# The effects of parameter settings on the performance of Genetic Algorithm through experimental design and statistical analysis

Farhad Kolahan, Assistant Prof.
Dept. of Mechanical Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
kolahan@um.ac.ir

Marziyeh Hassani Doughabadi
Dept. of Industrial engineering
Sadjad Institute of Higher Education
Mashhad, Iran
hasanymarzieh@gmail.com

*Abstract*—**Genetic algorithm (GA) is a meta-heuristic inspired by the efficiency of natural selection in biological evolution. It is one of the most widely used optimization procedure which has successfully been applied on a variety of complex combinatorial problems. The main drawback of GA, however, is its several tuning variables which need to be correctly set. The performance of GA largely depends on the proper selection of its parameters values; including crossover mechanism, probability of crossover, population size and mutation rate and selection percent. The objective of this research is to evaluate the effects of tuning parameters on the performance of genetic algorithm using the data collected as per Central Composite Design (CCD) matrix. To gather the required data, the proposed approach is implemented on a well-known travelling salesman problem with 48 cities. Then, regression modeling has been employed to relate GA variables settings to its performance characteristic. Analysis of Variance (ANOVA) results indicate that the function can properly represent the relationship between GA important variables and its performance measure (solution quality).**

***Keywords: Genetic Algorithm; ANOVA; Regression modeling; Design of Experiments***

## I. INTRODUCTION

In the past few decades optimization problems are becoming even more complex involving many decision variables and high non-linearity. Exact algorithms are usually ineffective and too slow to solve such large and complex problems. In addition, many optimization problems belong to the class of combinatorial optimization problems, for which no exact algorithm exists.

In recent years, with advent of computer capabilities, many heuristic algorithms have been proposed to deal with such large-sized problems. Among these, Genetic algorithm (GA) is one of the oldest and most widely used optimization procedures [1, 2]. Due to its several advantages, GA has become one of the most favorite evolutionary techniques in combinatorial optimization. Nevertheless, one of the challenging aspects of this algorithm is its numerous tuning parameters which have significant impacts on GA's solution quality and computational time. Each GA parameter may be considered in several levels and hence there are almost infinite numbers of possibilities. This combinatorial explosion on GA factors and its values makes it extremely difficult to evaluate the effects of parameters settings on GA performance. Therefore, there is a need for more profound and effective way to determine the influence of each parameter so as its proper values may be determined.

Work on GA parameters is a well established research area (e.g. Ghrayeb and Phojanamongkolkij [3] Kaya [4], Albayrak and Allahverdi [5]). The details of the other works on GA operators and parameters are well documented in the related literatures [6-9]. In general, in most existing research there is a lack of joint consideration of all important GA parameters simultaneously. The main objective of this work is, therefore, to investigate the mutual influences of GA's prominent parameters through statistical analysis and mathematical modeling. The proposed procedure is applied on a well-known benchmark TSP for 48 (att48) cities [10]. It is noted, this approach may be used for any other problem with minor modifications.

## II. GENETIC ALGORITHM

Genetic Algorithms are stochastic search techniques for approximating optimal solutions within complex search spaces. The technique is based upon an analogy with biological evolution, in which the fitness of individual determines its ability to survive and reproduce. Each solution in GA is represented in the form of a string of numbers or symbols, resembling chromosomes and their associated genes. The genes are then randomly combined to produce a population of chromosomes. Genetic operations are performed on chromosomes that are randomly selected from the population. This produces offspring. The fitness of these chromosomes is then measured and the probability of their survival is determined by their fitness. GA's major parameters and operations include population size (P), number of generations (G), crossover operator (COP) probability of crossover (%C), and mutation rate (%M). New chromosomes are created by crossover which is the probabilistic exchange of values between two selected chromosomes; or mutation, generating a new random chromosome by such means as random replacement of values in a chromosome. Mutation provides randomness within the chromosomes to increase coverage of the search space and help prevent premature convergence on a local optimum. Chromosomes are then evaluated according to a fitness function, with the fittest surviving and the less fit being eliminated. To avoid losing good solutions, the most fitted ones, called elites, are copied directly to the next

978-1-4244-9265-7 /11/$26.00

generation. The result is a new population that evolves over time to produce better and fitter solutions to the problem on hand. GA is stochastic iterative processes and is not guaranteed to converge on an optimal solution. Thus, search process typically terminates when a pre-specified fitness value is reached, a set amount of computing time passes or until no significant improvement occurs in the population for a given number of iterations [11]. The details of this algorithm and its diverse applications can be found in related literatures [e.g. 1, 2 and 12].

