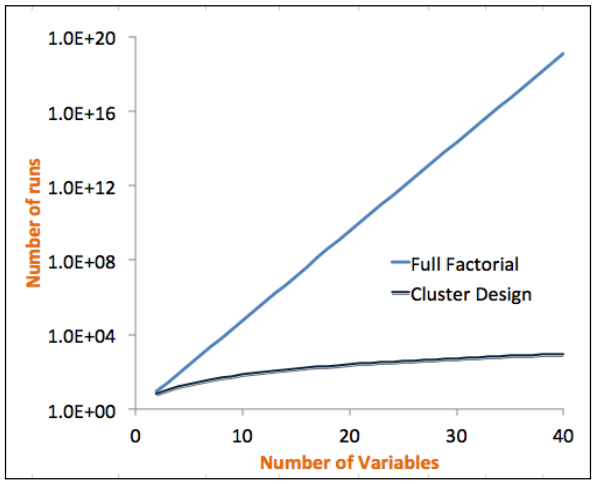

Figure 1. Growth in the number of runs as a function of the number of variables for the cluster design and the full factorial design.

Looking at **Figure 1**, it is clear that –if feasible- the cluster design would be convenient to explore tens of variables. However, a limitation also becomes apparent. The first step of the initial version of the method requires a full factorial enumeration, thus it would become computationally inconvenient at some point. This observation, corroborated by a series of tests, lead to the following modified version of the method.

4.2 Clustering Design Method: Modified Version

The first step of the original method required the generation of a full factorial enumeration, which would become computationally inconvenient at some point as shown previously. A slower growing enumeration would help alleviate this situation. The following modification was then introduced:

1) Generate a cluster design D1 of moderate size, say one to explore v=10 variables, using the original version of the method. D1 will have n runs.

2) Generate a second cluster design D2 as in the previous step. This second design will be different due to the random realization in step (ii). D2 will also have n runs.

3) Concatenate every run in D1 together with every run in D2. The resulting enumeration contains n^2 runs with $2v$ variables.

With this new enumeration in place, steps (ii) through (vi) can then be applied to generate a design for up to $2v$ variables. **Figure 2** shows the enumeration growth compared to the cluster design and the full factorial design.

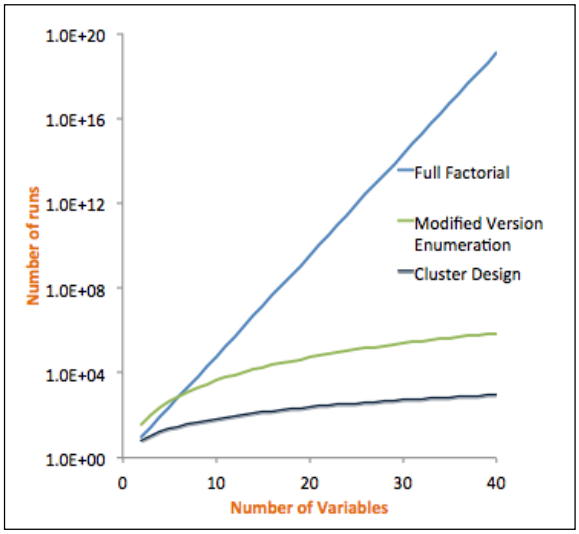

Figure 2.Growth on number of runs of the modified version enumeration compared to the full factorial enumeration.

4.3 Random Walk Method: Linear Congruential Generator Design

The clustering design method requires an initial enumeration as previously discussed. This makes it inconvenient when the number of variables increases. Experimental designs based on randomness, specifically in pseudo random methods are explored. The aim is to create a middle point between convenience and control of the resulting statistical properties. The idea behind the random walk generator is a path that initiates in a known point and jumps in a determined direction

with a given probability. The Linear Congruential Generator and the Mersenne Twister algorithm are explored to generate designs under this category.

The linear congruential method is a pseudo random number generator calculated with a linear equation as shown below:

$$Z_i = (aZ_{i-1} + c)(mod(m)) \quad (4)$$

Where *a* and *c* are the multiplier and increment parameters respectively, *m* is the module and Z_i is the remnant integer from the ration in the right-hand side of equation (4).

To generate an experimental design, each variable in the design was initialized setting Z_0 as a random integer number in a range from 1 to 3, where each value has a probability of 1/3. Multiplier parameter *a* and modulus *m* were set to values of 1 and 3 respectively. The increment parameter *c* was defined as a function of the random number generated (rng) as follows: *c = (0 if rng < 1/3, 1 if rng ≥1/3 and rng <2/3, or 2 if rng ≥2/3) +1*. A series of numbers were then generated to match the number of necessary regression coefficients to fit a full quadratic model plus one. A balanced design -with as many columns as design variables and as many rows as regression coefficients plus one- is generated with this method.

4.4 Random Walk Method: Mersenne Twister Design

The Mersenne Twister is derived from the generalized feedback shift register (GFSR) generator **[20]**. This algorithm has excellent statistical properties, including independence, uniformity and competitive equidistribution **[20]**. It also has a large period length of $2^{19937} - 1$. This algorithm generates uniform random numbers in the range of [0, 1], and has been programmed in many software packages, including R-Project, which is of free distribution **[31-32]**.

The generation of the experimental design was carried out in R-Project using the package called ‘rngSetSeedas', where the initial seed was set to a value of 5. The idea of this method is to focus on repeatability. When the same seeds are selected, the resulting designs will be identical.

5 Comparison of the methods

In this chapter a comparison of the proposed strategies to generate experimental designs for 10, 20 and 50 variables is presented. The comparison is based on two different aspects: (1) statistical properties of the designs generated with the methodologies identified in this work and (2) a cost evaluation based on computing time and the requirement of statistical software to generate the designs, for the case of 50 variables.

5.1 Statistical Properties Approach

The comparison among all competing strategies to generate experimental designs for tens of variables was carried out by artificially building a response through the addition of a known function and a random error. The known function was a full quadratic model, in the first case for 10 variables, 20 variables for a second case, and 50 variables for a third case, with all regression coefficients arbitrarily set equal to 10. The random error came from a normal distribution with 0 mean and standard deviation of 1.5 units.

The idea behind having an artificial response is to provide a controllable expected value and a random noise around it. The idea is focused in verification: if true experimental data can be effectively modeled with a full quadratic regression model, it will look very similar to our artificial response. If we control the artificial response, then we can measure the performance of our method when approaching it.

Experimental design from each strategy (i) random design, (ii) Full Factorial, Central Composite Design and D-Optimal Design, and (iii) the proposed Clustering Design, LCG Design,

and the Mersenne Twister Design, were used to sample and then to estimate the artificial response described previously. The following indicators were measured: (M1) number of runs, (M2) mean square error (MSE), (M3) number of regression coefficients estimated, (M4) the trace of $(X'X)^{-1}$, that is, the trace of the inverse of the so-called design information matrix, which is proportional to the covariance of the regression coefficients, and (M5) the determinant of $(X'X)^{-1}$. **[25]**

Residual analysis was also considered in this comparison to assess the assumptions of normality, independence and constant variance. This is carried out mostly through hypothesis testing. The residual is computed for the ith data point in n data points as $e_i=Y_i-\tilde{Y}_i$; i= [1, 2.....n], where Y_i is an actual observation and $\tilde{Y}_i$ is the corresponding fitted value from the regression model **[24-25]**.

A design with the lowest possible number of runs, the lowest MSE, capable to estimate all regression coefficients, with the lowest value of the trace and determinant of $(\mathbf{X'X})^{-1}$, and which complies with the residuals assumptions, would clearly dominate any other option.

Furthermore, it was important to assess how easy was to generate a design under each strategy. This last was done qualitatively by necessity. Finally, it was decided to tabulate the frequency of the coefficients by their percentual deviation from the target value. The results of the comparison are shown next for 10, 20, and 50 variables.

5.1.1 Experimental Designs for 10 Variables: Statistical Comparison Results

In this section, only the initial version of the cluster design was included. **Table 1** summarizes the comparative results for M1-M5 and **Table 2** shows the distribution of the percentage deviation from the intended regression coefficient value. The D-Optimal design seems to be an overall robust and sensitive alternative according to these results, with a minimum number of runs, the second lowest MSE, the capability to estimate all coefficients, and performing well in goodness-of-fit.

The full factorial and the central composite designs, even at 10 variables, start to seem impractical in terms of number of runs. This behavior was expected to be more drastic with larger numbers of variables. Looking at **Table 2**, it is evident that at this number of variables, the central composite design and the D-optimal design are the most competitive options.

Table 1.Comparative results for different experimental design for 10 variables

	Full Factorial	Central Composite	D-Optimal Design	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
Experimental runs	59,049	158	71	71	71	71	71
MSE	2.2558	1.2351	0.0774	0.0902	0.0944	0.0697	0.0968
Estimated coefficients	66/66	66/66	66/66	66/66	66/66	66/66	66/66
Trace of $(X'X)^{-1}$	0.065	86.24	140.56	2854.7	670.1	1462.54	3366.64
Determinant of $(X'X)^{-1}$	0.00	6.33E-124	1.495E-96	3.65E-89	5.39E-66	3.59E-68	5.84E-63

Table 2.Comparative results for the coefficients estimation by the different experimental design for 10 variables

	Full Factorial	Central Composite	D-Optimal Design	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
±5%	66	52	42	9	22	10	12
(5%-10%]	0	6	8	13	18	17	14
(10%-15%]	0	2	5	11	5	15	10
(15%-20%]	**0**	1	3	8	2	10	6
>20%	**0**	5	8	25	19	14	24

Qualitatively speaking, the easiest options to generate (Random design, Mersenne Twister design, and LCG design) show low values of MSE; although, the cost seems to come in terms of coefficient variance. At a competitive number of runs and an adequate performance in coefficient variance, the proposed clustering design at this point seemed like it could be improved to become

a competitive option for larger numbers of variables. From running this comparison, it was experienced that both solving for the D-optimal design as well as carrying out the clustering procedure can be consuming in terms of computing resources. Devising a way to use a more efficient clustering procedure as well as to reduce the dependency on a complete enumeration as a starting point would help to importantly improve the proposed strategy.

Table 3 shows the results of the residual analysis and **Figure 3** show residual plots for all designs under comparison. Normality was assessed with the Kolmogorov-Smirnov test, and independence with the Signs test. Variance homogeneity was assessed graphically and by measuring the percentage of residuals falling within a distance of two standard deviations of the estimated mean. Regarding the residuals' normality test, the Random design, the Mersenne Twister design, and the LCG design showed varying degrees of deviation from normality, while independence did not seem a concern for any design.

Table 3.Comparative results of the residual analysis for different experimental design for 10 variables.

