INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Research and Publications

An empirical investigation into randomly generated Euclidean symmetric traveling salesman problems

Megha Sharma
Diptesh Ghosh

W.P. No. 2006-06-03
June 2006

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD – 380015
INDIA

An empirical investigation into randomly generated Euclidean symmetric traveling salesman problems

Megha Sharma
Diptesh Ghosh

P&QM Area, Indian Institute of Management, Ahmedabad 380015, INDIA
{meghas,diptesh}@iimahd.ernet.in

Abstract

The traveling salesman problem is one of the most well-solved hard combinatorial optimization problems. Any new algorithm or heuristic for the traveling salesman problem is empirically evaluated based on its performance on standard test instances, as well as on randomly generated instances. However, properties of randomly generated traveling salesman instances have not been reported in the literature. In this paper, we report the results from an empirical investigation on the properties of randomly generated Euclidean traveling salesman problem. Our experiments focus on the properties of the edge lengths and the distribution of the tour lengths of all tours in instances for symmetric traveling salesman problems.

Keywords: Euclidean Traveling Salesman Problems, Random instances, Empirical distributions, Generalized Beta distributions

1 Introduction

The traveling salesman problem (TSP) is one of the best known problems in combinatorial optimization. In a TSP instance, we are given a weighted graph $G = (V, E, c)$, where V represents a set of n vertices, $E = V \times V$ represents the set of arcs in the graph, and $c : E \rightarrow \Re$ represents the cost of each arc in E. The objective in a TSP instance is to find the shortest simple cycle in the graph covering all the vertices in V. Such simple cycles are called tours in the TSP context. The number n is referred to as the size of the TSP instance. The TSP is a collection of all TSP instances. If for any two vertices i and j, $c(i, j) = c(j, i)$, then the instance is called symmetric, and the arcs (i, j) and (j, i) are collectively referred to as an edge between vertices i and j. If for any three vertices i, j , and k in a symmetric TSP (STSP) instance,

$c(i,j) \leq c(j,k) + c(k,i)$, then the STSP instance is called Euclidean (ESTSP). The TSP is known to be NP-Hard (see Karp [3]).

Since the TSP is so well-known, many solution algorithms exist for it (see for instance, Gutin and Punnen [1]). Some are exact algorithms, which are guaranteed to generate an optimal i.e., shortest tour, and others are heuristics, which generate near optimal tours. The algorithms are compared with each other based on their performance on benchmark instances, such as the ones in TSPLIB (see Reinelt [5]), or on randomly generated TSP instances. To generate a random instance of the ESTSP of size n, one generates n points at random in a square of pre-determined size, and creates the edge cost matrix by measuring the distance between each pair of points. The distances are measured as per the Euclidean norm, which means that between any two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, the distance is given by

$$d(P_1, P_2) = [(x_1 - x_2)^2 + (y_1 - y_2)^2]^{1/2}.$$

Since it is computationally more convenient to deal with integers rather than with floating point numbers, the distances are rounded down in empirical studies. Thus, the cost function $c(i,j)$, where vertex i is represented by the vector (x_i, y_i) and the vertex j is represented by the vector (x_j, y_j) is

$$c(i,j) = \lfloor [(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2} \rfloor.$$

To the best of our knowledge, there is no literature on the properties of such randomly generated instances. Some pertinent questions that arise are the following.

- What is the distribution of the lengths of the edges in randomly generated ESTSP instances, and are these lengths dependent on the size of the problems generated?
- What is the distribution of the tour lengths of all the tours in a randomly generated ESTSP instance?
- How are the tour lengths of optimal solutions to various randomly generated ESTSP instances of the same size distributed?
- How difficult is it to solve an ESTSP instance?

We carry out an empirical investigation in search of solutions to these questions. While similar questions are also unanswered for randomly generated TSP instances using other distance metrics, we have no reason to suspect that the answers would be very different from the ones that we obtain here.

The remainder of the paper is organized along the following lines. In Section 2 we study the distribution of the lengths of edges in randomly generated ESTSP instances. Next in Section 3, we study the distribution of the lengths of optimal tours to randomly generated ESTSP instances. We also study the distribution of

the lengths of all feasible tours in such instances. The results that we obtain in Sections 2 and 3 is summarized in Section 4. This section also points to a direction for future research.

