International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:2, No:9, 2008

Entropy Based Spatial Design: A genetic Algorithm Approach (Case Study)

Abbas Siefi, Mohammad Javad Karimifar

***Abstract*—**We study the spatial design of experiment and we want to select a most informative subset, having prespecified size, from a set of correlated random variables. The problem arises in many applied domains, such as meteorology, environmental statistics, and statistical geology. In these applications, observations can be collected at different locations and possibly at different times. In spatial design, when the design region and the set of interest are discrete then the covariance matrix completely describe any objective function and our goal is to choose a feasible design that minimizes the resulting uncertainty. The problem is recast as that of maximizing the determinant of the covariance matrix of the chosen subset. This problem is NP-hard. For using these designs in computer experiments, in many cases, the design space is very large and it's not possible to calculate the exact optimal solution.

Heuristic optimization methods can discover efficient experiment designs in situations where traditional designs cannot be applied, exchange methods are ineffective and exact solution not possible. We developed a GA algorithm to take advantage of the exploratory power of this algorithm. The successful application of this method is demonstrated in large design space.

We consider a real case of design of experiment. In our problem, design space is very large and for solving the problem, we used proposed GA algorithm.

***Keywords*—** Spatial design of Experiments, Maximum Entropy Sampling, Computer experiments, Genetic Algorithm.

I. Introduction

A. Design of Experiments

EXPERIMENT design is an activity that seeks to select experiments that are optimal in some sense. Scientist and engineers perform experiments to increase their understanding of a particular phenomenon. Design of experiments (DOE) has been of both great theoretical and practical interest. Generally, design generation methods can be divided in two families (1) model related designs and (2) model free designs. Obvious example for model related designs is construction of approximate model of a system. Construction of an approximate model (Meta-modeling) and calibration of deterministic computer model are examples of model free designs. In this article we focus on model free design methods for design generation. Because of considering computer models instead of physical phenomena, we use the term "computer experiment" for model free designs.

A. Seifi., is with Amirkabir University of Technology, he is now with the Associate Professor in Department of Industrial Engineering, 424 Hafez Avenue, Tehran, Iran (phone:0098+21-66413034 Fax:0098+21-66413025 e-mail*: aseifi@aut.ac.ir*).

M. J. Karimifar., is with Amirkabir University of Technology, he is now Master student in Department of Industrial Engineering, 424 Hafez Avenue, Tehran, Iran (e-mail: Karimifar.ie@gmail.com).

B. Spatial design of experiments

Sampling theory and *optimum experimental design theory* are two large branches in theoretical statistics that have developed separately, though with considerable theoretical overlap, both of them providing methods for efficient site positioning. Whereas *sampling theory* is a basically model-free methodology essentially oriented towards restoring unobserved data, in *optimum design theory* the aim is to estimate the structure of the data generating process, e.g. the parameters of an assumed (regression) model or functions of these parameters (Müller [1]). In this paper emphasis is on the first branch.

Spatial statistics is the collection of statistical methods in which spatial locations play an explicit role in the analysis of data. One of the characteristic features of geostatistical problems is: data consist of responses Y_i associated with locations xi which may be non-stochastic, specified by the sampling design, or stochastic but selected independently of the process $Y(x)$ in principle, Y could be determined from any location x within a continuous or discrete spatial region A. Spatial data occur in many fields such as agriculture, geology, environmental sciences, and economics. They have been recorded and analyzed probably as early as men started to make maps, however the origins of their statistical analysis as we understand it today must probably be attributed to the work of Matheron [2]. Spatial data has the distinctive characteristic that, attached to every observation, we have a set of coordinates that identifies the (geographical) position of a respective data collection site. The set of locations of those data collection sites influences decisively the quality of the results of the statistical analysis. Usually in choosing the design the aim is to ensure continuous monitoring of a data generating process or to allow for point prediction of present or future states of nature. Cox *et al.* [3] have listed current and future issues in this research area.
Spatial sampling is equivalent to take observations in a predefined area. Observations may be obtained by means of measurements in the field, or on samples taken to laboratories.

Spatial studies focusing on environmental, ecological and agricultural phenomena require a proper and carefully designed strategy for collecting data. Data can be difficult or expensive to collect, and both the sampling design and the quality of the data may affect the final qualities of an analysis (Cochran [4]; Müller [1]).

The main reason to use statistical sampling schemes is that such sampling guarantees scientific objectivity; if the same area is resampled, perhaps in the future, the results will be comparable in a well-defined sense (Stuart [5]).

The basis of statistical sampling is the interpretation of differences and similarities between two or more measurements (Wollum [6]; Wendroth et al. [7]). Drawing inference from a single measurement is highly risky. There is little reason to assume that another observation in the same area will resemble in any way the first observation. Drawing inference from two observations in the same area will exhibit some of the variation to be expected; the range of the two collected observations falls within the range of all possible observations. Taking a third, a fourth observation, etc., gives increasingly more information of the area under study, like an expression for the mean μ, the standard deviation σ and the standard error of the mean σe. It is decided by means of the objective of the study and the amount of prior information how many observations one should take. Stein and Ettema [8] focused on how to determine in advance the number of observations to be taken, as well as their observation locations. Generally, both depend on the actual purpose with which the sampling is carried out in their spatial context.

Several considerations can be made to choose the optimal design. In practice, this usually depends upon the amount of prior information, like available data, boundaries on the area and priority setting by the researcher.

The focus of this paper is primarily on the statistical aspects associated with selecting an appropriate spatial sample.

As with most large-scale sample design problems, the central challenge is how to allocate sampling resources across space (and time) to maximize the information available, which can then be used to make reliable and credible inferences or predictions about the response(s) of interest. Of course, as with any large-scale monitoring program, it is essential that the sampling or statistical design and the response (or operational) design are considered simultaneously to meet stringent resource constraints.

In this paper our emphasis is on the sampling or statistical design, and in particular on the spatial aspects of that design, which we call the spatial design.

