Advanced Correlation Analysis of Operators for the Traveling Salesman Problem

John Dzubera and Darrell Whitley

Computer Science Department, Colorado State University, Fort Collins, CO 80523

Abstract. An extension to the correlation analysis of Manderick et al. and Mathias and Whitley used to study TSP operators is introduced. Part correlation and partial correlation further refine the relationship between an operator's preservation of genetic information and its overall fitness. The fitness landscapes of ten genetic operators are studied in isolation and in conjunction with the *2-opt* local improvement technique in a GENITOR environment. Five variables (parent fitness, child fitness, child failure rate, insertion rate and increased tour length due to failure) are considered in the analyses. Results suggest that certain operators (Edge Family, MPX and Schleuter crossover (SCHX)) maintain higher correlations over several *interesting* variable relations, especially the parent to child (PC) relation, than do other genetic operators.

1 Introduction

With their introduction of correlation analysis, Manderick et al. [3] have provided a useful metric for determining the fitness of a genetic operator. Borrowing from statistics theory, Manderick proposed the *correlation coefficient,* ρ_{op}, as a method of analyzing the correlation between parent and child strings.

$$\rho_{op}(P, C) = \frac{Cov(P, C)}{\sigma(P)\sigma(C)}$$

where *op* is the genetic operator, P and C are variables that represent the fitnesses of the parent and child respectively, $Cov(P, C)$ is the covariance between P and C and σ represents the standard deviation. With respect to most non-flat landscapes, a larger value of ρ_{op} implies a greater preservation of genetic material between parents and offspring as well as an increased ability to exploit the fitness landscape [4]. The goal of this research is to extend this work by introducing the concepts of *part correlation* and *partial correlation* as more accurate metrics for genetic operator analysis. Unlike traditional correlation methods which determine the simple relationship between two variables, part and partial correlation consider the degree of association between two variables *while the effects of other variables are held constant.* This "parting out" or "partialling out" of possible extraneous influences allows the true relationship of the variables to be ascertained. Although computationally explosive as the number of variables is increased, part and partial correlation supply the means by which mathematically stronger statements about these relationships may be made. To facilitate this analysis, ten genetic operators were examined in isolation and in conjunction with the *2-opt* local improvement operator on the Padberg 532 city TSP. In addition to the primary research, two other issues considered were *statistical inference* and *statistical significance.* Manderick [3] neglects to address whether his results, which are based on sample data, are applicable to the entire population. Although his conclusions may be correct primarily due to adequate sample size [5], significance tests add additional rigor to findings based on sample data. All differences presented in this work have been tested and found to be statistically significant at a .05 level.

2 Fundamental Concepts

Genetic algorithms, in their purest form, may be thought of as *blind* search methods. That is, they do not exploit any local information to construct solutions. Several researchers [4] [8] have recently explored combinations of genetic algorithms and local improvement techniques (*hybridized genetic algorithms.*) In the work reported here, the *2-Opt* algorithm was utilized as the local search operator. Given a TSP tour, 2-Opt chooses two edges to remove from the string. One of the two segments is then reversed. If the fitness of the modified tour is "better" than the original, the change is kept; otherwise it is discarded. 2-Opt continues in this manner for all $\frac{(n)(n-1)}{2}$ edge pairings. To limit the expense of 2-Opt, a *single-pass* of 2-Opt was used. Since more improvement occurs during the first pass than subsequent passes [4], *one-pass 2-Opt* provided a suitable balance between performance and expense.

With respect to partial correlation analysis, several variables were studied in both hybridized and non-hybridized environments. In addition to parent fitness (P) and child fitness (C), *insertion rate* (R), *failure rate* (F), and *accumulated length due to failures* (L) were also considered. Insertion rate refers to offspring insertion frequency, or how often the newly constructed child fitness is greater than the least fit member of the population. This provides insight into the ability of the genetic algorithm to find improved points in the solution space [4]. Failure rate denotes the number of *foreign edges* introduced into the offspring during recombination. A failure or foreign edge occurs when an edge (e.g. city A - city B) is present in the child, but does not appear in either of its parents. Closely related to failure rate is accumulated length due to failures. This variable records the additional tour length added to the child because failures have occurred. Both failure rate and accumulated length due to failures measure how well the crossover operator preserves information from parent to child and how serious a penalty is incurred when a foreign edge is introduced.