## III. PROBLEM STATEMENT AND COMPUTATIONAL RESULTS

There are several types of benchmark problems for assessing the performance of a given optimization technique in terms of computational speed and solution quality. Travelling salesman problem (TSP) is one of the most famous combinatorial optimization problems. In its classical form, TSP consists of a set of N nodes or cities for which a closed tour with minimum distance should be constructed. For a problem with N cities, there are N! possible solutions and hence TSP like problems are classified as non-polynomial (NP)-complete problems. It means that the required computational effort increases exponentially with the number of cities. This property makes exact algorithms based on enumeration, extremely time consuming and inefficient.

In this work, the problem with 48 cities is used as a benchmark to model and evaluate significant parameters in GA. The structure of this problem and its optimal tour are given by Germany Heidelberg University database [10]. The objective is to investigate the effects of GA parameters settings and the types of operators on its solution quality. The parameters under study include population size (Pop), probabilities of crossover (Cr), mutation (Pm) and selection (Sr), as well as crossover operators. In our computational experiments, all GA parameters are studied in five levels, while three types of crossovers (PMX[1], OX[2], and Heuristic) are investigated.

The proposed approach is based on regression modeling on the data gathered through CCD Design of Experiment technique. To obtain required data for modeling, Design of Experiments (DOE) approach has been employed. Experimental design consists of a group of techniques used in the empirical study of relationship between one or more measured responses and a number of input variables [13]. In our study, Central Composite Design (CCD) is employed to reduce the number of computational executions needed to investigate the influences of GA parameters. In its basic form, CCD is a design requiring 5 levels of each parameter (0, ±1, ±a). The selected designed matrix is a standard central composite rotatable four-factor five-level factorial design with 31 experiments. To facilitate design matrix construction, a coding system is employed to indicate different ranges of parameters [14]. The values of GA parameters, given by this coding scheme, are shown in Table I. In this table, Pop, Cr, Pm, Sr are the number of population, probability of crossover, mutation probability and selection rate, respectively.

TABLE I. SETTING OF PARAMETERS LEVELS FOR GA IN CCD MATRIX

| Parameters | -2 | -1 | 0 | +1 | +2 |
|---|---|---|---|---|---|
| **Pop** | 50 | 150 | 250 | 350 | 450 |
| **Cr** | 0.3 | 0.45 | 0.6 | 0.75 | 0.9 |
| **Pm** | 0.001 | 0.026 | 0.050 | 0.075 | 0.100 |
| **Sr** | 0.35 | 0.5 | 0.65 | 0.8 | 0.95 |

The computer code was prepared using Matlab 7.1 software. For comparison purposes, in all runs computational experiments were performed for the same amount of CPU times. The algorithm was run five times for each combination of parameters and the mean of results was used as the final solution. Since, computational experiments have been performed for three types of crossover, there are a total of 93 (31×3) solutions.

### *A. Selecting the crossover mechanism*

To select the best crossover mechanism, all tests were performed for each of the three operators. The results were then compared in terms of solution quality. These pairwsie comparisons are schematically illustrated in Fig. 1 and 2. As shown, in all 31 runs OX performs better than PMX and Heuristic in terms of solution quality. Therefore, this crossover is selected in our future analysis and the mathematical model is developed based on computational results given by OX as the crossover operator.

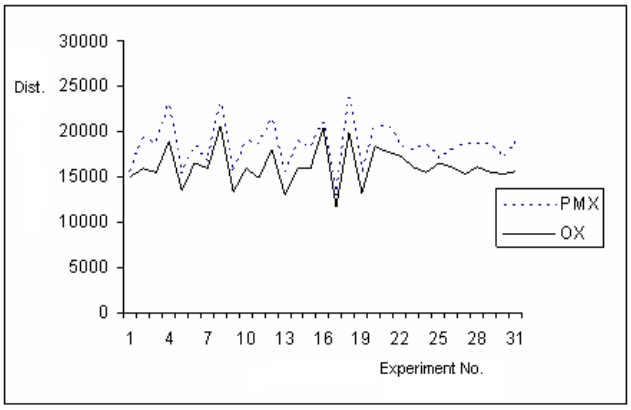


Figure 1. The comparison between PMX and OX crossovers.