There are four popular schools of thought for choosing the spatial design that three of them are statistically-based philosophies [9]:

1. Geometric approaches are typically based on heuristic arguments and include regular lattices, triangular networks, or space-filling designs. These approaches are typically used for exploratory purposes e.g. Muller [10].
2. Probability-based approaches select sites via a probability sample and use survey sample methods to make inferences about the population of interest or some characteristic of it.
3. Optimum experimental design or Model-based approaches base the inference about the target population on an explicit specification of the relationship between the selected sites and the population in terms of a statistical model (aims at parameter estimation, based upon information matrix). Model-based design can also include implicit model selection methods such as targeted sampling, representative sites and convenience sampling, and will not be discussed further in this paper.

And another one is based upon probabilistic structure:

4. Information theoretic approach (usually an entropy criterion).

There is considerable discussion of the contrasts between model-based and design-based inference in the literature (e.g. [11]; [12]). Brus & De Gruitjer [13, 14] and De Gruitjer & Ter Braak [15] focus on the differences in spatial inference.

As a compromise between these two broad strategies, Cressie et al [16] discuss design for data collection in ecological studies from a statistical modeling perspective. They show how probability-based sampling designs can be incorporated into statistical models, resulting in what is termed model-assisted design based inferences. The design necessary for making reliable spatial predictions in a region may be quite different to the design required to report on a distributional quantity such as the mean for a region. It is essential that there is clarity of purpose if we want to make informed decisions about the spatial design.

Geometric approaches consider how well a set of design points covers the domain. There is no dependence on the spatial covariance or the stochastic model. The design criterion is based on geometry and the distance between both current and potential sample locations. Royle & Nychka [17] and Nychka & Salzman [18] describe space-filling designs.

Probability-based designs assume a fixed underlying process and use probability sampling to select the monitoring sites. This contrasts with model-based designs, where the stochastic element is embedded in the model process. The use of probability sampling is critical to design-based inference. A probability sampling design for an explicitly-defined resource population is a means to certify that the data collected are free from any selection bias, conscious or not. A probability sampling design has three distinguishing features [9]:

• The population being sampled is explicitly described;
• Every element in the population has the opportunity to be sampled with known probability; and
• The selection is carried out by a process that includes an explicit random element.

These features provide mathematical foundations for statistical inference. Randomization is particularly important as it avoids bias and ensures the sample is representative.

Consider a problem of finding optimal sampling designs for dependent spatial data. Random selection of units can be extremely inefficient, since it doesn't take into consideration spatial nature of the locations which can be strongly correlated. Systematic selection is inflexible to irregular features of the space, such as stratification, inhomogeneous variances and anisotropy. In principle, we would like to model the process by a spatial random field, incorporate prior knowledge and select the best subset of points of desired cardinality to best represent the field in question. The motivation is a need to interpolate the observed behavior of a process at unobserved locations, as well as to design a network of optimal observation locations which allows accurate

representation of the process, Such goals are especially important in geophysics, meteorology and environmental sciences, since it is usually costly, unfeasible or impossible to sample the entire area.

Selecting the best subset translates to optimizing the criterion function, the cost of selecting that subset. The choice of cost function largely depends on the objectives. One of the common choices is maximum entropy sampling.

Selecting the best subset can be computationally intensive, since there is $\binom{N}{S}$ possible candidates, where N is the number of all possible sampling locations and S is the cardinality of the desired subset. Alternative strategies search for the solution close to the optimal. One of the most common strategies is sequential or forward selection, where we look for the next best point to be included until we get the set of the desired cardinality [2]. Configurations are nested, which is generally a false assumption, so the solution is only approximate. Local searches try to surpass this problem by searching for a better solution in the neighbourhood. The result is not necessarily globally optimal.

The MES principle allows us to find optimal design of experiments. This result is particularly valuable when the experimenter is asked to design experiments based on a small number of observations.

Spatial dependence is playing a crucial role for constructing spatial design. Of particular significance is the Gaussian case, for which, without considering the effect of the deterministic trend defined by the mean, the stochastic spatial dependence is determined by the covariance structure, which can be derived from a specific model or represented by an empirical function. This problem can be formulated differently depending upon the situation and, of course, on the purpose. The design problem is to find a set of sampling locations (optimum under some specific criterion) either observing x or some related variable (random field) Y, with or without assuming any restrictions, and with or without considering any prior sample or model information. Different approaches have been introduced in the literature (see, for example, De Gruijter and Ter Braak [19]). Cressie [20] summarizes the main aspects of the general geostatistical approach. A more random-field focused formulation can be found in Christakos [21]. Using an information theory approach, Caselton and Hussian [22] propose the choice of a network which maximizes the entropy of the random variables at gauged sites. Along the same lines, Caselton and Zidek [23] consider the problem in a Bayesian framework, formulating it as a decision problem. Their optimal choice maximizes the information in the random variables at gauged sites on the random variables at ungauged sites. Caselton et al. [24] assume that the random vector depends on a parameter with a prior distribution and their purpose is also to reduce uncertainty about this parameter. Thus, they select stations to be observed that minimize the residual uncertainty. In the Gaussian case, Ko et al. [25] provide an upper bound for the entropy and develop an exact algorithm based on this bound for solving the design problem.

C. Computer Experiments

Computer modeling is having a profound effect on scientific research. Many processes are so complex that physical experimentation is too time consuming or too expensive; or, as in the case of weather modeling, physical experiments may simply be impossible. As a result, experimenters have increasingly turned to mathematical models to simulate these complex systems. Advances in computational power have allowed both greater complexity and more extensive use of such models. Virtually every area of science and technology is affected. Computer models (or codes) often have high dimensional inputs, which can be scalars or functions. The output may also be multivariate. In particular, it is common for the output to be a time-dependent function from which a number of summary responses are extracted. In the design of complex systems, computer experiments are frequently the only practical approach to obtaining a solution. Typically, a simulation model of system performance is constructed based on knowledge of how the system operates. If a performance measure is not straightforward to calculate, such as one that involves an integral, then *sampling* via computer experiments may be employed to estimate the measure. If the simulation model is computationally expensive, then the optimization may instead rely on a *metamodel*, i.e., a mathematical model surrogate of system performance, to approximate the relationship between system performance and the design parameters. In metamodeling, there are two basic tasks that must be conducted: (i) select a set of sample points in the design parameter space (i.e., an experimental design); and (ii) fit statistical model(s) to the sample points. Our focus in this article is on the first task. Methods for the first task may be used to conduct sampling in general.