3 Part and Partial Correlation

In this paper, part correlation (or *semi-partial correlation*) theory will be developed from an arithmetic perspective with some references to linear regression interpretations [5] [7]. Consider a sample of strings from a random population that is measured according to three variables which are considered important in understanding how an underlying genetic operator functions. These variables are the parents' fitness (P)[1], the child's fitness (C), and the failure rate (F). Suppose that one area of interest is to determine the extent of the relationship between P and C when variation in F, which is related to variation in C, is held constant. The technique for solving this problem is called *part correlation,* as the area of interest is the relationship between two variables (P and C) with the effect of a third (F) "parted out" of *one* of the other two, in this case C. Symbolically:

$$r_{P(C \cdot F)} = \frac{r_{PC} - r_{PF}\, r_{CF}}{\sqrt{1 - r_{CF}^2}},$$

where r_{PC} is the regular correlation coefficient between P and C. An informal way to think about part correlation is as follows. Consider a plot of the distribution of F on a single axis. *Group* the data into significantly different clusters, with all points in a cluster fitting a normal curve (this could be inferred by choosing a large enough cluster (≥ 30) and applying the *Central Limit Theorem*). Next construct a set of plots, one for each F cluster, consisting of raw C scores. Finally, suppose that the C graphs are constructed on transparencies and that they are merged to form a single plot. This last graph represents C's *deviation scores* with F held constant, denoted $(C \cdot F)$. Thus, to calculate the part correlation of P and C with the influence of F removed from C, one would correlate P's raw scores with C's deviation scores.

While the use of part correlation may provide greater insight into the nature of data relationships than simple correlation, its power is limited. Not only does it lack the ability to "part out" a third variable from both of the variables under consideration, but it also does not admit higher-order relationships. *Partial correlation* addresses both of these shortcomings.

3.1 First-Order Partial Correlation

In the previous example, a relationship was determined between the variables of parent fitness (P) and child fitness (C) with the effect of failure rate (F) removed from C. Now suppose that one wishes to find the relationship between P and C when any association between P and F, as well as between C and F, is removed. Unlike part correlation, *partial correlation* removes the effects of a third variable from both variables under consideration ("F partialled out of P and C"), rather than just one. The formula for the above relationship is:

$$r_{PC \cdot F} = \frac{r_{PC} - r_{PF}\, r_{CF}}{\sqrt{1 - r_{PF}^2}\sqrt{1 - r_{CF}^2}}$$

where, again, each r on the right side of the equation is a normal correlation.

[1] Since two parents are used in the construction of an offspring, P corresponds to the *average* of the parents' fitness.

3.2 Second- and Higher-Order Partial Correlation

Although the above example considered only three variables, there is no theoretical need for such limits. Suppose now that a fourth variable is introduced, accumulated length due to child failures (L). Assume that the area of interest is to determine the relationship between P and C when the effects of variation in F and L are removed from both P and C. Again, a theoretically simple equation solves the problem:

$$r_{PC \cdot FL} = \frac{r_{PC \cdot F} - r_{PL \cdot F}\, r_{CL \cdot F}}{\sqrt{1 - r^2_{PL \cdot F}}\sqrt{1 - r^2_{CL \cdot F}}}.$$

However, each r term on the right side of the equation is now a *first-order* partial correlation, not a normal correlation as before.

The partial correlation concept is not limited to any finite number of variables; one could add the child's insertion rate R into the calculation, producing a *third-order partial correlation*:

$$r_{PC \cdot FRL} = \frac{r_{PC \cdot FR} - r_{PL \cdot RF}\, r_{CL \cdot RF}}{\sqrt{1 - r^2_{PL \cdot RF}}\sqrt{1 - r^2_{CL \cdot RF}}},$$

where each correlation on the right hand side is now a *second-order* partial correlation. Indeed, one might consider fourth, fifth, or even higher partials; however, such partials are not usually studied due to computational expense.