[1] Partially Mapped Crossover

[2] Ordered Crossover

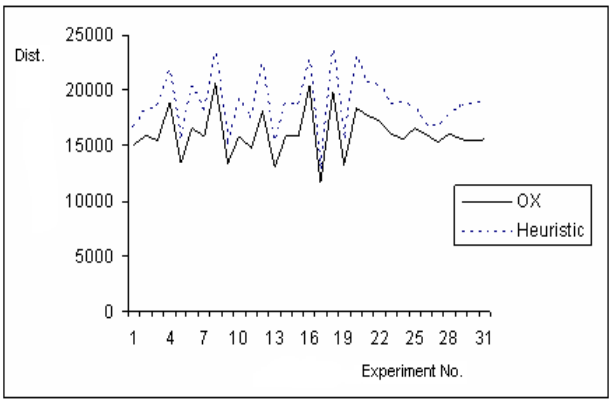


Figure 2. The comparison between OX and Heuristic crossovers

### B. *Regression modeling and Analysis of Variance (ANOVA)*

A regression model is an approximate fit to a set of sample data in a way that the sum of the square errors is minimized [13]. In this study, curvilinear function has been fitted on the experimental data to establish the relationships between GA parameters and its performance characteristic. The general form of the function is as follows:

$$Dist=b_0 +b_1Pop +b_2Cr +b_3Pm +b_4Sr +b_{11}Pop^2 +b_{22}Cr^2 +b_{33}Pm^2 +b_{44}Sr^2 +b_{12}Pop\times Cr +b_{13}Pop\times Pm +b_{14}Pop\times Sr +b_{23}Cr\times Pm +b_{24}Cr\times Sr +b_{34}Pm\times Sr \quad (1)$$

In the above equation, Dist is the length of the tour for a given TSP. The GA parameters values are stated by Pop, Cr, Pm and Sr. Finally, $b_0$ is the intercept term; while $b_1$, $b_2$… $b_{34}$, $b_{44}$ are coefficients of variables. Based on experimental data for the att48 TSP example, the mathematical model representing the relationship between GA parameters and its performance measure can be stated by:

$$Dist= 20786 - 7.9928\ Pop -5852.5\ Cr -181811\ Pm -6278\ Sr -0.0001\ Pop^2 +263.16\ Cr^2 +666074\ Pm^2 +257.61\ Sr^2 +26.829\ Pop\times Cr +131.57\ Pop\times Pm +4.2458\ Pop\times Sr +111183\ Cr\times Pm +2863.9\ Cr\times Sr +30583\ Pm\times Sr \quad (2)$$

This model can predict GA solution (final length of the tour) for any given set of parameter settings. They may also give insight into the relative importance of each GA parameters.

To assess the quality of the proposed model and to determine their adequacies, Analysis of Variance (ANOVA) has been performed within the confidence limit of 95%. Table II shows the Pr and F values resulted from ANOVA. Generally, the higher value of the correlation coefficient $R^2$ the higher significance of the model. The correlation coefficient of curvilinear model is 97%. This means it can predict GA performance with very good accuracy.

TABLE II. ANALYSIS OF VARIANCE (ANOVA) FOR THE QUADRATIC MODEL

| Source | DF | Sum of Squares | Mean Squares | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 14 | 128903136 | 9207367 | 37.17 | <.0001 |
| **Error** | 16 | 3963586 | 247724 | | |
| **Corrected Total** | 30 | 132866722 | | | |

The significance of each parameter in curvilinear model is determined using t-test and P-values which are listed in Table III. Student's t-test is employed to determine the mean square error which can be obtained by dividing each coefficient by its standard error.

A large t-value implies that the coefficient is much greater than its standard error. The P-values are necessary to understand the pattern of the mutual interactions between the test variables. For any parameter, larger t-value and smaller P-value indicate that the factor is very significant.

As indicated in Table III, the most important parameter affecting GA's solution quality is the probability of mutation (Pm). The t-test analysis also reveals that both first order and second order of Pm are highly significant since their respective P-values are very small. Moreover, the interactions between the population size and crossover probability (Pop-Cr), population size and mutation probability (Pop-Pm), probabilities of crossover and mutation (Cr-Pm) are also significant. All these interactions have positive effects on GA's response characteristic. The approach presented here may further be used to optimally determine the parameter settings so as the efficiency of GA is maximized.

