Evaluating the Effects of Parameters Setting on the Performance of Genetic Algorithm Using Regression Modeling and Statistical Analysis

Marziyeh Hasani Doughabadi [1] , Hossein Bahrami [2] and Farhad Kolahan *[2]
[1] Faculty of Industrial engineering, Sadjad Institute of Higher Education, Mashhad, Iran
[2] Department of Mechanical Engineering, Ferdowsi University of Mashhad, Iran
(Received 12 December 2010, Accepted 24 July 2011)

Abstract

Among various heuristics techniques, Genetic algorithm (GA) is one of the most widely used techniques which has successfully been applied on a variety of complex combinatorial problems. 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. In this paper, based on Design of Experiments (DOE) approach and regression modeling, the effects of tuning parameters on the performance of genetic algorithm have been evaluated. As an example, GA is applied to find a shortest distance for a well-known travelling salesman problem with 48 cities. The proposed approach can readily be implemented to any other optimization problem. To develop mathematical models, computational experiments have been carried out using a 4-factor 5-level Central Composite Design (CCD) matrix. Three types of regression functions models have been fitted to relate GA variables to its performance characteristic. Then, statistical analyses are performed to determine the best and most fitted model. Analysis of Variance (ANOVA) results indicate that the second order function is the best model that can properly represent the relationship between GA important variables and its performance measure (solution quality).

Keywords: ANOVA, Design of experiments, Genetic algorithm, Optimization, Regression modeling

Introduction

With the advent of computer technology and growing complexity of engineering problems in the past few decades, there has been much research to develop and use heuristic algorithms that can solve large scale optimization problems efficiently and effectively. Since late 1980s, a large number of optimization algorithms based on principles of natural and physical phenomena have been proposed. Genetic algorithm (GA) [1, 2], Simulated Annealing [3], Ant Colony Optimization [4], and Tabu Search [5] are some of the well-known heuristics used in combinatorial optimization. Among these, Genetic algorithm (GA) is one of the oldest and most widely used optimization procedures. Now days, there are several versions of Genetic Algorithms (GAs) that have successfully been applied to a variety of optimization problems [1, 2]. Due to its several advantages, GA has become one of the most favorite evolutionary techniques in combinatorial optimization. GA performs multiple directional searches using a set of candidate solutions; while most conventional methods conduct single directional search. It deals directly with the solutions to the problem instead of problem itself. It requires no domain knowledge and uses stochastic transition rules to guide the search. Nevertheless, one of the challenging aspects of this algorithm is its numerous tuning variables. GA's major parameters include population size (P), number of generations (G), crossover operator (COP) probability of crossover (%C), and mutation rate (%M). In terms of time and solution quality, the performance of the search, to a large extend, depends on its parameters settings. Moreover, when the problem size grows large, this technique faces difficulties to find

* Corresponding author: Tel: +98- 511- 8615100 Fax: +98- 511- 8436432 Email: kolahan@um.ac.ir

the global or near global solutions. Traditionally, time-intensive trial and error runs were used to determine the best parameter settings according to the nature of problem domain. However, these methods were limited in the sense that they were case dependent and would not give much insight into the effects of each parameter on the search performance. 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 set the proper levels through enumeration or any other trial and error approaches. Therefore, there is a need for more profound and effective way to find the influence of each parameter so as its proper values may be determined.

Work on GA parameters is a well established research area. Todd [6] investigated the performance of fourteen crossover and five mutation operators within GA applied to five problem sizes of TSP. However, some important parameters such as probabilities of crossover (%C) and mutation (%M) were overlooked. For basic flow-shop scheduling problem, Pongcharoen et al. [7] used a Design of Experiments (DOE) approach to find appropriate setting of GA parameters. They have taken into account such GA parameters as the combination of population size and number of generations, probabilities of crossover and mutation as well as different crossover and mutation operators. Ghrayeb and Phojanamongkolkij [8] have employed DOE along with Analysis of Variance (ANOVA) approaches to investigate the effects of GA parameters on its performance in solving job shop scheduling problem. Although, some important parameters including population size, number of generations, and probabilities of crossover and mutation have been considered on their work, they failed to account for two key GA operators; namely (COP and MOP). More recently, attempts have been made to improve GA performance by the means of more effective crossover (Kaya [9]) and mutation mechanisms (Albayrak and Allahverdi [10]). These works, however, did not consider the effects of other GA parameters on its performance. The details of the other works on GA operators and parameters are well documented in the related literatures [11-14].