2 Edge Lengths

In all our experiments, ESTSP instances with n vertices are generated by randomly scattering n vertices on a square of side 100 units. The Euclidean distances between each pair of vertices are then rounded down to obtain the edge lengths. The edge lengths thus lie in $[0, 141]$. Since the vertices are scattered randomly, the edge lengths in a randomly generated ESTSP are expected to be identically distributed.

Since the Euclidean distance between two vertices in the square is a reasonably complicated function of the four coordinates involved, and also since we round down the floating point values of the inter-vertex distances to obtain integer edge lengths, obtaining analytical expressions for the distribution of the length of an edge is intractable. In this section therefore, we obtain the shape of the distribution empirically, and examine whether it can be approximated using any known distribution function. In order to study the distribution of edge lengths empirically, we measured the lengths of ten million randomly generated edges. Each edge was generated by randomly generating two vertices in a square of side 100 units and the length of the edge was obtained by measuring the Euclidean distance between them. The probability distribution for the length of such a random edge is shown in Figure 1. It had an expected value of 51.6411 units, a standard deviation of 24.7939, and a skewness is 0.18493. This experiment of generating ten million random edges and finding the

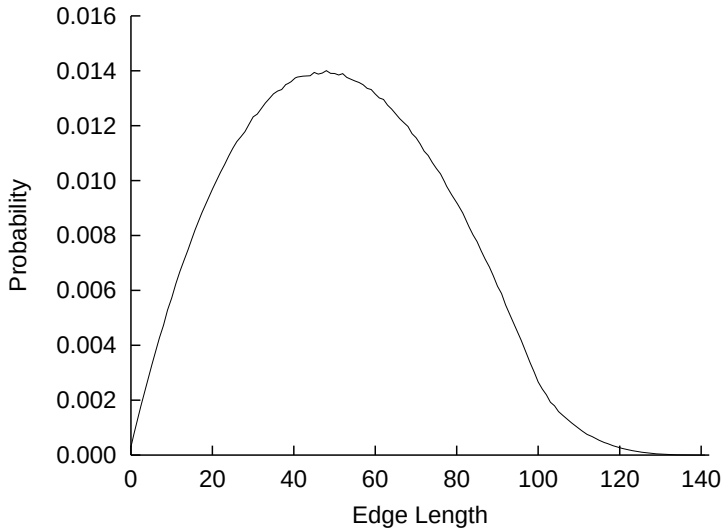


Figure 1: Probability distribution of edges in a random ESTSP

probability distribution of a random edge was repeated thirty times. Each time the

expected value of the distribution, its standard deviation, and skewness was noted. Table 1 presents the summary statistics of each set of thirty values obtained from the experiments. Notice that the mean of the expected values of the distributions is 51.6433 units, not very different from the 51.6411 units that we had obtained in our first experiment. In fact, for all the three parameters, the values obtained in the thirty experiments were very close together, as is evident from the small values of standard deviations, and the nearness of the maximum and minimum values. Thus, given a square of size 100 units, we conjecture that the lengths of the edges would be 51.64 units on average, if the lengths are measured using the Euclidean norm.

Table 1: Summary statistics of the mean, standard deviation and skewness of the distribution of edge lengths in a random ESTSP described in a square of side 100 units.

Parameter	Mean value of the parameter	Std Dev of the parameter	Min value of the parameter	Max value of the parameter
Expected value	51.6433	0.00684	51.6295	51.6563
Standard Deviation	24.7949	0.00429	24.7871	24.8059
Skewness	0.1844	0.00050	0.1832	0.1852

To test this conjecture, we generated ESTP instances of size 5, 10, and 15. For each problem size, we generated ten million instances of ESTSPs on squares of side 100 units. For each problem size, we generated the empirical probability distribution of the lengths of each of the edges from the ten million instances, and computed the expected value, the standard deviation, and skewness of the distributions. We noticed that there was no change in the parameters with increasing problem size; in fact the values of the parameters never deviated by more than 0.05 from the values quoted in Table 1. We repeated these experiments thirty times, and noted that the parameter values that we quote here were indeed very stable.