	Central Composite	**D-Optimal Design**	**Random Design**	**Clustering Design**	**Mersenne Twister Design**	**LCG Design**
P-value of Kolmogorov Smirnov	> 0.15	> 0.15	0.046	0.138	<0.010	<0.010
P-value of Signs	0.3010	0.4764	0.6350	0.8124	0.780	0.183
Standard deviation	1.1149	0.2802	0.3024	0.3094	0.2659	0.3133
$\mu - 2\sigma < \varepsilon < \mu + 2\sigma$	149/158; 94%	69/71; 97%	66/71, 93%	67/71; 94%	69/71; 97%	67/71; 94%

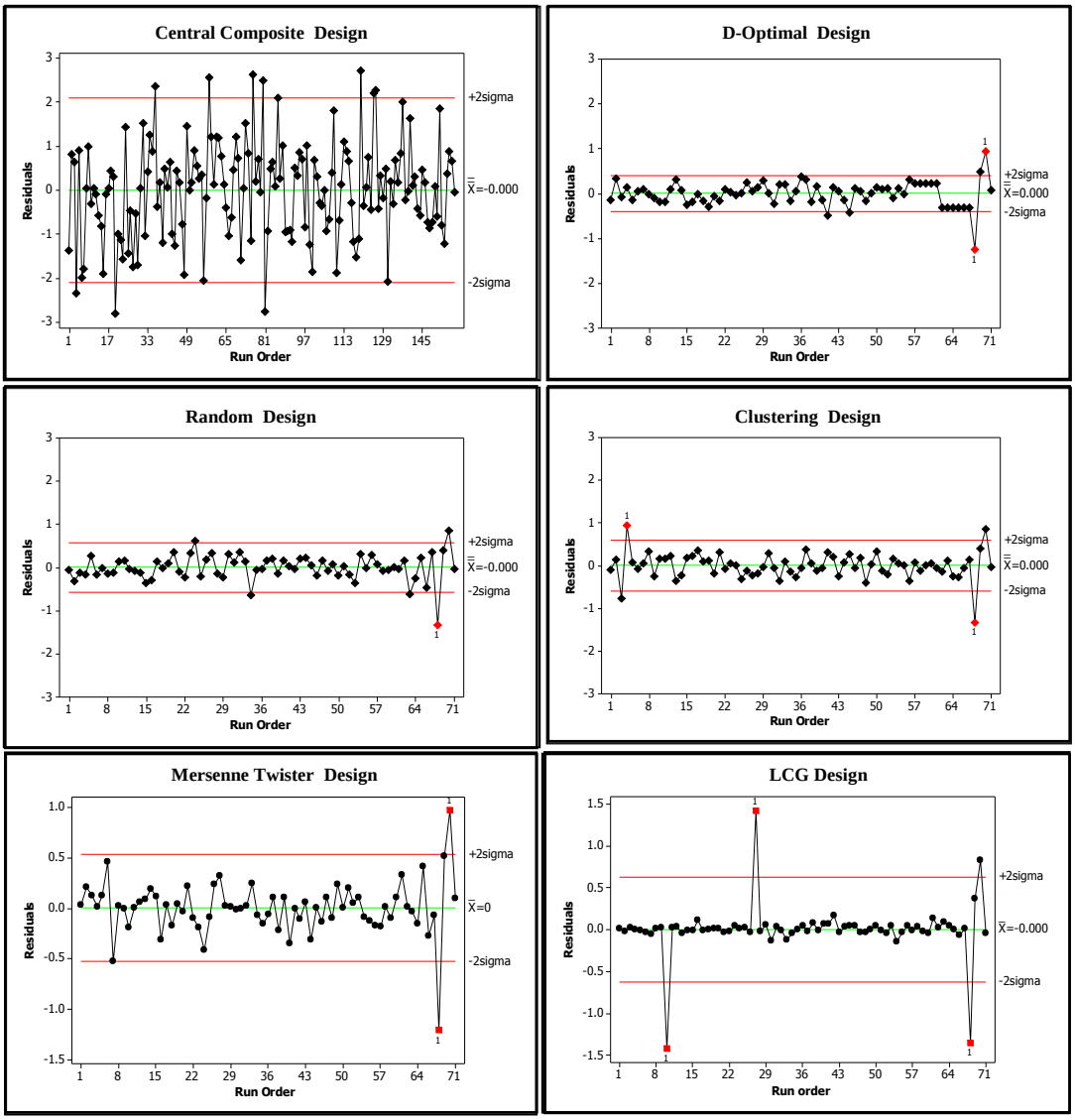

Figure 3. Plot of residuals in time sequence for 10 variables.

5.1.2 Experimental Designs for 20 Variables: Statistical Comparison Results

In this second set of results, the treatment of 20 variables was attempted. The modified version of the cluster design was included in this experiment. Also, the D-optimal design was generated in two ways: one with R-Project and an initial enumeration as in the modified clustering design, and the other with the commercially available statistical software JMP.

Table 4 summarizes the comparative results for M1–M5 and **Table 5** shows the distribution of the percentual deviation from the intended regression coefficients' values. In this case, the random design presented the lowest MSE and has the capability to estimate all coefficients, but the precision for the estimates of the coefficient is lower than the D-optimal design using R-Project and the commercial software JMP (**Table 4**). The clustering design has a competitive value of MSE, and has the capability to estimate all coefficients, but the precision for the estimates of the coefficient is less than the one obtained by the random design and the D-optimal design. The LCG design and the Mersenne Twister design have a competitive performance in terms of MSE; they are capable to estimate all regression coefficients but precision is still a challenge.

The full factorial and the central composite designs were not used in this comparison since at 20 variables, they are not practical. The Full Factorial Design, for 20 variables at three levels each, requires 3^{20} = 3,486,784,401 runs. The Central Composite design requires 1,048,617 experimental runs in its worst case. An important result is that of the D-Optimal paired with the shortened initial enumeration as proposed in this work it becomes feasible and is a competitive alternative for larger number of variables.

As in the previous case, a residual analysis was carried out. **Table 6** shows the results of the hypothesis tests and the assessment of the variance and **Figure 4** presents selected residual plots. The D-Optimal design (JMP) and the Mersenne Twister design showed some deviation in terms of normality. The D-Optimal design (R-Project) and the LCG Design showed problems with independence.

Table 4. Comparative results for different experimental design of 20 variables.

	D-Optimal (R-Project)	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
Experimental runs	232	232	232	232	232	232
MSE	0.0453	0.0108	0.0036	0.0142	0.0924	0.0383
Estimates coefficients	231/231	231/231	231/231	231/231	231/231	231/231
Trace of $(X'X)^{-1}$	5100.65	251.1	15867.87	50480.98	153902.2	13148.6

Table 5.Comparative results for the coefficients estimation by the different experimental design for 20 variables.

	D-Optimal (R-Project)	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
±5%	145	194	81	32	4	17
(5%-10%]	37	9	65	26	3	19
(10%-15%]	12	8	38	32	5	10
(15%-20%]	6	4	23	25	5	14
>20%	31	16	24	116	214	171

Table 6.Comparative results of the residual analysis for different experimental design for20 variables.

	D-Optimal (R-Project)	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
P-value of Kolmogorov Smirnov	0.133	0.027	> 0.15	> 0.15	<0.010	> 0.15
P-value of Signs	0.000	0.795	0.895	0.595	0.480	0.026
Standard deviation	0.2130	0.09910	0.0591	0.1195	62.81	0.1961
$\mu - 2\sigma < \varepsilon < \mu + 2\sigma$	221/232; 95%	221/232; 95%	202/232; 87%	220/232; 95%	221/232; 95%	222/232; 96%

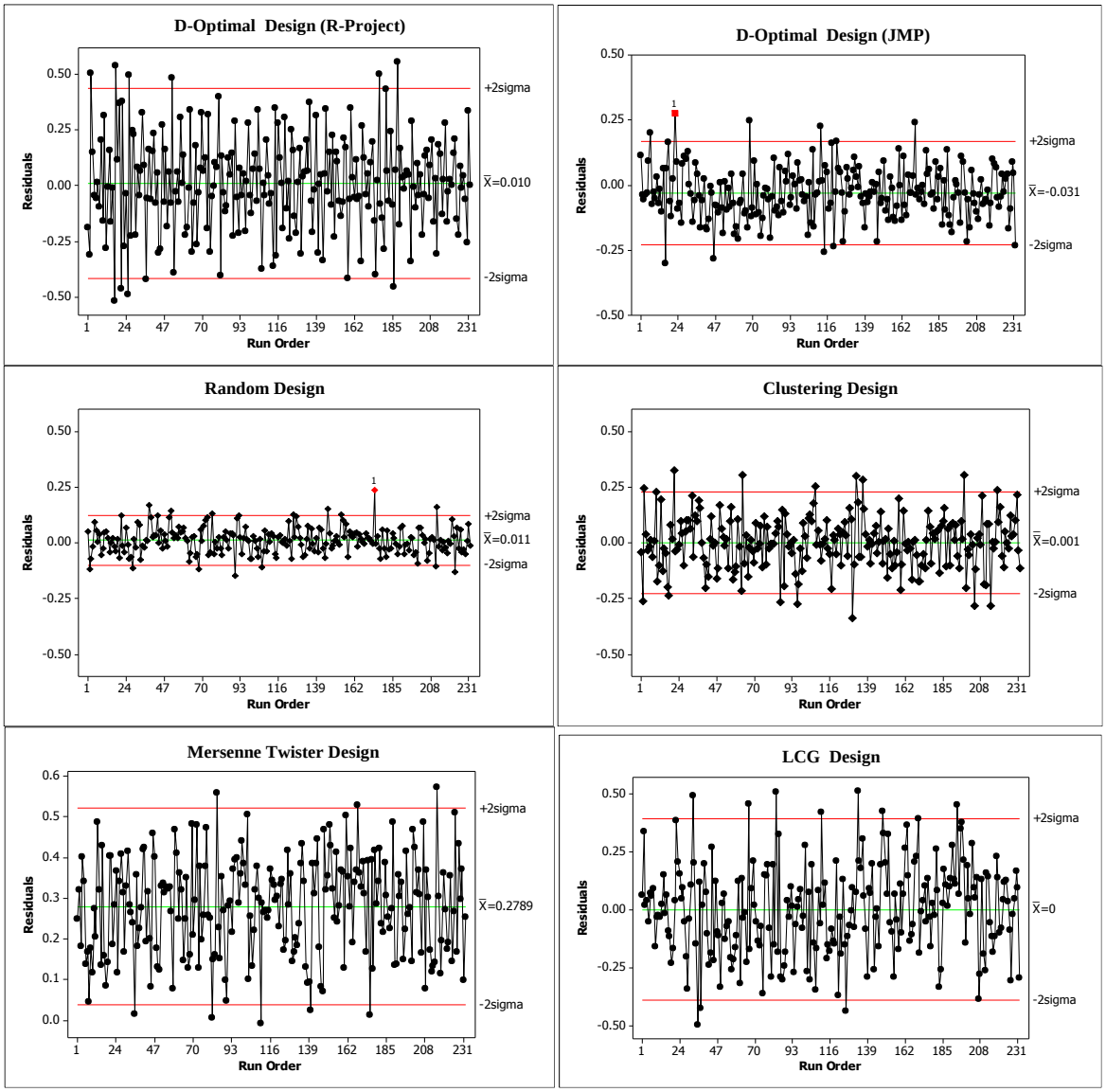

Figure 4. Plot of residuals in time sequence for 20 variables

5.1.3 Experimental Designs for 50 Variables: Statistical Comparison Results

In this third set of results, the treatment of 50 variables was attempted. **Table 7** summarizes the comparative results for M1–M5, and **Table 8** shows the distribution of the percentual deviation from the intended regression coefficients' values for experimental designs. For the development of experimental design for 50 variables, the modified version of the clustering method was used

in combination with the D-optimal design; however, it was not possible to complete the design due to lack of independence in the resulting enumeration, as detected by the software.