3.3 Significance Tests

The Fisher z-transformation [1] will be utilized to convert the calculated correlation coefficient to a statistical form which possesses a normal distribution. This is necessary because the significance tests used in this research depend upon normally distributed data. The formula for the z-transform is:

$$z_f \;=\; (1.1513)\;\; log[(1 + r)/(1 - r)].$$

To complete the conversion of the correlation to a normal form, the converted data must be divided by the standard error of the z-transform corresponding to the given partial. Thus, the calculated partial z-transform is:

$$pz_f \;=\; \frac{(1.1513)\;\; log[1 + r/1 - r]}{[1\,/\,\sqrt{N - (m + 3)}]},$$

where N is the sample size and m is the number of variables "partialled out." The calculated value of pz_f is then tested at a .05 level of significance to determine the relationship (if any) between the sample value and population value.

The other significance test needed is one which determines if two different population partial correlation coefficients derived from two sets of sample data are significantly different. The following formula calculates the relevant statistic d, which is then tested at a .05 level of significance:

$$d \;=\; \frac{z1_f - z2_f}{\sqrt{\frac{1}{(N_1 - m - 3)} + \frac{1}{(N_2 - m - 3)}}}, \quad N_1, N_2 > m + 3.$$

4 Edge-4 And Schleuter Crossover (SCHX)

Most of the operators studied in this research have been previously described [3] [4] [6] except for Edge-4 and SCHX. Edge-4 is a new operator presented here for the first time, while SCHX has been considered in only brief detail [2]. Both operators give good results with respect to several *interesting* correlations, but neither perform consistently as well as MPX in the context of hybridized search.

4.1 Edge Crossover

The Edge family of operators emphasizes adjacency information of the cities in the tour [6]. Briefly, Edge works by collecting the edge information contained in both parents in an *edge table*. The rows of the table consist of a city (e.g. A) followed by a list of cities which are connected to A in the two parent tours.

To construct an offspring, the original Edge algorithm first chooses a random starting city (e.g. A) and removes all instances of A from the "link" lists in the edge table. The next city is then chosen by finding the city connected to the current city with the least number of remaining connections. If all connecting cities have the same number of links, the next city is chosen randomly. This city is now removed from all of the remaining cities' link lists, and the algorithm proceeds as before. If at any time, a city is chosen such that its link list is empty, a *failure* occurs. In this case, the original Edge operator does not attempt to fix the failure, but instead chooses a new city, adds it to the child, and continues the method described above. The algorithm halts when all cities have been included in the offspring.

Successive versions of Edge have attempted to improve the amount of genetic information preserved between parent and child by limiting the number of failures. Edge 2 added the ability to preserve common subtours between parents by "flagging" duplicate link list entries with a minus (-) sign and giving priority to those entries when the next city in the tour is chosen. Although a minor change, the flagging operation improved the Edge algorithm's performance [6].

Edge-3 and Edge-4 enhanced the algorithm further by attempting to minimize failures (the primary source of increased tour length) through intelligent failure resolution methods. Unlike its ancestors, when Edge-3 encounters a failure it does not randomly choose a new city. Instead, it reverses the current tour segment and attempts to extend the tour by adding cities to the other end of the segment. The algorithm continues to add cities until another failure is reached, at which point a true failure occurs. Since the tour cannot be reversed again, the default behavior of choosing a random city to continue the tour is employed.

Edge 4 takes this idea still further by considering the end cities after failures have occurred on both ends. Since the end city has resulted in a failure, it must be the case that all of its links are already contained in the partially built tour. Edge 4 now attempts to reverse *only part* of the tour such that the end city will be adjacent to one of its original links, thereby postponing and possibly averting random failure resolution. For example, suppose that the original edge table entry for A contains edges to B, C, D, and E, where the partial tour consists of

$$\text{Q X } \underline{\text{B}} \text{ Y Z } \underline{\text{C}} \text{ W } \underline{\text{D}} \text{ R } \underline{\text{E}} \text{ A.}$$

Both A and Q are failure cities. The Edge 4 algorithm then inspects the link lists of the cities adjacent to the cities in A's link list. For example, the link lists of Y, W, and R would be examined and the one with the smallest number of entries would be chosen. Suppose Y is picked using this criteria. Edge 4 now reverses the subtour Y Z C W D R E A to produce

$$\text{Q X } \underline{\text{B}} \text{ A } \underline{\text{E}} \text{ R } \underline{\text{D}} \text{ W } \underline{\text{C}} \text{ Z Y.}$$