Making a number of runs at various input configurations is what we call a computer experiment. The design problem is the choice of inputs for efficient analysis of the data. The computer models we address in this article are deterministic; replicate observations from running the code with the same inputs will be identical. It is this lack of random error that makes computer experiments different from physical experiments, calling for distinct techniques. When the simulation of the system is stochastic, then we may consider performance measures that involve expected values (means) and/or variances. Repeated runs at the same settings may be used to mitigate the effects of random noise in experimental outcomes.

Computer experimenter, like the physical experimenter, can have many purposes in mind. We see three primary objectives:

- Predict the response at untried inputs.
- Optimize a functional of the response.
- Tune the computer code to physical data.

These objectives prompt basic statistical questions:

- The design problem: At which input "sites" $S = \{S_1 \ldots S_N\}$ should data y (S_1)... y (S_N) be collected?
- The analysis problem: How should the data is used to meet the objective?

In each of computer simulation, the user must specify the values of some governing variables. The deterministic

computer experiments differ substantially from the physical experiments performed by agricultural and biological scientists of the early 20th century. Their experiments had substantial random error due to variability in the experimental units. Relatively simple models were often successful.

Apparently, McKay, Conover and Beckman [33] were the first to explicitly consider experimental design for deterministic computer codes. They introduced Latin hypercube sampling, an extension of stratified sampling which ensures that each of the input variables has all portions of its range represented. Latin hypercube are computationally cheap to generate and can cope with many input variables. Iman and Helton [34] compared Latin hypercube sampling with Monte Carlo sampling of a response surface replacement for the computer model. Despite some similarities to physical experiments, then, the lack of random (or replication) error leads to important distinctions. Lest the reader wonder whether statistics has any role here, we assert that:

- The selection of inputs at which to run a computer code is still an experimental design problem. Statistical principles and attitudes to data analysis are helpful however the data are generated.
- There is uncertainty associated with predictions from fitted models, and the quantification of uncertainty is a statistical problem.
- Modeling a computer code as if it were a realization of a stochastic process, the approach taken below, gives a basis for the quantification of uncertainty and a statistical framework for design and analysis.

Selecting an experimental design is a key issue in building an efficient and informative model. The design of deterministic computer experiments has been partly addressed in the literature. For example, Sacks and Ylvisaker [26, 27], Welch [28] and references mentioned therein have considered nonparametric systematic departures from regression models. For the most part, however, the designs used for fitting predictors have been those developed for physical experiments. Such designs typically have appealing features of symmetry and are often optimal in one or more senses in settings f which include random noise.

II. DEFINITION

A. Problem Definition

We consider experimental situations in which we wish to make statistical inferences regarding a set of random variables from observations of a subset of these variables. In practice, the variables may be dispersed over space and/or time. For example, the random variables may correspond to potential observations of meteorological or environmental monitoring stations. Other examples occur in statistical geology where the observations may be collected at different points in space. Maintaining and operating all possible observation points or stations is costly, and one may want to select only a subset of them. In such circumstances, it may also be required, for a variety of other scientific, historical, or political reasons, that certain specified points be included in the resulting subset. We study the problem of selecting a "best" such subset with specified size. Formally, we are given a set **N** of **n** points, called the design space, and a design size s, such that $s \leq n$. Our goal is to choose a set S of s points satisfying $S \subset N$, called a feasible design, such that observations taken at these points will be as valuable as possible, and our goal is to choose a feasible design that minimizes the resulting uncertainty. To measure this uncertainty, we associate with the design space N a symmetric positive definite n x n matrix A, for example, a covariance matrix. Then the entropy associated with any s-element subset S of N is the logarithm of the determinant of the s × s principal sub matrix A[S] with row and column indices in S. A Ds-optimal design is a feasible design that has maximum entropy. As the logarithm is an increasing function, it is an equivalent problem to maximize det (A[S]) among the set of feasible designs.

Among the numerous applications of Ds-optimal designs (see, e.g., Mitchell [29], the papers collected in Dodge, Fedorov and Wynn [30], and the references therein), the optimal design of spatial sampling networks has received considerable recent attention (e.g., Shewry and Wynn [31], Fedorov and Hackl [46], the references therein, and those below).

(Ko, C.-W et al. [25]) show that our problem is NP-hard, implying that it is unlikely that an efficient algorithm can be found to optimally solve all instances of our problem. The methods used by statisticians for finding Ds-optimal designs consist of complete enumeration and heuristics, based mostly on exchanges. The best known of these is probably the DETMAX method of Mitchell [29]; also see the above references for related methods. A greedy constructive algorithm is sometimes used to construct an initial solution (e.g., Guttorp et al. [32]). Ko, C.-W et al. [25] also introduce a robust methods for producing truly optimal solutions for instances of moderate size.

Complete enumeration may be used when n and s are not too large.

B. Definition and Notation

We are given a set N of n points, Let N = (1, 2... n), where n is a positive integer. Throughout, A denotes a real symmetric positive definite matrix with rows and columns indexed by **N.** Hence $A_{ij} = A_{ji}$ for all $i, j \in N$, and $x^t Ax \rangle 0$ for all $x \in R^n$. For an s-element subset S of N ($1 \leq S \leq n$), Let A[S] denote the principal sub matrix of A having rows and columns indexed by Y. Similarly, we write $A[S, S]$ to denote the sub matrix of A having rows indexed by S_1, and columns indexed by S_2. We note that A[S, S] = A[S], and the symmetry of A implies that the transpose of A $[S_1, S_2]$ is A $[S_2, S_1]$. Throughout, we omit the braces around the element of single-element sets. Hence, A [i, j] denotes A_{ij}. We write det (A[S]) to denote the determinant of A[S]. The properties of A imply that det (A[S]) is always nonnegative. Our optimization problem P (max, ||) is to determine v (A, s) = max det (A[S]), $S \subset N$.