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 [15]. It is noted, this approach may be used for any other problem with minor modifications.

Genetic Algorithm

Genetic algorithm (GA) is a meta-heuristic inspired by the efficiency of natural selection in biological evolution. In GA the concepts of natural evolution are used to direct the search toward areas of high expected performance. This evolution is based on the past information which is summarized using a coding scheme.

Each solution in GA is represented in the form of a string of numbers or symbols, resembling chromosomes and their associated genes. The algorithm works by generating a population of numeric vectors (called chromosomes), each representing a possible solution to the problem. The individual components within a chromosome are called genes. 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 vector. 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 (or objective) 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 generation. The result is a new population that evolves over time to produce better and fitter solutions to a problem. 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 [16]. In its general form, GA works through the following steps:

1. Start: Generate a set of feasible random population of chromosomes
2. Fitness: Evaluate the fitness of each chromosome in the population. The fitness value assigned to each individual is determined by the fitness function defined by the problem being solved.
3. Check: If the termination criterion is reached, stop the search and show the best chromosomes of the current population as the final solution; otherwise proceed to the next step.
4. Next generation: Create a new population using crossover, mutation and elitism operators and go to step 2.

The details of this algorithm and its diverse applications can be found in related literatures [e.g. 1, 2 and 17].

Problem statement and computational results

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. In other words, the salesman is expected to visit each city exactly once and return to the point of his departure with minimum total travelling distance. Many optimization problems can be transferred to a TSP, including manufacturing scheduling, transportation, facility layouts, etc. 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 hence inefficient.

Computational research on the TSP began in earnest with the classic paper of Dantzig et al. [18], where the cutting-plane algorithm was used to calculate an optimal TSP tour through 49 cities in the United States. TSP has received considerable attention over the last two decades [19] and still is an attractive ongoing research topic. 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 [15]. The objective is to investigate the effects of GA parameters and operators on its solution quality while solving this kind of problems. This is done by regression modeling on the data gathered through Design of Experiment approach.

The parameters under study include population size, probabilities of crossover (%C), mutation (%M) and selection (%S), as well as crossover operators (COP). In our computational experiments, all GA parameters are studied in five levels, while three types of crossovers: partially mapped crossover (PMX), ordered crossover (OX) and heuristic crossover are used. To obtain required data, Design of Experiments (DOE) approach has been employed. DOE is a powerful technique used for exploring new processes, gaining knowledge of the existing processes and/or optimizing these processes for achieving desirable performance. 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. There are different DOE techniques including full factorial design, fractional factorial design, etc [20]. It is shown that Central Composite Design (CCD) matrix is the most popular design when there are large

numbers of parameters that provides equal precision of estimation in all directions. Therefore, CCD matrix is selected in this study for experimental runs.

Central Composite Design is a rotatable matrix that provides equal precision for fitted response at points (factor level combinations) that are of equal distance from the centre of the factor space. 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. The upper and lower limits are coded as +2 and −2, respectively. The intermediate values are calculated using Eq. (1).

$$X_i = \frac{2[2X - (X_{\max} + X_{\min})]}{(X_{\max} - X_{\min})} \qquad (1)$$

Where, X_i is the coded value of variable X ranging between X_{min} and X_{max}. For each parameter under study, +2 and -2 correspond to the upper limit (X_{max}) and lower limit (X_{min}) respectively. The values of GA parameters, given by this coding scheme, are shown in Table 1.

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

Table 1: GA parameters levels in CCD matrix

For four-factor five-level CCD matrix there are a total of 31 experiments out of which 16 are factorial points, 8 axial (star) points and 7 replicates at the centre points, as shown in Table 2. In this table, Pop, Cr, Pm, Sr are the number of population, probability of crossover, mutation probability and selection rate, respectively. Also the last three columns are associated with the three types of crossovers used in this research.

The computer code was prepared using Matlab 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, as shown in the last three columns of Table 2. In this table, each solution represents the length of a tour found by GA using the corresponding parameters setting. These 93 experimental runs are sufficient to gather required data for regression modeling relating the total distance of each tour to GA's tuning parameters.

A. Comparing types of crossover

The pairwise comparisons between different crossover operators are shown in Figures 1and 2. As illustrated, in all 31 runs OX is superior to PMX and Heuristic in terms of solution quality. Therefore, this crossover is selected in our future analysis and the mathematical models are developed based on computational results using OX as the crossover operator.