The effect of this reversal is twofold. First, A is no longer a failure city as the edge BA was present in one of the original parents. Second, the tour may now be extended if Y still contains cities in its link list. If this reversal does not admit any extension of the tour, another A link will be chosen and a different reversal will be attempted. For example, if C is chosen, the tour would then be

$$\text{Q X } \underline{\text{B}} \text{ Y Z } \underline{\text{C}} \text{ A } \underline{\text{E}} \text{ R } \underline{\text{D}} \text{ W}$$

and the algorithm would try to extend the tour from W. Backtracking would continue until an extension is found or all other edge swaps (excluding the adjacent edge E) produce no extension. Next, the algorithm would use the same method starting from the other end of the tour. If extensions still could not be constructed, the algorithm chooses a random city and proceeds as before.

4.2 Schleuter Crossover

The Schleuter crossover operator (SCHX) [2], is a variation on Mülenbein's MPX operator, although it has features similar to those found in Davis' order crossover [4] and it includes partial tour reversals. (The source code of Schleuter Crossover was derived from Pascal code received from E. Pesch.) SCHX begins by denoting the two selected parent strings as the *donor* and the *receiver*. A random city and a random segment length ($3 \leq |segment| \leq (\frac{1}{3})tourlength$) are then chosen in the donor. For example, suppose the randomly chosen donor city is A (at position 10) and the length of the segment is 81 cities, ending in B (at position 90). Expressing the cities between A and B as $x_1, x_2, \cdots, x_{79}$, the donor is then

$$(\ \cdots \ A(10) \ \ x_1(11) \ \ x_2(12) \ \ \cdots \ \ x_{79}(89) \ \ B(90) \ \ \cdots \ )$$

where the number in parentheses represents the position of that city in the donor. The algorithm then finds the cities on either end of the segment (A and B) and locates them in the receiver. Suppose A and B are found in the receiver at positions 25 and 245 respectively. The receiver can be represented as:

$$(\ \cdots \ A(25) \ \ y_1(26) \ \ y_2(27) \ \ \cdots \ \ y_{219}(244) \ \ B(245) \ \ \cdots \ ).$$

Next two copies of the receiver are made (temp1 and temp2), but they are altered in the following way. In temp1, the partial tour $y_1 \ \ y_2 \cdots y_{219} \ \ B$ is reversed to relocate city B adjacent to city A:

$$(\ \cdots \ A(25) \ \ B(26) \ \ y_{219}(27) \ \ \cdots \ \ y_2(244) \ \ y_1(245) \ \ \cdots).$$

Similarly, in temp2 the tour is reversed to move city A adjacent to city B:

$$(\cdots \ y_{219}(25) \ y_{218}(26) \ \cdots \ y_1(243) \ A(244) \ B(245) \ \cdots).$$

Finally, two offspring are constructed as follows. First, copy the donor segment into the first offspring at the position where city A was located in the receiver:

$$(\cdots \ A(25) \ x_1(26) \ x_2(27) \ \cdots \ x_{79}(104) \ B(105) \ \cdots).$$

The remainder of the first child is constructed by examining each city after the AB edge in temp1 and adding it to the child after B if the city is not already contained in the tour. Similarly, the second offspring receives the donor segment such that city B from the donor corresponds to city B in the receiver:

$$(\cdots \ A(165) \ x_1(166) \ x_2(167) \ \cdots \ x_{79}(244) \ B(245) \ \cdots).$$

To complete the construction of the second child, each city after the AB edge in temp2 is examined and added to the child after city B if it does not exist in the offspring. In both cases, the *relative order* of the temporary strings is preserved. Finally, both children are evaluated and the more fit child is retained.