In this case, the random design presented the lowest MSE and had the capability to estimate all coefficients, but the precision for the estimates of the coefficient is lower than the D-optimal design using (JMP), MT, and the LCG Design (**Table 7**). The D-Optimal design (JMP) has an intermediate value of MSE, it has the capability to estimate all coefficients, and has the lowest value of the trace of $(XX)^{-1}$. It also has the best precision for the estimates of the coefficient.

Table 7. Comparative results for different experimental design of 50 variables.

	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
Experimental runs	1327	1327	1327	1327	1327
MSE	0.002	0.0002	8522329.079	0.006	0.001
Estimates coefficients	1326/1326	1326/1326	1273/1326	1326/1326	1326/1326
Trace of $(X'X)^{-1}$	382.55	104,474.89	0	121,595.42	36,945.58

Table 8.Comparative results for the coefficients estimation by the different experimental design for 50 variables.

	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
±5%	1255	211	352	389	338
(5%-10%]	26	206	98	311	285
(10%-15%]	14	206	17	228	213
(15%-20%]	3	172	4	151	187
>20%	28	531	802	247	303

As in the previous case, a residual analysis was carried out. **Table 9** shows the results of the hypothesis tests and the assessment of the variance and **Figure 5** show residual plots for all designs under comparison. The clustering design is the only that showed some deviation in terms of normality, while independence did not seem a concern for any design.

Table 9. C**omparative results of the residual analysis for different experimental design for 50 variables.**

	D-Optimal (JMP)	Random Design	Clustering Design	Mersenne Twister Design	LCG Design
P-value of Kolmogorov Smirnov	> 0.15	> 0.15	<0.010	> 0.15	>0.15
P-value of Signs	0.764	0.672	0.745	0.391	0.799
Standard deviation	0.0337	0.013	1392.76	0.078	0.032
$\mu - 2\sigma < \varepsilon < \mu + 2\sigma$	1276/1327; 96%	1269/1327; 96%	1285/1327; 97%	1271/1327; 96%	1257/1327; 95%

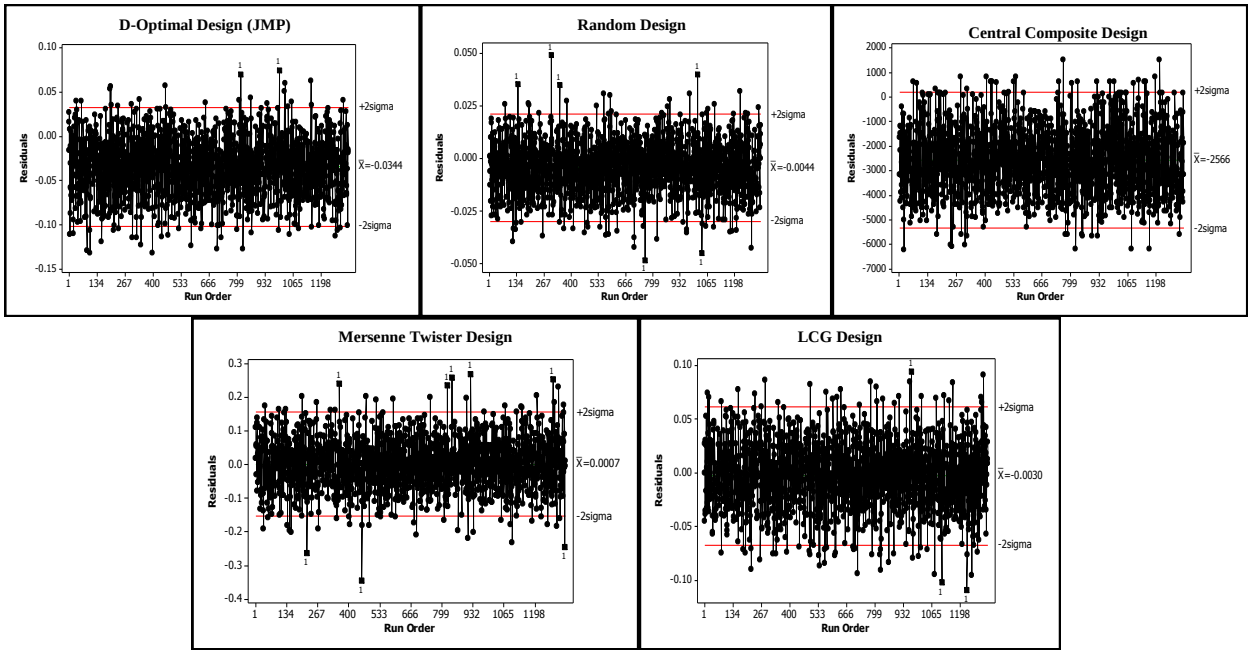

Figure 5. Plot of residuals in time sequence for 50 variables

5.2 Cost Approach

In this section a costing approach is proposed to compare the different strategies previously described in section 3 and 4 of this thesis. The identified strategies to generate experimental designs capable to analyze tens of variables at a time using a full quadratic regression model with the minimum number of necessary runs are shown in **Figure 6**, for 10, 20, and 50 variables. As the number of variables increases, many of these strategies become unfeasible. The designs for 50 variables are the focus of analysis here due to the potential they offer for system characterization, modeling, and optimization.

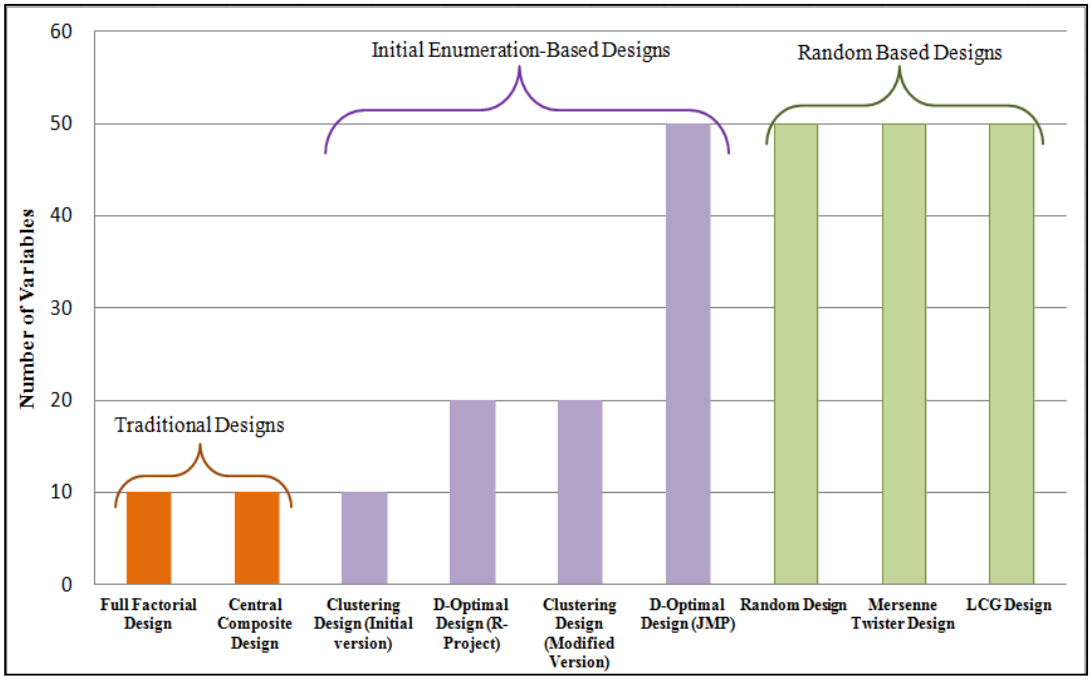

Figure 6. Resulting experimental designs for 10, 20 and 50 variables.

The costing approach was developed based on computing time and the cost associated to the purchase of necessary software to generate the design. The discussion at this point is limited to software available to the authors at the time of performing the comparison. The idea is to use

the cost model to evaluate other alternative combinations as they become available. Furthermore, the cost model includes computational time for accounting precision purposes. The resulting designs were classified in two categories: (1) those based in an initial enumeration and (2) those based on random processes. **Table 10** shows the estimated associated costs. It is important to note that, in the second category, it is possible to reduce the software cost to zero with the use of freely-distributed electronic spreadsheets, such as those included in LibreOffice and OpenOffice. For this analysis it was assumed that the necessary equipment (i.e. computer) to generate the designs is available, as well the necessary knowledge and the experience in the use of different statistical software and diverse methodologies for generating designs. Therefore the cost related to the acquisition of equipment and personnel training were not considered.

Table 10. Cost estimates for the generation of experimental design for 50 variables.

	D-Optimal Design	***Random Design***	***LCG Design***	***MT Design***
Software	$1,470	$1,495	$139.99	$139.99
Computational Time	10-15 minutes	0	0	20-25minutes
Computational Cost (1$/minute)	$15	0	0	$25
Total Cost	$1,485	$1,495	$139.99	$164.99

5.2.1 Initial Based Enumeration

In this category, at the 50 variables mark, the D-optimal design using JMP software is the only feasible alternative out of the strategies included in this study. To generate a design with this strategy, it is necessary to have specialized professional software, such as JMP in this case. The cost of the license of this software is $1,470 annually. The professional license has a cost of $14,900 annually **[35]**.

In terms of computing time, these designs are coded in the software, requiring that the user selects the number of variables and the levels for each of them, as well as the desired number of runs. The software internally runs the algorithm and presents the designs in a short-period of time. For 50 variables, the software generated the design in approximately 10-15 minutes.

For all designs it is assumed an arbitrary cost for the computational time of $1 per minute. Computational cost could be approached in the future through a parametrical study or based on models already found in the literature **[8, 30, 42]**.

Based on this assumption, the total cost associated to the generation of D-optimal design for 50 variables is $1,485 (**Table 10**). The designs were generated using a trial version of the software, and a computer with Intel CORE i5 processor, 64 bits. All designs were generated using the same computer.

5.2.2 Random Based Designs

This category contains the experimental designs based on random methods. These designs are discussed below.

Random Design

The generation of this design is, as its name indicates, completely random. It is not possible to control the design's statistical properties. In this method it is necessary to create a matrix with the levels of the variables to be investigated on the design and the probability for each level. The number of experimental runs is established by the user. This design was generated in this work using Minitab following a uniform discrete distribution. The perpetual license of this software has a cost of $1,495 without upgrades versions **[23]**.

The computing time for the generation of this design, in 50 variables, can be considered negligible because it only takes a few seconds. Therefore, the total cost associated to the generation of the random design is $1,495 (**Table 10**).