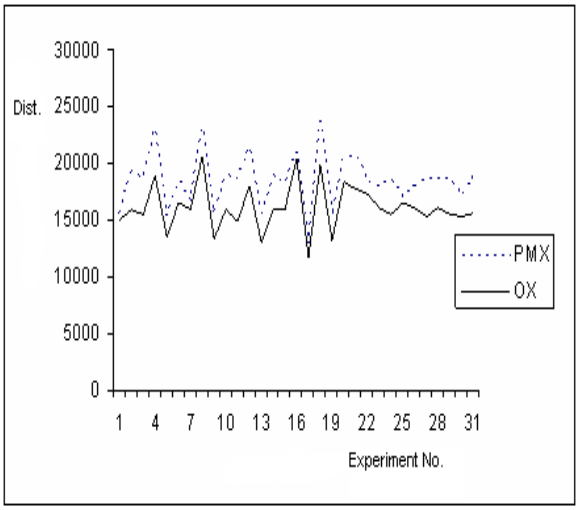

Figure 1: Comparison between PMX and OX crossovers

	Obs. No	Pop	Cr	Pm	Sr	Fitness fun. of PMX	Fitness fun. of OX	Fitness fun. of Heuristic
Factorial point	1	-1	-1	-1	-1	15558	14996	16764
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	16	+1	+1	+1	+1	21074	20397	22820
Axial point	17	-2	0	0	0	12938	11642	12794
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	24	0	0	0	+2	18566	15505	18915
Center point	25	0	0	0	0	16941	16558	18495
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	31	0	0	0	0	18753	15616	18998

Table 2: CCD matrix for process variables in coded and real units along with the observed responses

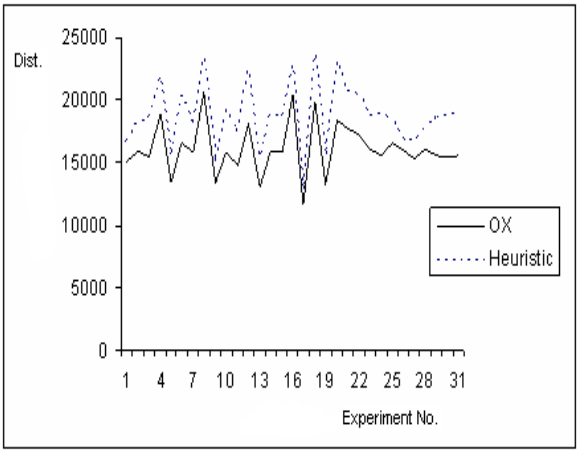

Figure 2: Comparison between OX and Heuristic crossovers

B. Model development and analysis

A model is a mathematical relationship that states changes in dependent variable when there are changes in independent variables. In many instances, it is of interest to model and explore the relationship between dependent and independent variables. The relationship between these variables is usually characterized by a regression model. The regression model is an approximate fit to a set of sample data in a way that the sum of the square errors is minimized [20]. In this study, Linear, Curvilinear and Logarithmic functions have been fitted on the experimental data to establish the relationships between GA parameters and its performance characteristic (solution quality). The general forms of the three functions are as follows:

Linear model: $Dist_1= b_0 +b_1Pop +b_2Cr +b_3Pm +b_4Sr$ (2)

Curvilinear model: $Dist_2= b_0 +b_1Pop +b_2Cr +b_3Pm +b_4Sr +b_{11}Pop^2 +b_{22}Cr^2 +b_{33}Pm^2 +b_{44}Sr^2 +b_{12}Pop\times Cr +b_{13}Pop\times Pm +b_{14}Pop\times Sr +b_{23}Cr\times Pm +b_{24}Cr\times Sr +b_{34}Pm\times Sr$ (3)

Logarithmic model: $Dist_3= e^{b0} \times Pop^{b1} \times Cr^{b2} \times Pm^{b3} \times Sr^{b4}$ (4)

In the above equations, 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 models representing the relationship between GA parameters and its performance measure (solution quality), can be stated by:

Linear model: $Dist_1= 7263 + 17.5\ Pop + 8702\ Cr + 5612\ Pm - 1604\ Sr$ (5)

Curvilinear model: $Dist_2 = 20786 - 7.9928\ Pop - 5852.5\ Cr - 181811\ Pm - 6278\ Sr - 0.0001\ Pop^2 + 263.16\ Cr^2 + 666074\ Pm^2 + 257.61\ Sr^2 + 26.829\ Pop \times Cr + 131.57\ Pop \times Pm + 4.2458\ Pop \times Sr + 111183\ Cr \times Pm + 2863.9\ Cr \times Sr + 30583\ Pm \times Sr$ (6)

Logarithmic model: $Dist_3 = e^{8.56} \times Pop^{0.225} \times Cr^{0.305} \times Pm^{-0.0090} \times Sr^{-0.0599}$ (7)

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

C. Statistical analysis and model selection

To assess the quality of the proposed models and to determine their adequacies, Analysis of variance (ANOVA) has been performed within the confidence limit of 95%. Given the Pr and F values resulted from ANOVA, all models are considered adequate within the specified confidence limit, as tabulated in Table 3.