5 Experimental Method and Data Collection

To examine the part and partial correlation behavior and the fitness landscapes of the ten genetic operators with and without the *2-opt* local improvement operator[2], the following experimental setup was devised. First, four areas of interest were selected: ordinary GENITOR (called PC), GENITOR with an initially 2-opted population (P2C), GENITOR with 2-opt performed on the resulting child (PC2), and GENITOR with an initially 2-opted population and 2-opt performed on the resulting child (P2C2). For each interest area, two different sets of runs were performed. One set was used to generate correlations over 30 independent runs of 2500 recombinations each from an initial population. This provided insight into how each of the operators work when the population is diverse. The other set generated correlations over 2 independent runs of 250,000 recombinations each. This showed how the operators and their associated correlations performed *over time* as the population became more homogeneous and converged toward a solution. Each sample run began with the same initial population. For each pair of parents in the sample, 2-opt was performed on the parents (if applicable), followed by the recombination of two parents to produce offspring. 2-opt was then applied to the resulting child (if applicable), and statistical data was collected. This data included both parents' fitness, child fitness, the number of failures in the child, whether or not the resulting child was inserted in the GENITOR population, and the total TSP tour length increase due to failures in the child. Finally, the part and partial correlations were calculated and significance tests were performed on the *interesting* relations.

[2] 2-opt in this case means *one-pass 2-opt,* not *2-opt to convergence.* For an in-depth discussion of the differences among 2-opt operators, as well as the workings of MPX and Edge, see [4].

Correlation	MPX	SCHX
PC	0.299089	0.423785
$P(C \cdot F)$	0.493316	0.454906
$C(P \cdot F)$	0.275070	0.382705
$PC \cdot FRL$	0.634582	0.431662

Table 1. P2C Correlation Performance / 2500 Recombinations

5.1 Utility of Part and Partial Correlation

Part and partial correlation justify their expense over a series of related correlations, as can be seen from Table 1. When simple correlation is considered, SCHX appears to admit the greater correlation. A calculation of d (see Section 3.3) would show that the difference is significant and therefore SCHX possesses a higher PC correlation than MPX does for this environment. However, this conclusion would be misleading. Under the $P(C \cdot F)$ relation, MPX exhibits greater correlation than SCHX. Clearly, the failure rate in the child negatively masks part of the PC correlation and once that effect is removed, a much different picture develops. Conversely, the $C(P \cdot F)$ relationship demonstrates a different masking effect. Both operators' correlations are decreased when failure rate is removed from the parents; however, the effect is greater in MPX than in Schleuter crossover. Finally, when all other variables are "partialled out," MPX exhibits the greater correlation. From this analysis, two important pieces of information are obtained. First, the true correlation between parents and child is much higher in MPX than in SCHX, even though simple correlation analysis would have shown the reverse. Also, by examining a series of part and partial correlations, the masking effects of other variables on the correlation relationship may be determined.

5.2 Results

From the sample data collected, several significant trends are uncovered. As seen from Table 2, MPX appears to be the best overall genetic operator with respect to preserving information from parent to child. Not only does MPX perform well with respect to ρ_{op} in environments with no local optimization over a short initial sample (PC, 2500), but MPX is also the operator of choice when the initial population has been locally optimized (P2C, P2C2). There are two reasons for MPX's success. First, the operator performs a straight copy of one-third of the donor into the final child. This results in a correlation of one for all variables over the subtour domain. Second, MPX does not rely on simple relative order to complete offspring tours, but instead looks in both the donor and receiver to preserve edge and order information, thereby maximizing genetic preservation.

The Edge family, on the other hand, exhibits comparable or superior performance on the PC relation when no local optimization is performed. While the

Relation	# Recom.	Correlation	Rank
PC	2500	0.634582	1
PC	250000	0.996977	3
$P2C$	250000	0.951699	1
$P2C2$	250000	0.646876	1

Table 2. MPX $PC \cdot FRL$ correlation performance

Edge Corr. / $PC \cdot FRL$/ 250K Recom.				
Operator	PC Correlation	PC Rank	P2C Correlation	P2C Rank
Edge-4	0.998321	1	0.434894	6
Edge-3	0.997705	2	0.528044	4
Edge-2	0.996653	4	0.531802	3

Table 3. Edge family performance in local and non-local environments.

Edge family is superior in non-local environments, Edge-4's correlations may suggest that the point of diminishing returns has been reached. Edge-4 consumes greater resources than Edge-3, yet as Table 3 shows, Edge-4 provides only an .001 gain (the difference is statistically significant, however).