Random Walks Methods

Two methods were tried in this strategy. The first one generated the MT Design using R and MS Excel. R is an open access software, so it does not have a cost **[32]**. This is a well-known statistical software characterized by its computational capacity. The Microsoft Office suite that includes Excel, in its version "Office home and student 2013" has a cost of $139.99, while the "Office Home and Business 2013" has a cost of $219.99 **[22]**. Freely-distributed electronic spreadsheets, such as those included in LibreOffice and OpenOffice, could feasibly be used instead of further applications.

For this design the MT algorithm was coded in R **[31]**. The initial step is setting the seed to then generate vectors of magnitude *v*, where *v* represents the number of necessary regression coefficients to fit a full quadratic model plus one. These vectors were copied into Excel to be translated into practical levels. For this design k vectors were generated, where k represents the number of variables to be investigated. For each vector, it is necessary to establish a new seed. The computer time to generate each vector is negligible since it takes only a few seconds. The process of copying and adjusting each vector in Excel, with the translation to the selected levels, takes approximately 0.5 minutes. The expression of the total generation time is 0.5n. For 50 variables for example, the design generation time was calculated as 0.5(50) = 25 minutes approximately. Using this approximation, the total cost associated to the generation of this design is $164.99 (**Table 10**).

The second method was the LCG. To develop this design, it was necessary to use an electronic spreadsheet only. The associated cost for Excel was shown previously. For this design it is necessary to set the associated parameters in (4). The computer time is not significant for our analysis. The total cost associated to this design is $139.99 (**Table 10**).

6 Simulation Optimization Method

The capability of dealing with tens of variables simultaneously opens important analysis possibilities ranging from statistical characterization to optimization. To show this capability an illustrative example is presented in this section, where a simulation-optimization algorithm developed in our research group **[45]** was used.

This algorithm results in high quality solutions that can be achieved efficiently with a modest number of simulation runs. The algorithm starts with an initial design of experiments (DOE) from which an incumbent solution is obtained. In each iteration, a metamodel is obtained using the available set of points and is used to generate a new attractive point where a simulation is performed.

The simulated value of the new point is compared against the incumbent for updating purposes. A series of stopping criteria are evaluated and, if none is met, the new point is added to the existing set of points and a new iteration begins. Otherwise, the iteration stops. This algorithm is depicted in **Figure 7**, and a more detailed description is presented next.

Initialization:

1. Initial DOE: The initial DOE consists of n runs containing combinations of the v controllable variables of interest, $\mathbf{x}^i = (x_1, x_2, x_3, \ldots, x_v)^i$, as well as their evaluations $f(\mathbf{x}^i)$, where i=1,2,…,n. If a replicated DOE is used, the value of $f(\mathbf{x}^i)$ will be the average across the replicates.

2. Select incumbent*:* Considering a minimization instance, the DOE run with the minimum objective value is selected as the current best (incumbent) solution [$x_{k\text{-best}}$, $f(x_{k\text{-best}})$]. An iteration counter is initialized here at $k = 0$.

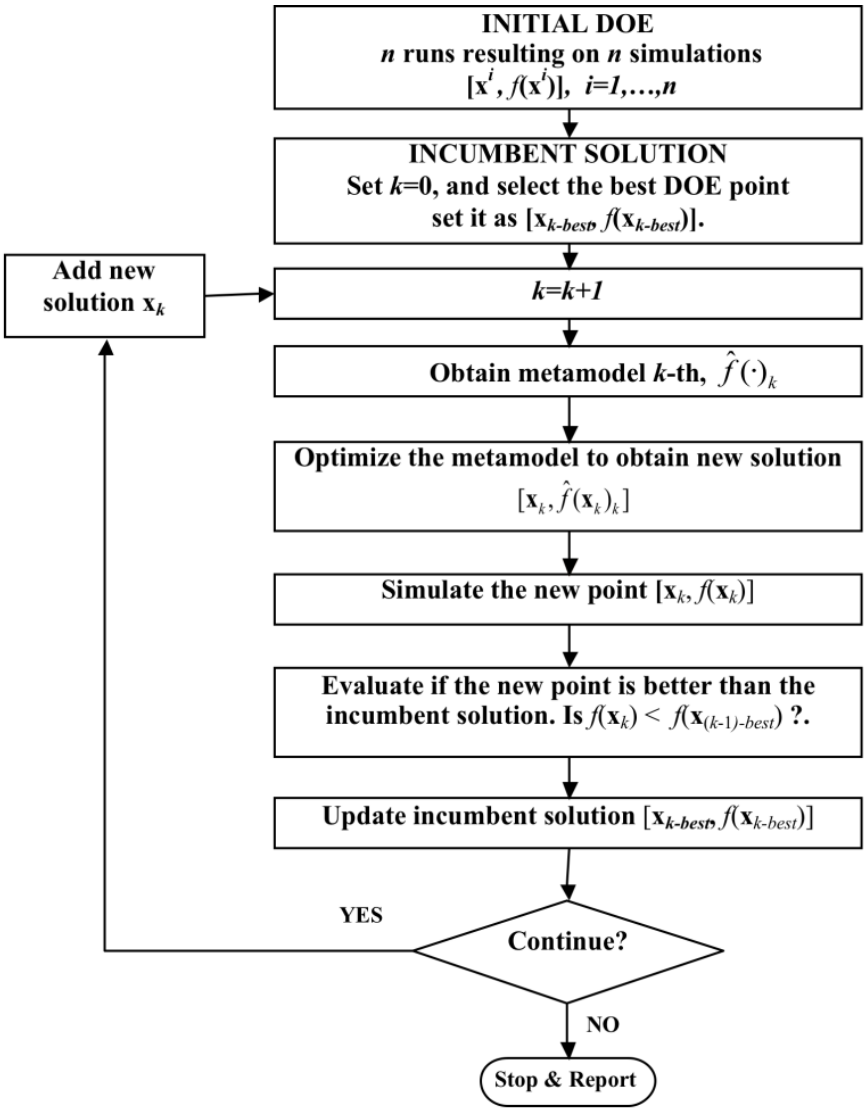

Figure 7.Simulation-based optimization algorithm.

Main Iteration:

3. Update counter: $k = k+1$
4. Obtain metamodel: Using the available points, build the k-th metamodel, $f(\cdot)_k$. In case of having only few variables, a saturated metamodel is preferred i.e. one that uses all available degrees of freedom, in this case a regression model with (n+k-1) coefficients.
5. Optimize metamodel: Using the metamodel as objective function in the optimization problem under analysis, a multiple-starting-points heuristic is used along with a local optimizer to obtain an attractive solution, $\mathbf{x}_k$.
6. Simulate the new point: Estimate, via simulation, the value of $f(\mathbf{x}_k)$ considering that if a replicated DOE was used, the same number of replicates is used for the new point and the mean value across them is reported.
7. Evaluate if the new point is better than the incumbent: In this case, evaluate if $\mathbf{x}_k$ has an objective value strictly lower than $\mathbf{x}_{(k\text{-}1)\text{-best}}$ i.e. if $f(\mathbf{x}_k) < f(\mathbf{x}_{(k\text{-}1)\text{-best}})$.
8. Update the incumbent: Update the incumbent according to the evaluation in the previous step. If $f(\mathbf{x}_k) < f(\mathbf{x}_{(k\text{-}1)\text{-best}})$, then the following is set $[\mathbf{x}_{k\text{-best}}, f(\mathbf{x}_{k\text{-best}})] := [\mathbf{x}_k, f(\mathbf{x}_k)]$, otherwise, the incumbent remains the same.
9. Evaluate the stopping criteria: Stop the algorithm if (1) $\mathbf{x}_k$ belongs to the initial DOE or is similar to any of the points generated on previous iterations; (2) if the coefficient of determination, $R^2 \geq \varepsilon$ (where ε is defined by the user); or (3) the maximum number of iterations has been reached. Both the ε and the maximum number of iterations are defined by the user.

If any of the stopping criteria is met, the method stops and the incumbent is reported as the final output. Otherwise, $\mathbf{x}_k$ and its simulated objective function value are added to the set of points available to build a new metamodel, and the main iteration is repeated. This algorithm has been empirically shown to converge in a moderate number of iterations even in the presence of several variables using global optimization test functions **[45]**.

6.1 Illustrative Example: Production Line with 50 Workstations

This example illustrates how a 50-variable simulation-optimization problem can be addressed aided by an experimental design with such capability. The strategies identified to generate experimental designs capable to analyze tens of variables at a time using a full quadratic regression model with the minimum number of necessary runs are shown in **Figure 6** for 10, 20, and 50 variables.

Consider a production line with 50 workstations simulated with the software package SIMIO. The simulation is run for 8 hours per day with 10 replicates. The simulation parameters of interest were the mean process time on each of the workstations (WSi). The process time in each workstation was assumed to follow a normal distribution with a mean that varied in three levels and a constant standard deviation of 0.25 minutes. It is further assumed that a particular user can choose the nominal process time. The response of interest was the system time defined as the period of time elapsed since a raw part to be processed enters the system until it exits as a finished product.

A simulation optimization method based on design of experiments and metamodeling techniques was used **[45]**. The method starts with an initial experimental design, which for 50

variables has 1327 experimental runs. **Figure 8** shows the ranges of values to be explored with the objective to minimize the system time per unit.

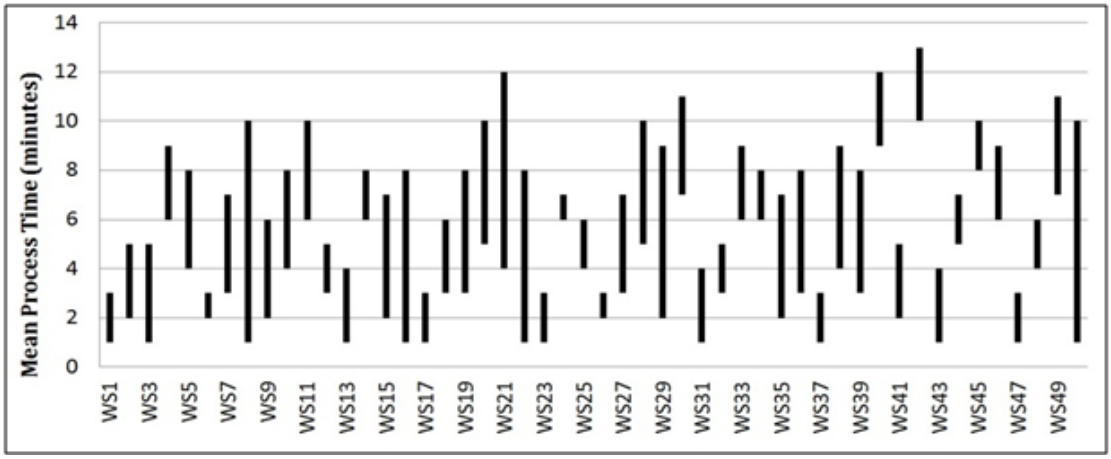

Figure 8. Range of values for the workstations' mean process time in simulation model.