Model	F Value	$P_r > F$	R^2	R^2-(adj)
Linear	45.88	0.000	87.6%	85.7%
Curvilinear	37.17	<.0001	97.0%	94.4%
Logarithmic	50.49	0.000	88.6%	86.8%

Table 3: ANOVA table for GA models

In regression modeling, the choice of the best model depends on the nature of initial data and the required accuracy. Generally, the higher value of the correlation coefficient R^2 the higher significance of the model. In Table 3, curvilinear model has the highest correlation coefficient of 97%. This means it can predict GA performance with the highest possible accuracy. To further investigate the adequacy of the selected model, tests of normal plot of residuals were also performed on the models. The spread of calculated and actual values of final tours around the regression lines for the three functions are shown in Figures 3 to 5. As illustrated, the best model to relate GA parameters settings and its performance characteristic, found to be second degree polynomial function. Therefore, further statistical analyses would be performed on this model only.

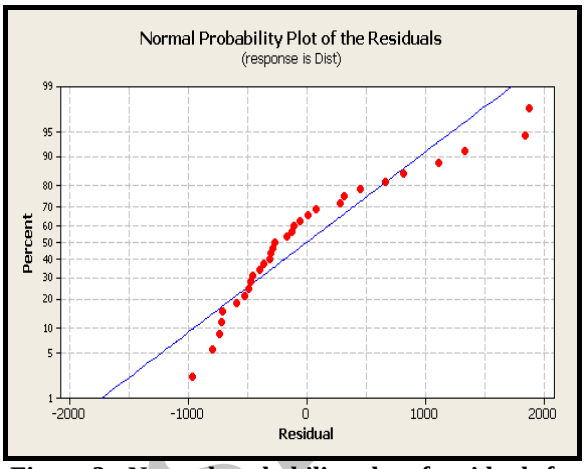

Figure 3: Normal probability plot of residuals for linear model

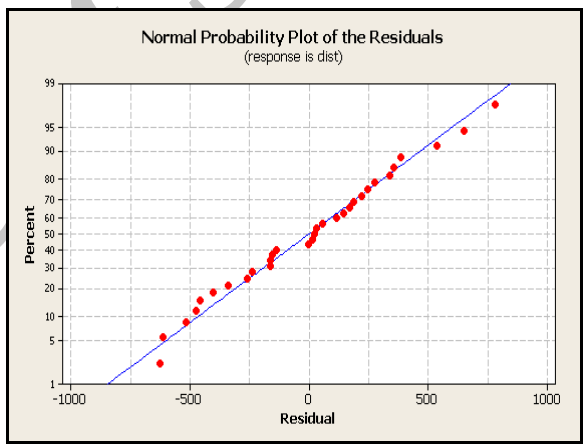

Figure 4: Normal probability plot of residuals for curvilinear model

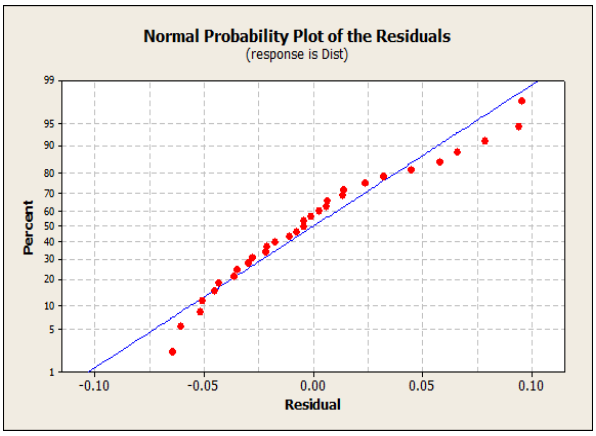

Figure 5: Normal probability plot of residuals for logarithmic model

The significance of each parameter in curvilinear model is determined using t-test and P-values which are listed in Table 4. 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.