Still, there is room for improvement in the Edge family as seen by their performance under local optimization. For the P2C relation, Edge's best correlation is 0.531802. This pales in comparison to MPX's correlation of 0.951699 and SCHX's 0.894806 in the same environment. One reason for Edge-4's poor execution may be its "decomposing" of the parent tours into component edge parts. Unlike MPX, Edge-4 possesses no direct copy mechanism from parent to child. Such a deficiency does not make a difference in the PC case, but when local optimization is used the absence of direct copy may severely hamper parent-child correlation. This suggests that a direct copy mechanism added to Edge-4 might improve its performance when used with optimized initial populations. The other possible cause for Edge-4's inferiority in hybridized environments may be the operator's exclusive emphasis on edge information instead of relative order information. It may be that after a failure occurs, relative order can help determine which city should be chosen to start the next tour segment.

Although not shown, SCHX's performance is comparable to MPX's. Whereas SCHX is ranked behind Edge in non-local environments, it performs better than Edge when 2-Opt is used. While not an optimal genetic operator, SCHX's combination of direct copy, internal swapping and relative order child completion allows it to distance itself performance-wise from the less fit genetic operators.

Although the above results highlight the importance of partial correlation, a central question remains: "Does PC correlation imply greater performance?" The answer is a qualified "yes." Mathias and Whitley [4] have shown the superiority of Edge in a environment without 2-opt and of MPX when 2-opt is used. Their results support the findings of this work. However, caution must be taken not to assume that a greater preservation of genetic material *always* corresponds to improved performance. Consider the genetic operator (Unity) that constructs a child by cloning one parent. Under correlation analysis, most relations would be unity since no change has occurred from parent to child. Yet no improvement would be obtained beyond the fitness of the best parent in the initial population. Thus, while correlation analysis provides insight into performance, it is clearly not the only feature to consider when choosing a genetic operator.

Another comparison concerns the third order FL correlations. Without exception, MPX exhibits a greater correlation between the number of failures and the accumulated length due to failures than the Edge family. While this implies that a certain number of failures will almost always map to the same failure length, it does not provide any insight into the magnitude of the length. For example, suppose that MPX exhibits a characteristic such that, for 99% of the time, when five failures are present in the child, length due to failure is 100K. For another operator, the length might be 10 for 5 failures. In either case, the correlations themselves would be equivalent. Thus, this relationship shows only that MPX's failures and failure length are more correlated than Edge's, not that this characteristic is necessarily positive or negative.

The other genetic operators (OrderDav, OrderSyz, Pmx, Cycle, Position) [6] considered in this work do not compare very well to SCHX, MPX or Edge. Under the PC relation, for example, these operators almost never rank higher than sixth, with the occasional exception of Davis' order operator. Overall, the results of this work tend to reinforce Manderick's and Mathias and Whitley's conclusion of the superiority of Edge and MPX over other genetic operators.

References

1. M. Fogiel. (1991) *The Statistics Problem Solver.* Research and Education Assoc.
2. Martina Gorges-Schleuter. (1989) "ASPARAGOS An Asynchronous Parallel Genetic Optimization Strategy." In *Proceedings of the Third International Conference on Genetic Algorithms.* J. David Schaffer, ed. Morgan Kaufmann.
3. B. Manderick, M De Weger, and P. Spiessens. (1991) "The Genetic Algorithm and the Structure of the Fitness Landscape." In *Proceedings of the Fourth International Conference on Genetic Algorithms.* R. Belew and L.Booker, ed. Morgan Kaufmann.
4. K. Mathias and D. Whitley. (1992) "Genetic Operators, The Fitness Landscape and the Traveling Salesman Problem." In *Parallel Problem Solving from Nature, 2* R. Manner and B. Manderick, ed. Elsevier Science Publishers.
5. D. F. Morrison. (1967) *Multivariate Statistical Methods.* McGraw-Hill.
6. T. Starkweather, S. McDaniel, K. Mathias, D. Whitley, and C. Whitley. (1991) "A Comparison of Genetic Sequencing Operators." In *Proc. of the Fourth Intl. Conf. on Genetic Algorithms.* R. Belew and L.Booker, ed. Morgan Kaufmann.
7. R. M. Thorndike. (1978) *Correlational Procedures for Research.* Gardner Press.
8. N. Ulder, E. Aarts, H. Bandelt, P. van Laarhoven, E. Pesch. (1991) "Genetic Local Search Algorithms for the Traveling Salesman Problem." In *Parallel Problem Solving from Nature* H.P. Schwefel and Reinhard Manner, ed. Springer/Verlag.