The minimum value for the average cycle time in the experimental design was identified and selected as the first best solution (first incumbent solution) (I-1). I-1corresponded to a value of 312.09 minutes for the D-optimal design (JMP), 317.16 minutes for the LCG Design, 317.16 minutes for the Random design, and 316.82 minutes for the Mersenne Twister design (**Table 11**). With the initial experimental design, a full quadratic regression metamodel was built and used as the objective function to be minimized to obtain a predicted competitive solution. A generalized reduced gradient optimization procedure along with a multi-start strategy was used for this purpose.

Using the process times prescribed for each workstation by the first predicted competitive solution, a simulation was performed and the simulated values were compared with the incumbent solution (I-1) for updating purposes. Each iteration of the algorithm follows a similar structure until either a solution that has already been visited is predicted, or a user-defined maximum number of iterations are met. For this example, a maximum of 40 iterations was used. The algorithm was

stopped once it maxed out the allowed number of iterations. The best solution corresponded to a system time of 278.80 minutes for the D-optimal design (JMP), 304.72 minutes for the LCG design, 295.61 minutes for the Random design, and 303.5 minutes for the Mersenne Twister design (**Table 11**).

Table 11.System Time for the incumbent solution for the production line with 50 workstations.

D-Optimal (JMP)			***Random Design***			***Mersenne Twister Design***			***LCG Design***		
Run	I-j	System Time (minutes)	Run	I-j	System Time (minutes)	Run	I-j	System Time (minutes)	Run	I-j	System Time (minutes)
166	I-1	312.09	**1168**	I-1	317.16	**551**	I-1	316.82	**863**	I-1	317.16
1328	I-2	291.10	**1329**	I-2	305.51	**1332**	I-2	312.32	**1341**	I-2	312.91
1329	I-3	284.27	**1334**	I-3	298.73	**1334**	I-3	311.94	**1346**	I-3	307.43
1331	I-4	281.72	**1356**	I-4	295.61	**1336**	I-4	307.47	**1355**	I-4	307.37
1335	I-5	281.24	-	-	-	**1348**	I-5	304.74	**1361**	I-5	305.55
1338	I-6	280.11	-	-	-	**1354**	I-6	304.74	**1364**	I-6	304.72
1341	I-7	279.55	-	-	-	**1356**	I-7	303.50	-	-	-
1364	I-8	278.80	-	-	-	-	-	-	-	-	-

When a comparison between the initial incumbent solution (I-1) with the final one (I-4) was performed, the result was that the system time decreased in 33.09 minutes for the D-optimal design (JMP), 12 minutes using the LCG design, 21.5 minutes for the random design, and in 13.33 minutes for the Mersenne Twister design (**Figure 9**). This represents a reduction of 10.67%, 3.9%, 6.8%, and 4.21%, respectively, in the system time per unit in the simulated production system.

Although many aspects are interesting in this example, it is important to emphasize that it was possible to run this simulation-optimization procedure because there existed an experimental design capable to build a full quadratic regression model for 50 variables with a low number of runs. Appendix A and B contain the analyses for 10 and 20 variables.

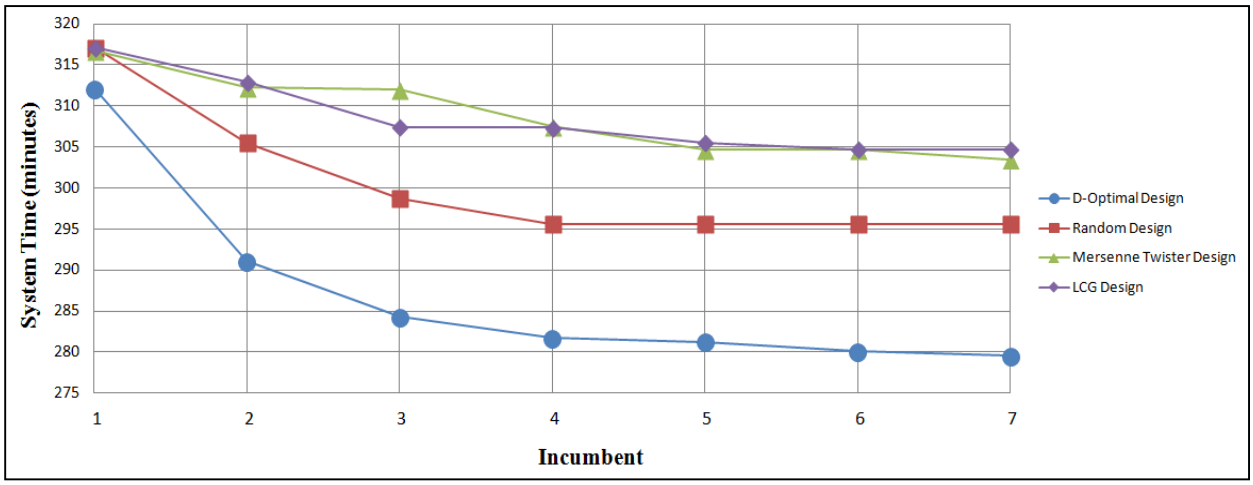

Figure 9.System time for the incumbent solutions of the simulation-optimization method.

7 Conclusion and Future Work

In this thesis the performance of different strategies to generate experimental designs is contrasted aiming to devise feasible options to explore tens of variables simultaneously in the future. An emphasis was made in using only a personal computer for the generation of the design. It was learned that a more efficient initial enumeration would improve the generation of the D-optimal design. It was also learned that the clustering design might be improved in terms of coefficient variance for it to be a competitor to the D-optimal design. Furthermore, at least the designs included in this preliminary comparison could be kept as benchmarks for future developments.

It was found that with 10 variables, the traditional design of experiment techniques such as the full factorial design and the central composite designs are the most competitive options. These designs, however, are already difficult to generate at 20 variables. Here is where computer generated designs such as the D-optimal design become competitive. At 50 variables, designs that require a large and well-crafted initial enumeration such as the D-optimal and the proposed clustering design become difficult to approach with a personal computer, although their generation is still possible and competitive. It is remarkable, however, that designs simple to generate such as the random design and the random walk- like methods become convenient options at such number of variables due to their overall feasibility. Further research on how to control the resulting designs from the latter seems promising in its own right.

An illustrative example with simulation optimization was used to show how important analysis are possible to do when having an experimental design that can be used to obtain a full quadratic model with the least possible number of experimental runs for tens of variables. This

encourages further research into the matter. In addition, it is envisioned that the designs resulting from this work be tested in real systems in the future.

Finally, the different experimental designs for 10, 20 and 50 variables and their assessment are made readily available online to those users interested in simulation-optimization based on experimental design. In this web page the designs, their statistical properties, the associated cost analysis, and the simulation-optimization example can be found. The website URL is: *http://yaileenmendez.wix.com/experimentaldesignlv*

Future work includes exploring cases with larger number of variables as well as improving the statistical properties of the random based designs to be inexpensive competitors to the D-Optimal Design. Furthermore, explore the possibility to address this design-generation process as a multiple criteria optimization problem, where a multicriteria evaluation of the designs previously generated will be developed.

A sequential experimentation generation via simulation-optimization currently is explored in our research group based on the proposed alternatives of experimental design identified in this work for 50 variables. The aim is to evaluate the possibility to converge to a best solution in an optimization problem with less number of experimental runs.

References

1. Alkhatib, M.F., Muyibi, S.A., and Amode, J. O. 2011. "Optimization of Activated Carbon Production from Empty Fruit Bunch Fibers in One-Step Steam Pyrolysis for Cadmium Removal from Aqueous Solution."*Environmentalist Journal* 31:349-357.
2. Anotai, J., Thuptimdang, P., Su, C.C., and Lu, M. C. 2012. "Degradation of O-Toluidine by Fluidized-Bed Fenton Process: Statistical and Kinetic Study." *Environmental Science and Pollution Research* 19:169-176.
3. Bell, C. M., Needham, M. D., and Szuster, B. W. 2011. "Congruence among encounters, norms, crowding, and management in a marine protected area." *Environmental Management, 48*(3), 499-513.
4. Cabrera-Ríos, M., Mount-Campbell, C.A., and Irani, S.A. 2002. "An Approach to the Design of a Manufacturing Cell under Economic Considerations." *International Journal of Production Economics* 78:223-237.
5. Christin, C., Smilde, A. K., Hoeflsoot, H.C., Suits, F., Bischoff, R., and Horvatovich, P.L. 2008. "Optimized Time Aligment Algorithm for LC-MS Data: Correlation Optimitized Warping Using Component Detection Algorithm-Selected Mass Chromatograms." *Analytical Chemistry* 80:7012-7021.
6. Chung, M., and Haber, E. 2012. "Experimental Design For Biological System." *SIAM Journal on Control and Optimization* 50:471-489.
7. Dixon, K.R. 2011. "Designing Simulation Experiments." *Modeling and Simulation in Ecotoxicology with Applications in Matlab and Simulink*. pp.147-158

8. Eboli, M. 2003. "Two Models of Information Costs Based on Computational Complexity." *Computational Economics*, 21:87-105.
9. Edwards, D. J., and Mee, R. W. 2011. "Fractional Box-Behnken Designs for One-Step Response Surface Methodology." *Journal of Quality Technology* 43:288-307
10. Gebhardt A. 2011. "Second Order Design." *Optimal Experimental Design with R*. Chapman and Hall. pp.263-278
11. Hassan Khani, M. 2011. "Statistical Analysis and Isotherm Study of Uranium Biosorption by Padina Sp. Algae Biomass." *Environmental Science and Pollution Research* 18:790-799.
12. Job, J., Sukumaran,R.K., and Jayachandranl, K. 2010. "Production of a Highly Glucose Tolerant B-Glucosidase by PaecilomycesvariotiiMG3: Optimization of Fermentation Conditions Using Plackett Burman and Box–Behnken Experimental Designs." *World Journal of Microbiology and Biotechnology* 26:1382-1391.
13. Laferriere, C., Ravenscroft, N., Wilson, S., Combrink, J., Gordon, L., and Petre, J. 2011. "Experimental Design to Optimize an Haemophilus Influenzae Type BConjugate Vaccine Made with Hydrazide-Derivatized Tetanus Toxoid." *Glycoconjugate Journal* 28:463-472.
14. Langner, H. W. 2003. "Genetic Algorithms for the Construction of D-Optimal Design." *Journal of Quality Technology* 35:28-46.
15. Larentis, A. L., Cunha Sampaio, H. C., Martins, O. B., Rodriguez, M. I., and Moitinho Alves, T. L. 2011. "Influence of Induction Conditions on the Expression of Carbazoledioxyfenase Components (CarAa, CarAc and CarAd) from Pseudomonas

Stutzeri in Recombinant Escherichia Coli Using Experimental Design."*Journal of Industrial Microbiology and Biotechnology* 38:1045-1054.