TABLE III. SIGNIFICANCE OF REGRESSION COEFFICIENTS GIVEN BY STATISTICAL ANALYSIS

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Intercept** | 1 | 20786 | 4170.45 | 4.98 | 0.0001 |
| **Pop** | 1 | -7.99 | 9.11 | -0.88 | 0.3932 |
| **Cr** | 1 | -5852.46 | 6722.05 | -0.87 | 0.3968 |
| **Pm** | 1 | -181811 | 35547 | -5.11 | 0.0001 |
| **Sr** | 1 | -6278.04 | 6895.83 | -0.91 | 0.3761 |
| **$Pop^2$** | 1 | -0.001 | 0.01 | -0.01 | 0.9898 |
| **$Cr^2$** | 1 | 263.16 | 4136.67 | 0.06 | 0.9501 |
| **$Pm^2$** | 1 | 666074 | 148920.00 | 4.47 | 0.0004 |
| **$Sr^2$** | 1 | 257.61 | 4136.67 | 0.06 | 0.9511 |
| **Pop_Cr** | 1 | 26.83 | 8.29 | 3.23 | 0.0052 |
| **Pop_Pm** | 1 | 131.57 | 49.77 | 2.64 | 0.0177 |
| **Pop_Sr** | 1 | 4.24 | 8.29 | 0.51 | 0.6158 |
| **Cr_Pm** | 1 | 111183 | 33181.00 | 3.35 | 0.0041 |
| **Cr_Sr** | 1 | 2863.89 | 5530.21 | 0.52 | 0.6116 |
| **Pm_Sr** | 1 | 30583.00 | 33181.00 | 0.92 | 0.3704 |

## IV. CONCLUSION

In this research, based on Design of Experiments (DOE) approach and regression modeling, the effects of tuning parameters on the performance of genetic algorithm, using a TSP benchmark problem (att 48) have been evaluated. Central composite design matrix with 31 experiments was used to gather the required data for regression modeling. Based on computational results, the effects of three types of crossover have also been studied. Results show that in all cases OX crossover is better than PMX and heuristic mechanisms. Analysis of Variance (ANOVA) for curvilinear function reveals that mutation probability as well as interaction effects between population and crossover, population and mutation and between mutation and crossover have significant influences on the performance of GA in terms of solution quality. In future research works, the proposed approach can readily be used to determine the best set of parameter settings for any optimization problem.

## REFERENCES

[1] C. Moon, J. Kim, G. Choi, Y. Seo, "An efficient genetic algorithm for the traveling salesman problem with precedence constraints," European Journal of Operational Research 140, 2002, pp. 606–617.

[2] F. Liu, G. Zeng, "Study of genetic algorithm with reinforcement learning to solve the TSP," Expert Systems with Applications 36, 2009, pp. 6995–7001.

[3] O. Ghrayeb and N. Phojanamongkolkij, "A study of optimizing the performance of genetic algorithms using design-of-experiments in job-shop scheduling application," International Journal of Industrial Engineering-theory Applications and Practice 12, 2005, pp. 37-44.

[4] M. Kaya, "The effects of two new crossover operators on genetic algorithm performance," Applied Soft Computing, "in press", 2010.

[5] M. Albayrak and N. Allahverdi, "Development a new mutation operator to solve the Traveling Salesman Problem by aid of Genetic Algorithms," Expert Systems with Applications, "in press", 2010.

[6] P. Pongcharoen, W. Chainate and P. Thapatsuwan, "Exploration of Genetic Parameters and Operators through Travelling Salesman Problem," Science Asia 33, 2007, pp. 215-222.

[7] E. Alfaro-Cid, E.W. McGookin, D.J. Murray-Smith, "A comparative study of genetic operators for controller parameter optimization," Control Engineering Practice, vol. 17, 2009, pp. 185– 197.

[8] G. Laporte, "The traveling salesman problem: An overview of exact and approximate algorithms," European Journal of Operational Research, vol. 59 (2) ,1992, pp. 231-247.

[9] S. Chatterjee, C. Carrera, L. A. Lynch, "Genetic algorithms and traveling salesman problems," European Journal of Operational Research, vol. 93 (3), 1996, pp. 490-510.

[10] www.iwr.uniheidelberg.de/iwr/comopt/soft/ TSPLIB95/TSPLIB.html

[11] C. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Wesley: Addison, 1989.

[12] L. H. Randy, Haupt Sue Ellen, Practical Genetic Algorithm, JOHN WILEY & SONS, 2004.

[13] D.C. Montgomery, Design and analysis of experiments – fifth edition, JOHN WILEY & SONS,2001

[14] F. Kolahan, M. M. Bagheri, "Modeling and prediction of process parameters in Submerged Arc Welding by Genetic Algorithm," Proc. of the 37th International Conference on Computers and Industrial Engineerin, October 20-23, 2007, Alexandria, Egypt.