Any s-element subset S of N satisfying $S \subset N$ is called a feasible solution.

III. CASE STUDY

The considered problem deals with one of oil tanks in south of the country and the purpose is to find some points of the oil tank from which have the maximum amount of output.

There is a simulator for this tank and the response of running in a special point of tank shows the amount of output oil in that point. This simulator is deterministic. Due to long time of consuming the simulation, the research group wishes to replace simulator by a surrogate model. To construct the model and to achieve this goal, one of the initial steps is to find the points of experimental design. In other hand, we seek for some points at which simulator can be run and achieved results can be used to construct approximate model (it is noted that this problem is in the field of computer experiments).in this paper, we focus on constructing proper experimental design for tank problem, and cases include the way of constructing surrogate model and its optimization are not surveyed here. The specifications of tank are as big as a field about 45 length and 37 width in a two dimension coordinate in which each coordinate (x, y) indicates one point of tank and each point of tank includes 95 blocks. In fact, the tank can be considered as a cube having 45×37×95 blocks (in real coordinate, each block has a dimension includes 200m length, 200m width and 1m height).By receiving the point of view of relevant professionals and making concise surveys,, effective factors on outputs oil were specified in every block of tank, included porosity and saturation of water.

As it mentioned, the final aim is fitting the best model and optimizing it to find the points having the maximum amount of output oil (Qacc); therefore it is important that the fitting model properly fits points with high Qacc (with no regard to how points with low Qacc are fitted). Thus, points of experimental design must be nominated among those points which have high Qacc. For tank an indicator is defined and called HIP that indicates the value of Hydrocarbon in each block of tank. Specification of this indicator implies that if Qacc is high in a point of tank, value of HIP will be high too. By using this indicator a noticeable survey on Qacc of every block will be possible.

HIP is defined as:

$$HIP = V_B \times (1 - S_W \times P_O) \quad (1)$$

In above equation, V_B is equal to the considered volume of every block and due to same volume of all blocks, we give V_B =1. Also P_O is porosity and S_W is saturation of water that the value of them changes between (0%, 100%).

According to former remarks, each point of tank is a statistical society including 95 blocks and the value of Qacc correlate together in different points of tank. Thus for construction of design of experiment, the correlation between points must be considered together. According to previous section, to do this, we need matrix of covariance between points of tank. By using of existing data of every block, we computed value of HIP for every block and then formed matrix of covariance for points of tank. As it mentioned, considered tank (Design space) includes 45×37 points. Our goal is to choose a set of s=25 points from the tank (the value of s=25 is selected according to several considerations), such observations taken at these points will be as valuable as possible, and our goal is to choose a feasible design that minimizes the uncertainty of results. To measure this uncertainty, we associate with the design space of 1665, a symmetric positive definite 1665 × 1665 matrix A, a matrix of covariance. Then the entropy associated with any n-element subset S of N is the logarithm of the determinant of the $S \times S$ principal sub matrix A [S] with row and column indices in S. As the logarithm is an increasing function, maximizing entropy design is equivalent to maximizing det ($A[S = 25]$) among the set of feasible designs. Thus our design of experiment is a spatial design and uses the Maximum Entropy Sampling (MES) for construction of design.

The maximum-entropy sampling problem is NP-Hard and exact algorithms are applied to compute a maximum-entropy design by using the "branch and bound" framework with upper bounds calculated by a variety of methods (Ko, C.-W et al. [25], Anstreicher et al. [35, 36], Lee [37–39], and Hoffman et al. [40]). Also Lee and Williams [41] introduced a new upper bound as the solution of a linear integer programming. Their bound depends on a partition of the underlying set of random variables. Also Anstreicher and Lee introduced a new "masked spectral bound" for the problem.However our case contains a very big design space (N=1665) and it is impossible to find an optimal solution for large scale problems. Thus for solving the problem, we developed a genetic algorithm that it will be surveyed in next section.

IV. ALGORITHMIC PROCESS

A. Description

A genetic algorithm (GA) is an evolutionary search strategy based on simplified rules of biological population genetics and theories of evolution.

– A GA maintains a population of candidate solutions using a sampling procedure to select the solutions that seem to work well for the problem (that is, optimizing an objective function).

– After this selection process, the most fit candidate solutions are combined or altered by "reproduction" operators to produce new solutions for the next generation.

– The process continues, with each generation evolving more fit solutions until an acceptable solution has evolved. (Introduction to GAs: Michalewicz and Zbigniew [49] and Haupt and Haupt [50]).

Terminology:

- A chromosome represents a potential solution to the problem of interest that will be represented by a string of encoded genes.

– Genes can be either binary encoded (0 or 1) or real-number encoded.

– Davis [44] found that GAs using real number representations outperformed GAs with binary representations in numerical optimization problems. This is also the opinion of Haupt and Haupt [43].

– Because real number representation (Goldberg [45], Davis [44], Michalewicz [42]) works effectively on mathematical optimization problems and allows for the use of numerical reproduction operators, it will be used in this research.

- The objective function *F* measures a chromosome's fitness as a solution and is the function we wish to optimize. *F* takes a chromosome as input and outputs a fitness value.

– GAs are attractive because they are relatively easy to implement and, mathematically, they do not require a differentiable objective function thereby reducing the chance of reporting local optima.