16. Lee, K.M., and Gilmore, D. F. 2006. "Statistical Experimental Design for Bioprocess Modeling and Optimization Analysis." *Applied Biochemistry and Biotechnology* 135:101-116.
17. Mahapatra S.S. 2009. "Modelling and Analysis of Erosion Wear Behavior of Hybrid Composites Using Taguchi Experimental Design." *Journal of Engineering Tribology* 224:157-168.
18. Martínez-cardeñas, J. A., de la Fuente-Salcido, N.,M., Salcedo-Hernández, R., Bideshi, D. K., and Barboza-corona, J. 2012. "Effects of physical culture parameters on bacteriocin production by mexican strains of bacillus thuringiensis after cellular induction." *Journal of Industrial Microbiology & Biotechnology, 39*(1), 183-9.
19. Marwa, H. A., Sammour, A., El-ghamryHanaa, A., and El-nahasHanan, M. 2011. "Optimizing Proniosomes for Controlled Release of Ketoprofen Using Box-Behnken Experimental Design." *International Journal of Pharmaceutical Sciences and Research* 2:2195-2205
20. Matsumoto, M., and T. Nishimura. 1998. "Mersenne Twister: A dimensionally equidistributed uniform pseudo-random number generator." *Journal ACM Transactions on Modeling and Computer Simulation* 8:3-30.
21. Meyer, R. K., and Nachtsheim, C.J. 1995."The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs." *Technometrics*, 37:60-69.
22. Microsoft Office Store. Accessed May 6, 2014.

http://www.microsoftstore.com/store/msusa/en_US/cat/Office/categoryID.62684700?Icid=Office_suites_redirect_021314.

23. Minitab Inc. Minitab 17 Pricing. Accessed May 6, 2014. http://www.minitab.com/en-us/products/minitab/pricing/
24. Montgomery D.C., and Runger, G. C. 2007. *Applied Statistics and Probability for Engineers.* 4th ed.New York: John Wiley & Sons, Inc.
25. Montgomery, D. C. 2009. *Designs and Analysis of Experiments*. 8th ed. New York: John Wiley & Sons, Inc.
26. Nair, V., Strecher, V., Fagerlin, A., Ubel, P., Resnicow, K., Murphy, S., Little, R., Chakraborty, B., and Zhang, A. 2008. "Screening Experiments and the Use of Fractional Factorial Designs in Behavioral Intervention Research." *American Journal Public Health* 98:1354-1359.
27. Nelofer, R., Ramanan, R. N., Zaliha, R. N., Basri, M., Ariff, A.B. 2012. "Comparison of the Estimation Capabilities of Response Surface Methodology and Artificial Neural Network for the Optimization of Recombinant Lipase Production by E. Coli BL21." *Journal of Industrial Microbiology and Biotechnology* 39:243-254.
28. Nobuyuki, R. M., Mello, P. S., Melo Santa Anna, L. M., and Pereira, N. 2010. "Nitrogen Source Optimization for Cellulase Production by Penicilliumfuniculosum, Using a Sequential Experimental Design Methodology and the Desirability Function." *Applied Biochemistry and Biotechnology* 161:411-422.
29. Rigas, F., Papadopoulou, K., Philippoussis, A., Papadopoulou, M., and Chatzipavlidis, J. 2009. "Bioremediation of Lindane Contaminated Soil by Pleurotusostreatus in Non Sterile

Condition Using Multilevel Factorial Design." *Journal of Water Air and Soil Pollution* 197: 121-129.

30. Rogers, R.O., and Skillicorn D.B. 1998. “Using the BSP cost model to optimise parallel neural network training,” *Future Generation Computer Systems*, 14: 409-424.
31. R Project for Statistical Computing. Package ‘rngSetSeed’. 2014. http://cran.r-project.org/web/packages/rngSetSeed/rngSetSeed.pdf.
32. R Project for Statistical Computing Package. Accessed May 6, 2014. http://www.r-project.org/
33. Sanchez, S. M., Lucas, T. W., Sanchez, P. J., Nannini, C. J., Wan, H. 2012 "Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security."*Design and Analysis of Experiments: Special Designs and Applications*, Vol. 3 (ed K. Hinkelmann), John Wiley & Sons, Inc., Hoboken, NJ, USA.
34. Sanchez, S. M., and Sanchez, P. J. 2005. "Very Large Fractional Factorial and Central Composite Designs." *ACM Transactions on Modeling and Computer Simulation* 15:362-377.
35. SAS Institute Inc. JMP Statistical Discovery From SAS. Accessed May 6, 2014. http://word.tips.net/Pages/T000273_Numbering_Equations.html.
36. Sayara, T., Sarra, M., and Sanchez, A. 2010. "Optimization and Enhancement of Soil Bioremediation by Composting Using the Experimental Design Technique." *Biodegradation Journal* 21:345-356.
37. Senthilkumaran, K., Pandey, P. M., and Rao, P. V. M. 2012. "Statistical Modeling and Minimization of Form Error in SLS Prototyping." *Rapid Prototyping Journal* 18:38-48.

38. Sowmya, R., Rathinaraj, K., and Sachindra, N. M. 2011. "An autolytic process for recovery of antioxidant activity rich carotenoprotein from shrimp heads." *Marine Biotechnology*, *13*(5), 918-927

39. Splendore, R., Dotti, F., Cravello, B., and Ferri, A. 2011. "Thermo-physiological comfort of a PES fabric with incorporated activated carbon." *International Journal of Clothing Science and Technology*, *23*(5), 283-293.

40. Sudhankar, P., and Nagarajan, P. 2011. "Optimization of Chitinase Production Using Statistics Based Experimental Designs."*Journal of Chemical, Biological and Physical Sciences* 40:352-362.

41. Sudheer Kumar, Y., Varakumar, S., and Reddy, O.V. 2010. "Production and Optimization of Polygalacturonase from Mango (Mangiferaindica L.) Peel Using Fusariummoniliformein Solid State Fermentation." *World Journal of Microbiology and Biotechnology* 26:1973-1980.

42. Turchenko, V., and Grandinetti, L. Investigation of Computational Cost Model of MLP Parallel Batch Training Algorithm. In Smposium on Industrial Electronics and Applications (ISIEA 2009).

43. Vaithanomsat, P., Songpim, M., Malapant, T., Kosugi, A., Thanapase,W., and Mori, Y. 2011. "Production of β-Glucosidase from a Newly Isolated Aspergillus Species Using Response Surface Methodology." *International Journal of Microbiology* 2011:1-9.

44. Vieira, H. Jr., Sanchez, S., Kienitz, K. H., and Neyra Belderrain, C.N. 2011. "Generating and Improving Orthogonal Designs by Using mixed Integer Programming." *European Journal of Operational Research* 215:629-638

45. Villarreal-Marroquín, M. G., Castro, J. M., Chacón-Modragón, O. L., and Cabrera-Ríos, M. 2013. "Optimisation Via Simulation: AMetamodelling-Based Method and a Case Study." *European J. Industrial Engineering* 7:275-294.
46. Vishwantha, K.S., AppuRao, A. G., and Singh, S. A. 2010. "Acid Protease Production by Solid-State Fermentation Using Aspergillusoryzae MTCC 5341: optimization of process parameters." *Journal of Industrial Microbiology and Biotechnology* 37:129-138.
47. Wass, J. A. 2011. "A Further Step in Experimental Design (III): The Response Surface." *Journal of Validation Technology* 17:54-62.
48. Zambare, V. 2011. "Optimization of Amylase Production from Bacillus Sp. Using Statistics Based Experimental Design." *Emirates Journal of Food and Agriculture* 23:37-47.
49. Zhou, W. W., He, Y.L., Niu, T. G., and Zhong, J. J. 2010. "Optimization of Fermentation Condition for Production of Anti-TMV Extracellular Ribonuclease by Bacillus Cereus Using Response Surface Methodology." *Bioprocess and Biosystems Engineering* 33:657-663.

Publications:

Refereed Journals:

- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., Pérez-Candelario, H., and Cabrera-Ríos M. 2014. "Affordable experimental design with tens of variables." *Production & Manufacturing Research: An Open Access Journal* 2:1, 658-673. DOI: 10.1080/21693277.2014.956903

- Rodríguez-Yañez, A.B., **Méndez-Vázquez, Y.M.**, Cabrera-Ríos, M. 2014 "Simulation-based process windows simultaneously considering two and three conflicting criteria in injection molding." *Production & Manufacturing Research: An Open Access Journal* 2:1, 603-623. DOI: 10.1080/21693277.2014.949359

Refereed Conference papers:

- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., Pérez-Candelario, H., and Cabrera-Ríos M. "Enabling Simheuristics Through Designs for Tens of Variables: Costing Models and Online Availability." *Proceeding of the 2014 Winter Simulation Conference* (Savannah, GA, December 2014).

- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., Pérez-Candelario, H., and Cabrera-Ríos M. "The Search for Experimental Design with Tens of Variables: Preliminary Results." *Proceeding of the 2013 Winter Simulation Conference* (Washington DC, December 2013).

- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., and Cabrera-Ríos M. "The Search of Experimental Design with Tens of Variables." ISERC, Institute of Industrial Engineers, (Puerto Rico, May 2013).

Non-Refereed Conference Proceedings (Abstract Only):

- Pérez-Candelario, H., **Méndez-Vázquez, Y.M.**, Ramírez, K.L., and Cabrera-Ríos, M. "Testing Different Strategies for Large Experimental Designs." 2014. Industrial & Systems Engineering Research Conference, Montréal, Canada, 2014
- Pérez Candelario, H., **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, KL., and Cabrera-Ríos, M. "The Search for Experimental Designs with Tens of Variables." 2014. UPRM Research Simposium, Mayagüez PR, May 2014
- **Méndez-Vázquez, Y.M.**, Ramírez Rojas, K.L., Pérez-Candelario, H., and Cabrera-Ríos, M. "Designs for tens of variables." 2nd Annual Symposium: Science & Technology, Innovation Through Research, Mayagüez PR, April 2014
- **Méndez-Vázquez, Y.M.**, Pérez Candelario, H., Ramírez Rojas, K.L., and Cabrera-Ríos, M. "The Search for Experimental Designs with Tens of Variables." 2014 Sigma Xi Science Day, Mayagüez PR, April 2014
- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., and Cabrera-Ríos, M. "The Search for Experimental Designs with Tens of Variables." 2013 Simposio de Investigación RUM, Mayagüez PR, May (2013).

- **Méndez-Vázquez, Y.M.**, Ramírez-Rojas, K.L., and Cabrera-Ríos, M. "The Search for Experimental Designs with Tens of Variables." 2013 SACNAS COHEMIS Ciencia & Tecnología, Innovación a través de la Investigación, Mayagüez PR, April (2013).

Appendix A

Illustrative Example: Production Line with 10 Workstations

This section deals with the treatment of 10 variables, the initial case of study in this thesis. Consider a production line with 10 workstations simulated with the software package SIMIO where the simulation model is illustrated in **Figure A.1**. The simulation is run for 8 hours per day with 10 replicates. The parameters were the mean process time on each of the workstations (*WSi*), which assumed to follow a normal distribution with a mean that varied in three levels (**Figure A.2**) and a constant standard deviation of 0.25 minutes.