The empirical distribution of edge lengths of random ESTSPs is seen to be smooth and unimodal, and with finite support. Hence it seems possible to model the distribution using a Generalized Beta (GB) distribution. A GB distribution supported on $[a, b]$ with with shape parameters α and β has the following density function (see, e.g., Johnson et al. [2]):

$$f(x; \alpha, \beta, a, b) = \frac{1}{B(\alpha, \beta)} \frac{(x-a)^{\alpha-1} + (b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}},$$

In our case, the natural choice of a and b are 0 and 141 respectively. However, the densities of a GB distribution at the endpoints of its support equal 0, while in our case, as is evident from the example which we have plotted in Figure 1, the density at the 0 is clearly positive. Hence we let $a = -1$ for fitting the distribution. We obtained the values of α and β using the method of moments, computing them from the mean and standard deviation of our sample of 10 million data points.

We fitted a GB distribution to the distribution of edge lengths obtained in our first experiment. The estimates of α and β that we obtained were 2.466 and 4.185 respectively. The estimated GB distribution underestimated the empirical distribution for both low values and for high values of edge lengths, but overestimated the empirical distribution for medium values of edge lengths. The amount of overestimation and underestimation was minor; the maximum deviation in the distribution functions of the empirical and estimated distributions never exceeding 0.01663. The distribution functions of the empirical distribution and the estimated distribution are shown in Figure 2. From the figure, it is evident that the GB distribution is a good candidate for modeling the lengths of individual edges in ESTSPs.

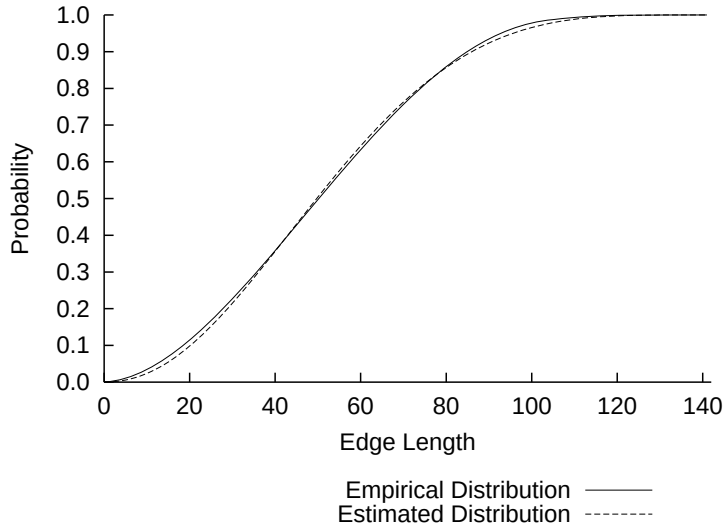


Figure 2: The estimated and empirical distributions of edge lengths

We also obtained the distribution of the length of an edge between two random vertices when the distances were measured according to the L_1 or Manhattan norm, and when the distances were measured according to the L_∞ norm. A comparison of these distributions with the distribution shown in Figure 1 is shown in Figure 3. The shapes of all three distributions are seen to be similar, although the spread of the distributions and their coefficient of variations are different. This evidence is in favor of our conjecture in the introductory section that the distance measure used does not affect the basic nature of the results that we obtain in this paper.

3 Tour Lengths

We performed two types of experiments with respect to tour lengths on randomly generated ESTSPs. The first type of experiments dealt with the distribution of the lengths of optimal tours to randomly generated ESTSPs. The second type of experiments looked into the whole set of tours for each instance of a randomly generated

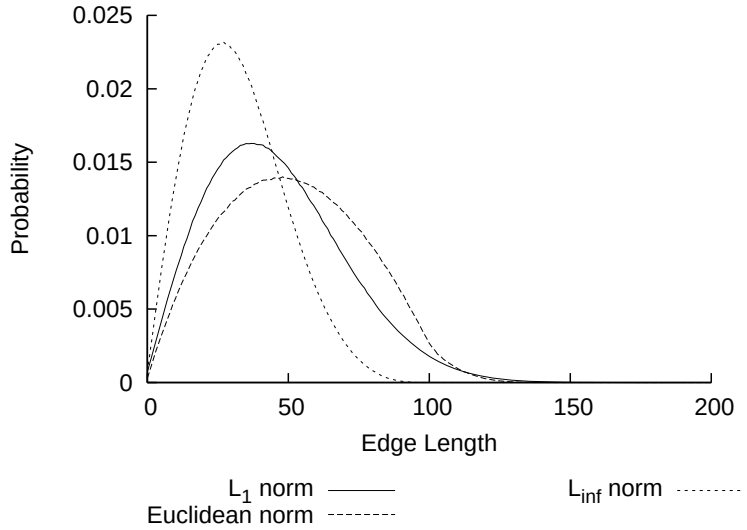


Figure 3: A comparison of the empirical distributions of edge lengths under different norms