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

Table 4: Least squares fit and parameters estimates (significance of regression coefficients)

With respect to the above results, the most important parameter affecting GA performance is the probability of mutation (P_m). Statistical analysis shows that both first order and second order of P_m 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-P_m), probabilities of crossover and mutation (Cr-P_m) are also significant. These interactions have positive effects on response variable.

Conclusion

In this research, the relations between input parameters and solution quality (output) of Genetic Algorithm have been established using a TSP benchmark problem (att 48). 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. Next, various functions were fitted on the data to model the optimization process. Among various regression function, curvilinear is found to be the best model based on correlation coefficient and Analysis of Variance (ANOVA) criteria. Statistical analyses show that mutation probability as well as interaction effects between population and crossover, population and mutation and between mutation and crossover are significant factors. The proposed approach is promising in the sense that it may be used to determine the proper set of parameter settings for a given optimization problem. Nevertheless, it should be noted that the performance of such procedures is case-dependent and may vary with problem size and its structure.

References:

1- Moon, C., Kim, J., Choi, G. and Seo, Y. (2002). "An efficient genetic algorithm for the traveling salesman problem with precedence constraints." *European Journal of Operational Research,* Vol. 140, PP. 606–617.

2- Liu, F. and Zeng, G. (2009). "Study of genetic algorithm with reinforcement learning to solve the TSP." *Expert Systems with Applications, Vol. 36,* PP. 6995–7001.

3- Meer, K. (2007). "Simulated Annealing versus Metropolis for a TSP instance." *Information Processing Letters,* Vol. 104, No. 6, PP. 216-219.

4- Blum, C. (2005). "Ant colony optimization: Introduction and recent trends." *Physics of Life Reviews,* Vol. 2, No. 4, PP. 353-373.

5- Tang, H., Miller-Hooks, E. (2005). "A Tabu Search heuristic for the team orienteering problem." *Computers & Operations Research*, Vol. 32, No. 6, PP. 1379-1407.

6- Todd, D. (1997). "Multiple Criteria Genetic Algorithms in Engineering Design and Operation." Faculty of Engineering, University of Newcastle upon Tyne, UK, Newcastle.

7- Pongcharoen, P., Stewardson, D. J., Hicks, C. and Braiden, P. M. (2001). "Applying designed experiments to optimize the performance of genetic algorithms used for scheduling complex products in the capital goods industry." *Journal of Applied Statistic,* Vol. 28, No. 3, PP. 441-55.

8- Ghrayeb, O. and Phojanamongkolkij, N. (2005). "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,* Vol. 12, PP. 37-44.

9- Kaya, M. (2010). "The effects of two new crossover operators on genetic algorithm performance." *Applied Soft Computing,* accepted for publication.

10- Albayrak, M.and Allahverdi, N. (2010). "Development a new mutation operator to solve the Traveling Salesman Problem by aid of Genetic Algorithms." *Expert Systems with Applications,* accepted for publication.

11- Pongcharoen, P., Chainate, W. and Thapatsuwan, P. (2007). "Exploration of Genetic Parameters and Operators through Travelling Salesman Problem." *Science Asia,* Vol. 33, PP. 215-222.

12- Alfaro-Cid, E., McGookin, E.W. and Murray-Smith, D.J. (2009). "A comparative study of genetic operators for controller parameter optimization." *Control Engineering Practice,* Vol. 17, No. 1, PP. 185-197.

13- Laporte, G. (1992). "The traveling salesman problem: An overview of exact and approximate algorithms." *European Journal of Operational Research,* Vol. 59, No. 2, PP. 231-247.

14- Chatterjee, S., Carrera, C. and Lynch, L. A. (1996). "Genetic algorithms and traveling salesman problems." *European Journal of Operational Research,* Vol. 93, No. 3, PP. 490-510.

15- www.iwr.uniheidelberg.de/iwr/comopt/soft/ TSPLIB95/TSPLIB.html

16- Goldberg, C. (1989). Genetic Algorithms in Search, Optimization and Machine Learning, Wesley: Addison.

17- Randy, L. H. and Haupt Sue Ellen (2004). *Practical Genetic Algorithm,* JOHN WILEY & SONS.

18- Dantzig, G., Fulkerson, R. and Johnson, S. (1954), "Solution of a largescale traveling-salesman problem." *Operations Research,* Vol. 2, No. 4, PP.393–410.

19- Bektas, T. (2006). "The multiple traveling salesman problem: an overview of formulations and solution procedures." *Omega* Vol. 34, No. 3, PP. 209-219.

20- Montgomery, D.C. (2001). *Design and analysis of experiments.* 5th. Ed. JOHN WILEY & SONS.