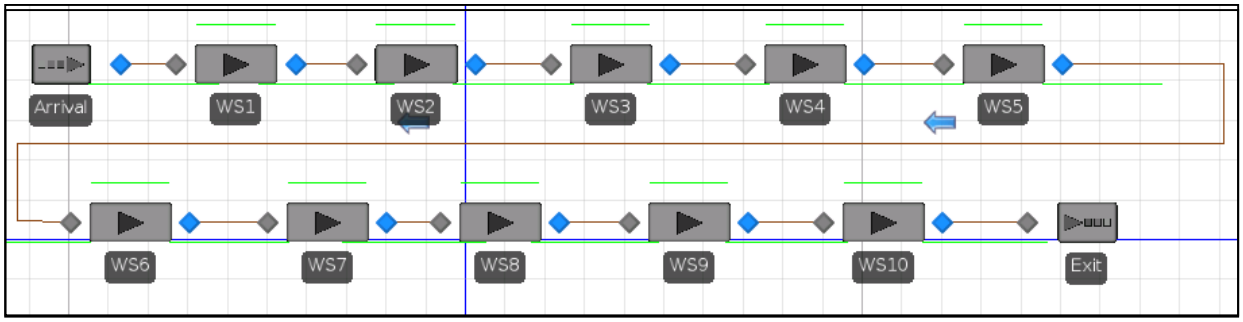

Figure A.1. Simulation model for a production line with 10 workstations.

For this example, the simulation optimization method described in **[45]** is used along with our original experimental design for 10 variables (**Figure 6**). The method starts with an initial experimental design, which for 10 variables has 71 experimental runs. The maximum number of iterations of the algorithm was 40 if the other stopping criteria were not met before **[45]**.

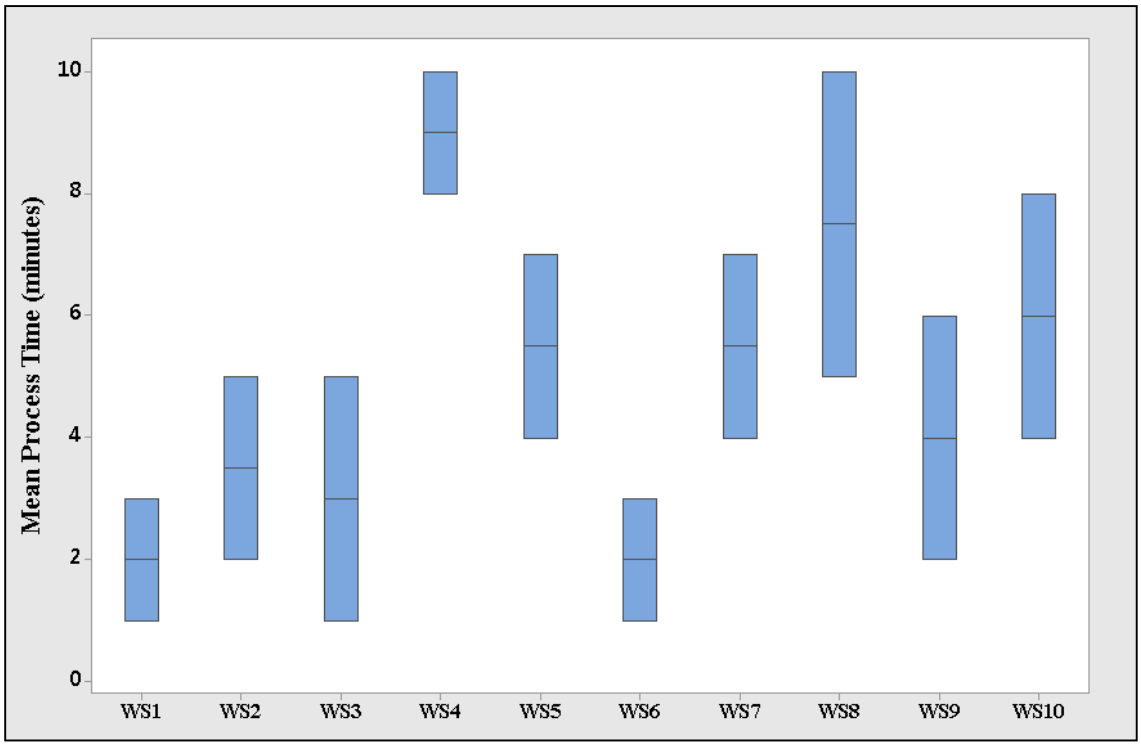

Figure A.2. Range of values for the workstations' mean process time in simulation model with 10 variables.

The minimum value for the average cycle time selected as the first best solution (first incumbent solution) (I-1), is identified from the initial experimental design and corresponded to a value of 176.36 minutes for the Random design, 173.02 minutes for the Mersenne Twister Design, 172.06 minutes for the Clustering Design, 170.59 minutes for the LCG Design, and 170.06 minutes for the D-Optimal (**Table A.1**). With the initial experimental design, a full quadratic regression metamodel was built and used as the objective function to be minimized to obtain a predicted competitive solution.

The algorithm followed the same structure as described in Section 6 of this thesis. In this example, the best solution when the algorithm stopped corresponded to a system time of 167.39 minutes for the Random Design, 165.41 minutes for the Mersenne Twister Design, 166.36 minutes

for the Clustering Design, 165.17 minutes for the LCG Design, and 167.02 minutes for the D-Optimal (**Table A.1**).

When a comparison between the initial incumbent solution (I-1) with the final one (I-9) was performed, the result was that the system time decreased in 8.97 minutes for the Random Design, 7.61 minutes for the Mersenne Twister Design, 5.70 minutes for the Clustering Design, 5.42 minutes for the LCG Design, and 3.04 minutes for the D-Optimal (**Figure A.3**). This represents a reduction of 5.1, 4.4, 3.3, 3.2, 1.8%, respectively, in the system time per unit in the simulated production system.

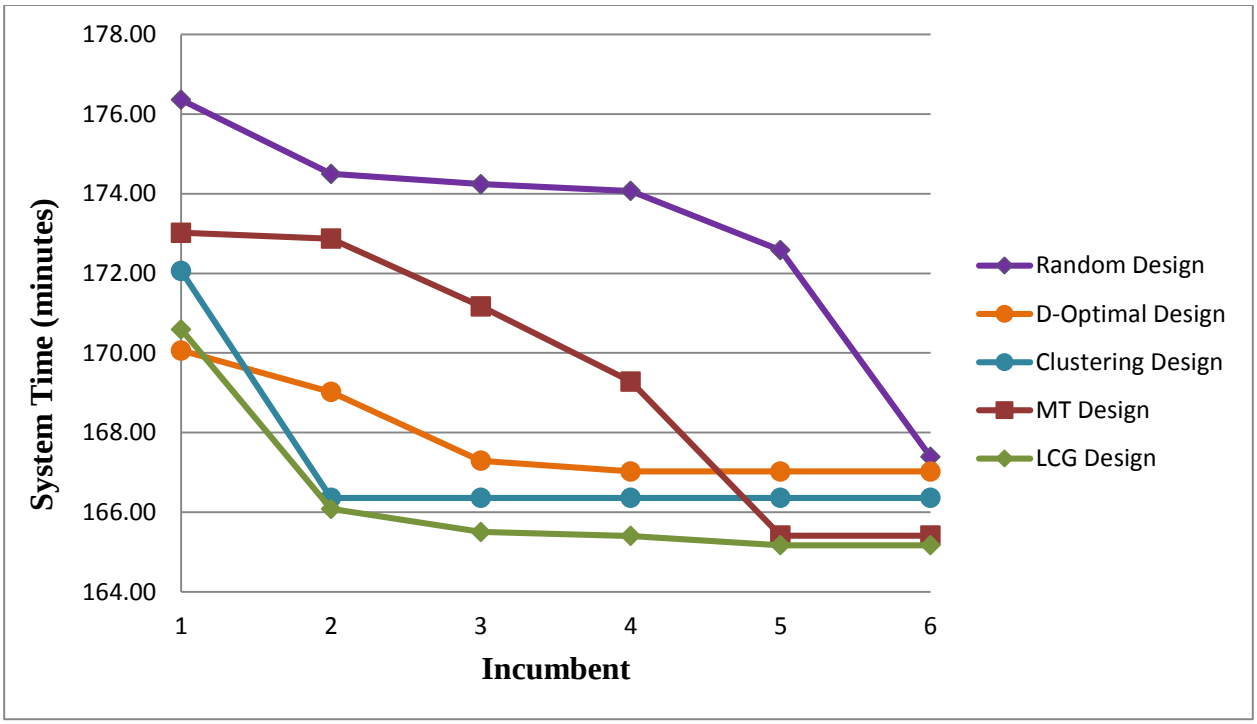

Figure A.3. System time for the incumbent solutions of the simulation-optimization method for 10 variables.

Table A.1. System Time for the incumbent solution for the production line with 10 workstations.

D-Optimal Design			***Clustering Design (Initial Version)***			***Random Design***		
Run	I-J	System Time (minutes)	Run	I-J	System Time (minutes)	Run	I-j	System Time (minutes)
55	I-1	170.06	**13**	I-1	172.06	**49**	I-1	176.36
72	I-2	169.02	**72**	I-2	166.36	**72**	I-2	174.50
80	I-3	167.29	-	-	-	**74**	I-3	174.24
86	I-4	167.02	-	-	-	**77**	I-3	174.04
-	-	-	-	-	-	**86**	I-4	172.58
-	-	-	-	-	-	**88**	I-5	167.39

Mersenne Twister (MT) Design			**LCG Design**		
Run	I-J	System Time (minutes)	Run	I-j	System Time (minutes)
66	I-1	173.02	**7**	I-1	170.59
83	I-2	172.87	**87**	I-2	166.08
84	I-3	171.87	**95**	I-3	165.50
85	I-4	169.28	**102**	I-4	165.40
87	I-5	165.41	**103**	I-5	165.17

Appendix B

Illustrative Example: Production Line with 20 Workstations

The simulation optimization example with 20 variables is presented here. A production line with 20 workstations is considered where the simulation model is illustrated in **Figure B.1**. The simulation is run for 8 h per day with 10 replicates, and the parameters were the mean process time on each of the workstations (*WSi*), which assumed to follow a normal distribution with a mean that varied in three levels (**Figure B.2**) and a constant standard deviation of 0.25 minutes.