In computer experiment area, because of using softwares for solving questions, we can design the experiment with very large design space and there is an ability to consider the questions with a high volume.Because of high volume of experiment points in such questions and the inefficiency for solving these questions with accurate methods, the need for innovative methods with a tolerable speed and capability in giving a result to the near optimum is affirmed. We offer an "Adaptive Genetic algorithm" for solving such questions.In our problem the total number of points is N (The design space consists of N points**)** and matrix A is an n×n matrix which indicates the covariance between these points. The purpose is to find n points out of N existing points that the n×n sub matrix, resultant these n points, has the maximum determinant out of all n×n possible sub matrixes.

B. The structure of algorithm

The structure of our genetic algorithm is represented by Fig. 1.

C. Initial population

A chromosome represents an experimental design that number of genes for a chromosome is equal to n (number of points for experiment) and each gene represent a point that experiment can be do in that point and value of each gene can be choose in range of (1,N). (A gene represents a point of design space). Therefore each chromosome is a 1×n matrix.
Generally, evolutionary algorithms generate the initial population at random. This initial population helps the algorithm to be representative from any area of search space. In this paper we generate *pop_size* chromosome using the Uniform random number.

D. Chromosomes evaluation

To determine the objective function for chromosomes in each generation, after constructing the covariance matrix for each chromosome, we calculate the determinant of this matrix as objective function. This covariance matrix is a sub matrix of matrix ***A*** and represents covariance between pints of each chromosomes (dimension of this matrix is $n \times n$ that ***n*** is length of chromosome or number of points of experiment). For determining the fitness of chromosomes in each generation, we used the proportional function.

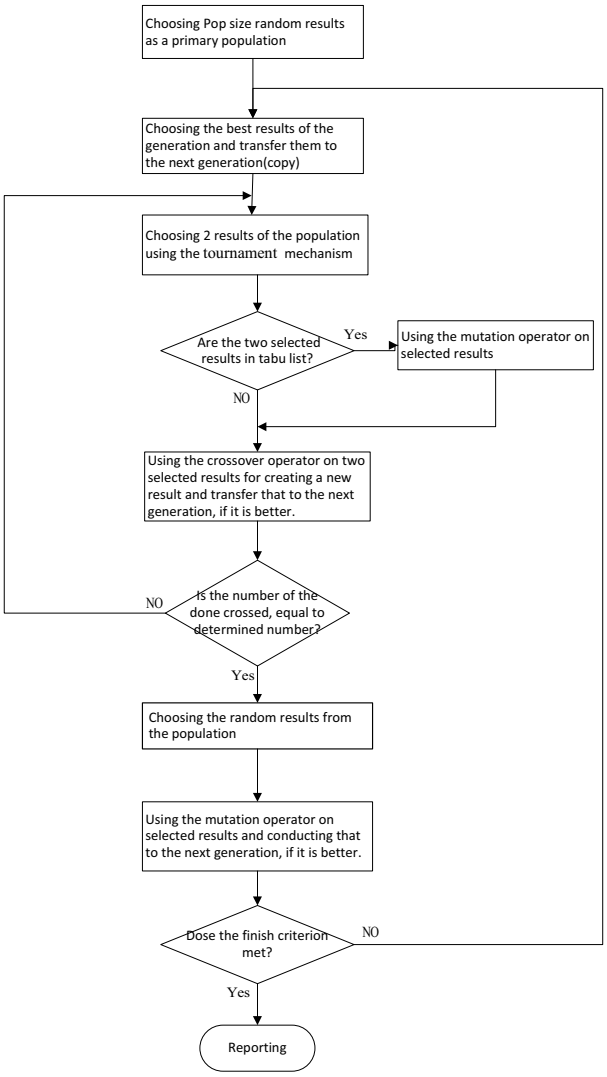


Fig. 1 Structure of the algorithmic process

E. Sorting

After being evaluated, chromosomes are sorted based on their fitness value, from highest fitness to lowest.

F. Elitism

Copy_size chromosomes with highest fitness are directly selected and are copied to the next generation.

G. Selection

Selection operator determines parents for construction of next generation. It's obvious that those chromosomes that having high fitness, have high chance for selection. In the proposed algorithm, selection is based on the tournament selection. The first step is a uniform random number less than one. The tournament selection chooses each parent by choosing n players at random and then choosing the best

individual out of that set to be a parent. The tournament size (or subgroup, n) must be at least 2. This size influences the selective pressure, i.e. more individuals in the subgroups increase the selection pressure on the better individuals.

H. Crossover or Recombination

In the research carried out on the traveling salesman problem (TSP), several crossover operators have been proposed for a permutation representation such as partial-mapped crossover (PMX), order crossover (OX), position based crossover (PBX), and order-based crossover (OBX). These operators can be viewed as an extension of two-point or multi-point crossovers of binary strings to the permutation representation. Generally, two-point crossovers yielded infeasible offspring in sense of two or more nodes may be duplicated. The repairing procedure is usually embedded in these operators in order to fix this problem. In this paper, we benefit from the well-known TSP crossovers with modifications to adapt them to our problem.

At each generation (*Cross_rate* × *pop_size*)/2 pairs parents are selected according to the selection operator.

For each parent two cross-points are selected randomly and the section between them is exchanged between the two parents. After this stage, if a chromosome contain duplicated number, as described, for the reason that we use a deterministic simulator and duplicate experiment in one point have same result, we use mutation operator in order to fix this problem. For do this one of the genes that contains duplicated number is selected and mutated. Fig 2 and Fig 3 depicts this.

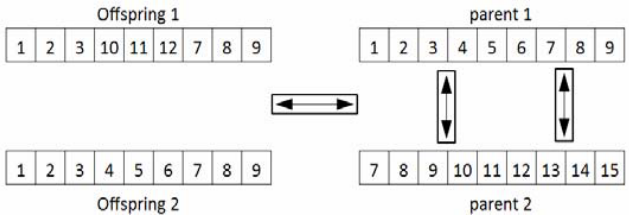


Fig. 2 Crossover when duplicated number not produced.