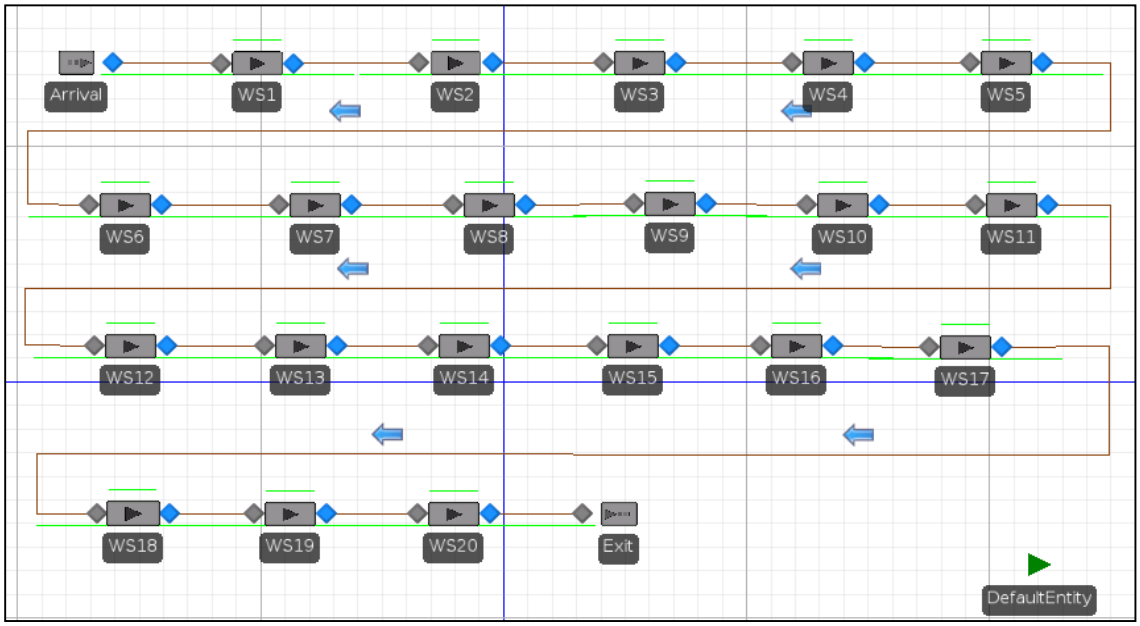

Figure B.1. Simulation model for a production line with 20 workstations.

For this example, the simulation optimization method described in **[45]** is used along with our own experimental design for 20 variables (**Figure 6**). The method starts with an initial experimental design for 20 variables with 232 experimental runs.

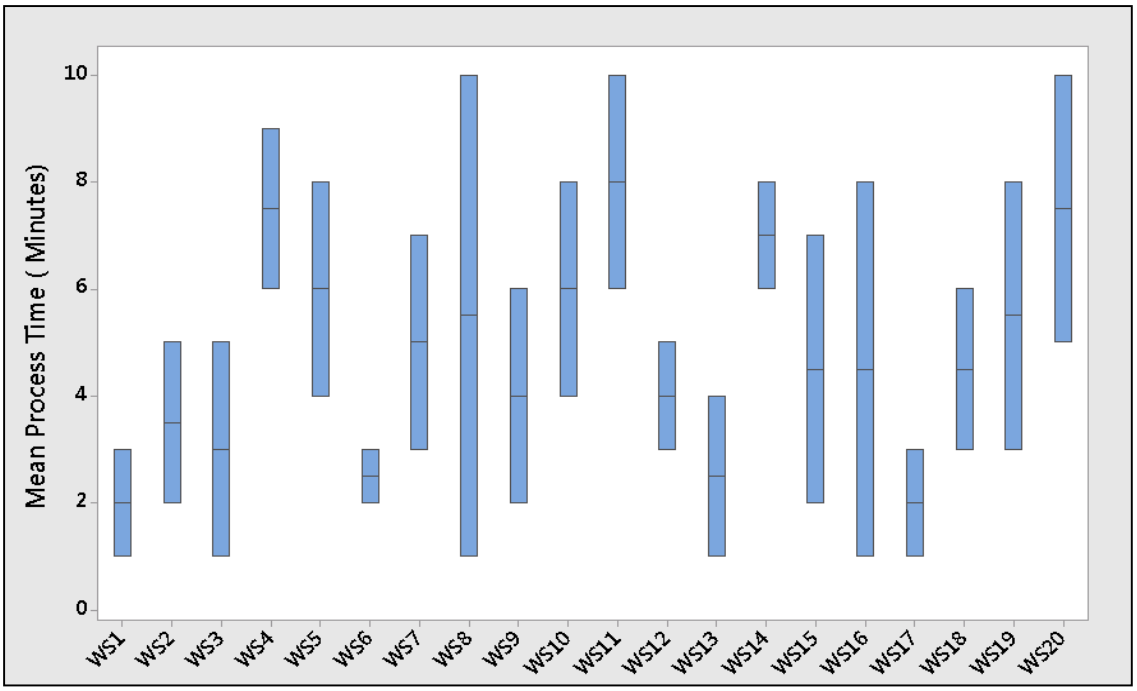

Figure B.2. Range of values for the workstations' mean process time in simulation model with 20 variables.

The minimum value for the average cycle time selected as the first best solution (first incumbent solution) (I-1), is identified from the initial experimental design and corresponded to a value of 197.7 minutes for the LCG Design, 194.5 minutes for the Mersenne Twister Design, 197.2 minutes for the D-Optimal Design (R-Project), 195.2 minutes for the Random Design, 197.2 minutes for the Clustering Design using the modified version, and 199.5 for the D-optimal design (JMP) (**Table B.1**). With the initial experimental design, a full quadratic regression metamodel was built, and used as the objective function to be minimized to obtain a predicted competitive solution.

The algorithm followed the same structure as described in Section 6 of this thesis. In this example, the best solution when the algorithm stopped corresponded to a system time of 175.4 minutes for the LCG Design, 181.7 minutes for the Mersenne Twister Design, 157.5 minutes for

the D-Optimal Design (R-Project), 153.5 minutes for the Random Design, 158.2 minutes for the Clustering Design using the modified version, and 162.1 for the D-optimal design (JMP) (**Table B.1**).

Table B.1. System Time for the incumbent solution for the production line with 20 workstations.

D-Optimal Design (JMP)			***Clustering Design (Modified Version)***			***Random Design***		
Run	I-J	System Time (minutes)	Run	I-J	System Time (minutes)	Run	I-j	System Time (minutes)
58	I-1	199.5	**184**	I-1	197.2	**11**	I-1	195.2
233	I-2	176.6	**235**	I-2	174.1	**24**	I-2	188.3
251	I-3	166.6	**236**	I-3	167.7	**24**	I-3	174.1
252	I-4	166.0	**237**	I-4	160.0	**25**	I-4	168.8
253	I-5	163.7	**263**	I-5	158.2	**25**	I-5	167.7
255	I-6	162.1	-	-	-	**26**	I-6	162.3
-	-	-	-	-	-	**26**	I-7	157.2
-	-	-	-	-	-	**27**	I-8	154.9
-	-	-	-	-	-	**27**	I-9	153.5

D-Optimal Design (R-Project)			**Mersenne Twister Design**			**LCG Design**		
Run	I-J	System Time (minutes)	Run	I-J	System Time (minutes)	Run	I-j	System Time (minutes)
184	I-1	197.2	**12**	I-1	194.5	**108**	I-1	197.7
233	I-2	188.2	**24**	I-2	188.1	**234**	I-2	195.8
235	I-3	169.9	**24**	I-3	185.3	**243**	I-3	194.4
237	I-4	164.4	**25**	I-4	182.4	**254**	I-4	193.8
241	I-5	161.4	**26**	I-5	182.1	**259**	I-5	184.7
245	I-6	157.5	**27**	I-6	181.7	**266**	I-6	180.10
-	-	-	-	-	-	**269**	I-7	180.09
-	-	-	-	-	-	**270**	I-8	176.7
-	-	-	-	-	-	**272**	I-9	175.4

When a comparison between the initial incumbent solution (I-1) with the final one (I-9) was performed, the result was that the system time decreased in 22.3 minutes for the LCG Design, 12.8 minutes for the Mersenne Twister Design, 39.7 minutes for the D-Optimal Design (R-Project),

41.7 minutes for the Random Design, 39 minutes for the Clustering Design using the modified version, and 37.4 for the D-optimal design (JMP) (**Figure B.3**). This represents a reduction of 11.3, 6.6, 20.1, 21.4, 19.8 and 18.7%, respectively, in the system time per unit in the simulated production system.

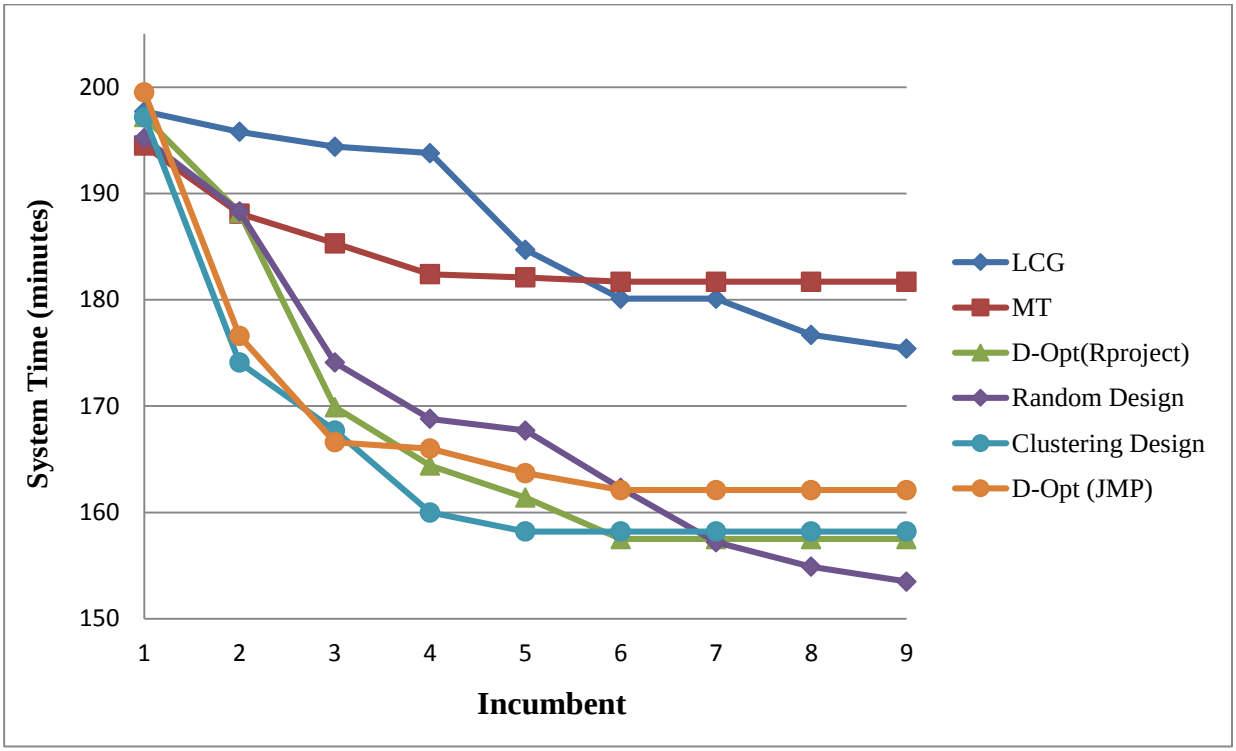

Figure B.3. System time for the incumbent solutions of the simulation-optimization method for 20 variables.