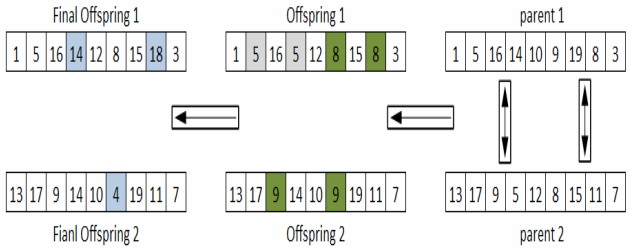


Fig. 3 Crossover when duplicated number produced.

İ. Mutation

After crossover, (*Mutation_rate* × *pop_size*) parents are mutated whit uniform mutation operator. One of the most popular operators for mutation is uniform mutation (Mühlenbein et al [020]). This mutating function operate in two stages: at first the algorithm determine position of mutation and then replace value of this position whit another value at random manner.

Keep in mind that the mutated chromosome should not have duplicate number.

J. Reporting

Finally, if the termination condition is satisfied, which is to repeat the algorithm for *max_generation* number of times, chromosome with highest fitness determined, and reported as the solutions of the problem.

V. Experiments and Computational Results

This section presents experimental results. In order to test the performance of the proposed genetic algorithm (GA), we have solved some problems in small design space with the exact method and then solved the same questions with the suggested algorithm. Then we consider some problem whit large design space that we can't solve them in exact manner.

A. Genetic algorithm implementation

The proposed genetic algorithm is programmed in the Matlab 7.0.4 and executed on a Pentium 4 2.8 MHz pc. We test the performance of our proposed algorithm with producing various values for N and n. In all test problems, the matrix that we used as covariance between points is a sub matrix of A (the described matrix in previous section that contains covariance between all of the points of design space).

B. Computational results

The proposed algorithm is implemented by the Matlab. The associated results are obtained by running the codes on test problems. Table 1 reports the results obtained by the proposed MA and the exact method (complete enumeration) on the set of instances. For each problem, we give the problem number, design space (N), number of desired pints for experimental design (n). For each test problem, Table 1 presents the following output:

1. The solution value that obtained from exact manner and GA method.
2. The computation time that spends by GA method and expressed in seconds.

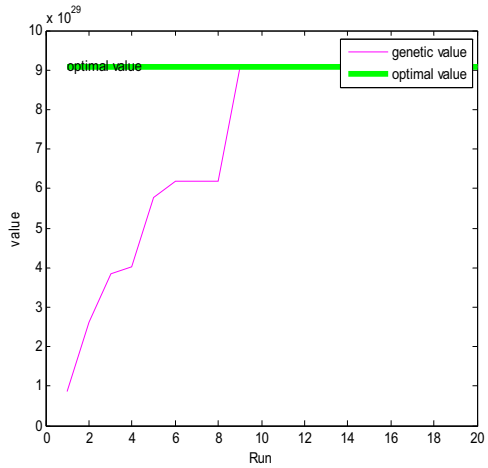
Figures 4, 5 are GA progress and exact solution in various Examples.

In all of our experiments, the time spent by the algorithm is negligible and our algorithm is quite effective on the data sets that we applied it to. As it is also obvious from the convergence graph, the algorithm convergence is very fast and it results to the optimum solution in the primary steps. It is possible to show the accuracy of the suggested algorithm by comparison between the suggested algorithm and the optimum solution (which is possible to be calculated for the questions with small size). So this algorithm could be used for solving the important questions in which it is impossible to calculate the optimum solution, and expect that the suggested algorithm would be able give very good results.

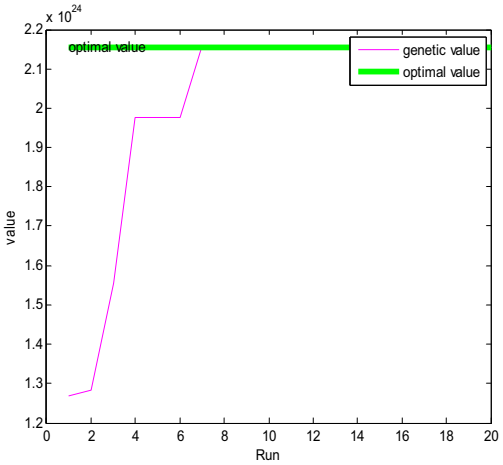
Thus we can use this algorithm for solving our problem. Whit using the proposed GA algorithm, we construct design of experiment for tank problem and reach to pest possible design.

TABLEI GA RESULTS FOR EIGHT PROBLEMS

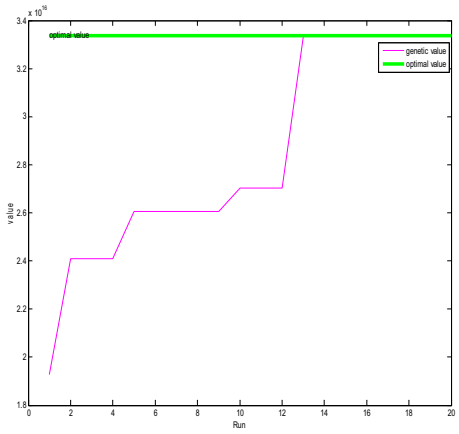
P	N	n	Exact solution	GA	Time(seconds)
1	16	8	2.1544e+024	2.1544e+024	2.4042
2	21	10	9.0661e+029	9.0661e+029	4.9539
3	26	6	2.2671e+019	2.2671e+019	2.2988
4	36	5	3.3362e+016	3.3362e+016	8.1747
5	100	25	--	4.8188e+074	45.1676
6	200	25	--	1.9060e+077	63.2114
7	500	25	--	1.8613e+079	107.6746
8	1000	25	--	8.4404e+080	138.7



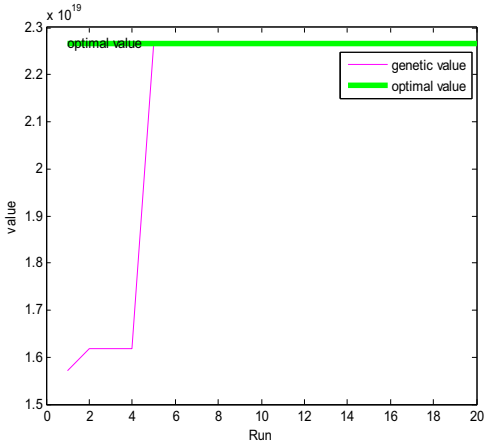
(a) N=21 n=10



(b) N=16 n=8

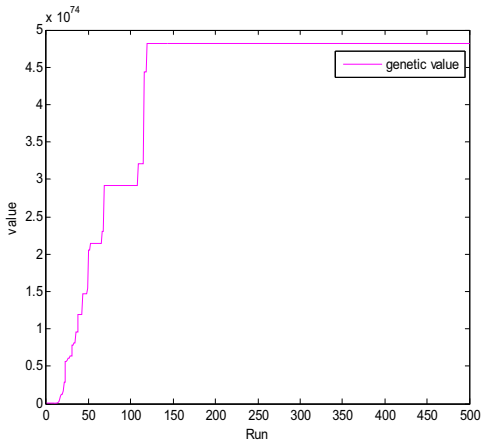


(d) N=35 n=5

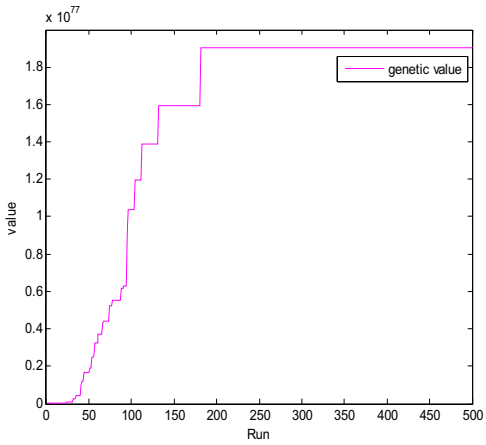


(c) N=26 n=6

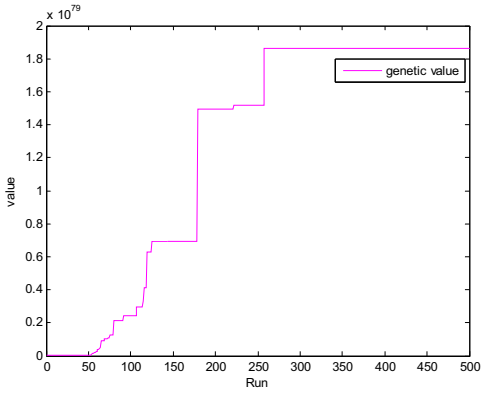
Fig. 4 GA progress and exact solution in various Examples



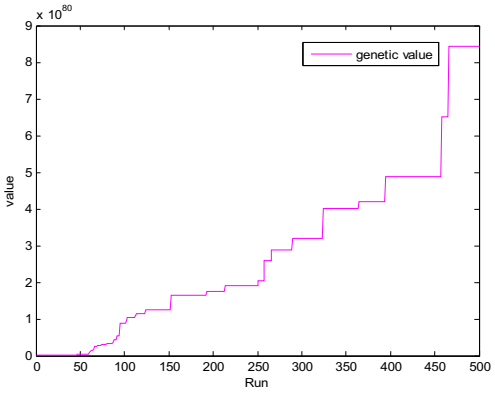
(e) N=100 n=25



(f) N=200 n=25



(g) N=500 n=25



(h) N=1000 n=25

Fig. 5 GATS progress in various Examples

VI. Conclusion

Heuristic optimization methods are useful for solving experiment design problems in cases where other methods cannot be applied, or are ineffective. The GA algorithm was demonstrated to be effective in this type of design of experiments. Also we demonstrate usefulness of spatial design in a real case.

With daily spread of the optimum design of experiment usage in various areas and the need for efficient methods in construction of optimum design, the necessity of paying attention to the optimum design of experiments is justifiable. [54].

One of the criterions which had been attended by scientists is the entropy criterion. And there is a vast research area for it. Among them we can point following:

- Using entropy for creating optimum design of operational examples in different field of science.
- Because for middle volume problems it is possible to reach the optimum solution by branch and bound method, another research area is research on finding efficient upper bands for branch and bound questions.
- Using entropy criterion for creating optimal designs of experiment in multi response experiment is another open area for research.

Hitherto there is not any conducted research in this area.

References

[1] Muller, W.G. (2007). Collecting Spatial Data: Optimum Design of Experiments for Random Fields. 3nd ed., Physica Verlag, Heidelberg .
[2] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.
[3] Cox, D.D., Cox, L.H. and Ensore, K.B. (1997). Spatial sampling and the environment: some issues and directions. *Environmental and Ecological Statistics*, 4:219–233.
[4] Cochran, W.G., 1977. Sampling Techniques, 3rd Edition. Wiley, New York.
[5] Stuart, A., 1984. The Ideas of Sampling. Griffin, London.

[6] Wollum, A.G., 1994. Soil sampling for microbiological analysis. In: Weaver, R.W., et al. (Eds), Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties. Soil Science Society of America, Madison, pp. 1–14.
[7] Wendroth, O., Reynolds, W.D., Vieira, S.R., Reichardt, K., Wirth, S., 1997. Statistical approaches to the analysis of soil quality data. In: Gregorich, E.G., Carter, M.R. (Eds.), Soil Quality for Crop Production and Ecosystem Health. Elsevier, Amsterdam, pp. 247–276.
[8] Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons, Agriculture, Ecosystems and Environment 94, PP. 31–47.
[9] Melissa J. Dobbie, Brent L. Henderson, and Don L. Stevens, Jr, 2008. Sparse sampling: Spatial design for monitoring stream networks, Statistics Surveys Vol. 2, PP. 113–153.
[10] Muller, W.G. (2000). Collecting Spatial Data: Optimum Design of Experiments for Random Fields. 2nd ed., Physica Verlag, Heidelberg.
[11] Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An evaluation of model dependent and probability sampling inferences in sample surveys. Journal of the American Statistical Association 78, 776–760.
[12] Sarndal, C. (1978). Design-based and model-based inference for survey sampling. Scandinavian Journal of Statistics 5, 27–52.
[13] Brus, D.J. and de Gruijter, J.J. (1993). Design-based versus modelbased estimates of spatialmeans: Theory and application in environmental soil science. Environmetrics 4, 123–152.
[14] Brus, D.J. and de Gruijter, J.J. (1997). Random sampling or geostatistical modeling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). Geoderma 60, 1–44.
[15] de Gruijter, J.J., and Ter Braak, C.J.F. (1990). Model free estimation from survey samples: A reappraisal of classical sampling theory. Mathematical Geology 22, 407–415.
[16] Cressie, N., Calder, C.A., Clark, J.S., Ver Hoef, J.M. and Wikle, C.K. (2007). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Department of Statistics Preprint No. 798, The Ohio State University.
[17] Royle, J.A. and Nychka, D. (1998). An Algorithm for the Construction of Spatial Coverage Designs with Implementation in Splus. Computers and Geosciences 24, 479–488.
[18] Nychka, D. and Saltzman, N. (1998). Design of air-qualitymonitoring networks. In Case Studies in Environmental Statistics, D. Nychka, W. Piegorsch, and L. Cox, Eds. Springer, New York, 51–76.
[19] De Gruijter, J.J. and Ter Braak, C.J.F. (1990) Model-free estimation from spatial samples: A reappraisal of classical sampling theory. Mathematical Geology, 22(4), 407-15.
[20] Cressie, N.A.C. (1991) Statistics for Spatial Data. Wiley, New York.
[21] Christakos, G. (1992) Random Field Models in Earth Sciences. Academic Press, San Diego.
[22] Caselton, W.F. and Hussian, T. (1980) Hydrologic networks: Information transmission. Journal of the Water Resources Planning and Management Division, A.S.C.E., 106 (WR2), 503-20.
[23] Caselton, W.F. and Zidek, J.V. (1984) Optimal monitoring network designs. Statistics and Probability Letters, 2, 223-7.
[24] Caselton, W.F., Kan, L., and Zidek, J.V. (1991) Quality data network designs based on entropy. In Statistics in the Environmental and Earth Sciences, P. Guttorp and A. Walden (eds), Griffin, London.
[25] Ko, C.-W., Lee, J., and Queyranne, M. (1995) An exact algorithm for maximum entropy sampling. Operations Research, 43, 684-91.
[26] Sacks, J. and Ylvisakerd. (1984). Some model robust designs in regression. *Ann. Statist*. 12, 1324-1348.
[27] Sacks, J. and Ylvisakerd. (1985). Model robust design in regression: Bayes theory. In Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer (L. M. Le Cam and R. A. Olshen, eds.) 2, 667-679. Wadsworth, Monterey, Calif.
[28] Welch, W. J. (1983). A mean squared error criterion for the design of experiments. *Biometrika* 70, 205-213.
[29] Mitchell, T. J. An algorithm for the construction of "D-optimal" experimental designs. *Technometrics* **16**, 1974, 203-210.
[30] Dodge, Y., Fedorov, V, V. and Wynn, H. P. Optimal Design and Analysis of Experiments. *Elsevier*, 1988.
[31] Shewry, M.C. and Wynn, H.P. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, 14, 165-170.
[32] Guttorpn, P., Le, N. D, Sampson, P. D. and Zidek, J. V. (1992). Using Entropy in the Redesign of an Environmental Monitoring Network. Technical Report #116, *Department of Statistics, The University of British Columbia.*
[33] Mckaym, D., Conoverw, J. and Beckmanr, J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21, 239-245.
[34] Imanr, L. and Heltonj, C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis* 8, 71-90.
[35] Kurt M. Anstreicher, Marcia Fampa, Jon Lee, and JoyWilliams. (1996). Continuous relaxations for constrained maximum-entropy sampling. In *Integer Programming and Combinatorial Optimization (Vancouver, BC, 1996)*, volume 1084 of *Lecture Notes in Computer Science*, pages 234–248. Springer, Berlin.
[36] Kurt M. Anstreicher, Marcia Fampa, Jon Lee, and Joy Williams. (1999). Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling problems. *Mathematical Programming, Series A*, 85(2):221–240.
[37] Jon Lee. (1998) Constrained maximum-entropy sampling. *Operations Research*, 46(5):655–664.
[38] Jon Lee. (2000). Semidefinite-programming in experimental design. In Henry Wolkowicz, Romesh Saigal and LievenVandenberghe, editors, *Handbook of Semidefinite Programming*, volume 27 of *International Series in Operations Research and Management Science*, pages 528–532. Kluwer.
[39] Jon Lee. Maximum-entropy sampling. (2001). In Abdel H. El-Shaarawi andWalterW. Piegorsch, editors, *Encyclopedia of Environmetrics*, volume 3, pages 1229–1234. JohnWiley & Sons Inc.
[40] Alan Hoffman, Jon Lee, and Joy Williams. (2001). New upper bounds for maximum-entropy sampling. In A.C. Atkinson, P. Hackl, andW.G.M¨uller, editors,*MODA 6—Advances in model-oriented design and analysis*, pages 143–153. Springer-Verlag.
[41] Jon Lee, Joy Williams. (2003). A linear integer programming bound for maximum-entropy sampling, Math. Program., Ser. B 94: 247–256.
[42] Michalewicz, Zbigniew. (1992). Genetic Algorithms + Data Structures = *Evolution Programs*, Springer-Verlag, New York.
[43] Haupt, Randy L. and Haupt, Sue Ellen. (1998). *Practical Genetic Algorithms*, Wiley, New York.
[44] Davis, Lawrence (Ed.). (1991). *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.
[45] Goldberg, David E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York.