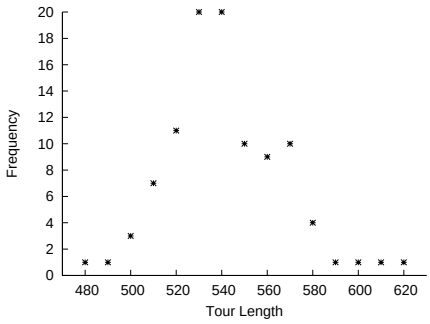
ESTSP, and examined the properties of the distribution of the lengths of the tours in this set. Recall from Section 2 that in our experiments, ESTSP instances with n vertices are generated by randomly scattering n vertices on a square of side 100 units.

3.1 Distribution of tour lengths of optimal tours

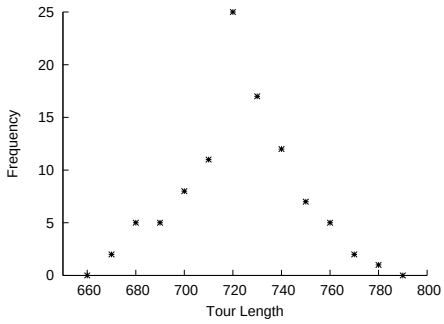
For our experiments regarding the distribution of the lengths of optimal tours to randomly generated ESTSP instances, we generated a hundred instances each of random ESTSPs of sizes 50, 100, 150, 200, and 250. We solved each of the instances to optimality using the CONCORDE TSP solver, implemented by Hans Mittelmann, and made available through the NEOS Server for Optimization [4]. For presentation purposes, the lengths of the optimal tours to the instances were grouped into classes of width 10 units. The results of the grouping are shown in Figure 4. Note that even when the problem sizes are relatively large, the distributions of the lengths of optimal tours are not smooth enough to reasonably fit known probability distributions. The shapes of the distributions also vary widely with changing problem sizes. On the other hand, the distributions have a reasonably wide spread. This means that, although the distribution of the lengths of edges in similar sized problems are similar, their optimal solutiosn could have widely different tour lengths. This makes randomly generated ESTSP instances a good test bed for comparing the performance of algorithms.

3.2 Distribution of tour lengths of all tours

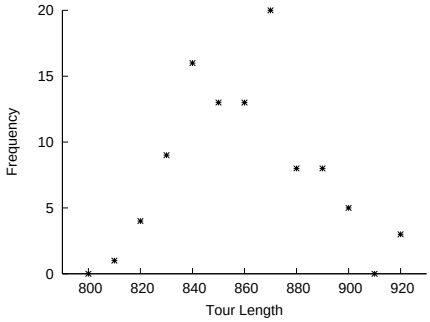
An ESTSP defined on n vertices admits $(n-1)!/2$ tours as solutions. In this section we examine the distribution of the lengths of these $(n-1)!/2$ tours. For small



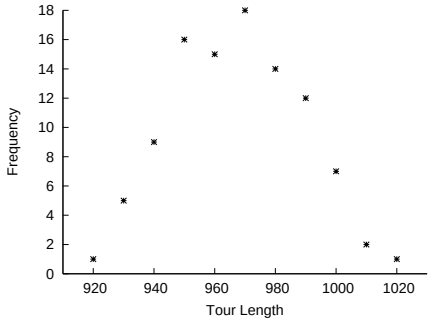
Instances with 50 vertices



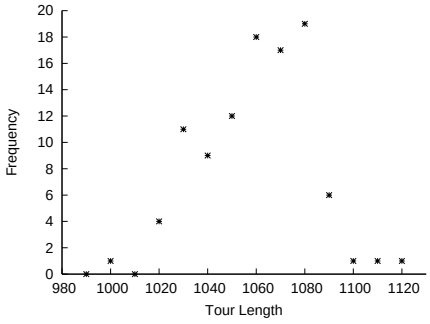
Instances with 100 vertices



Instances with 150 vertices



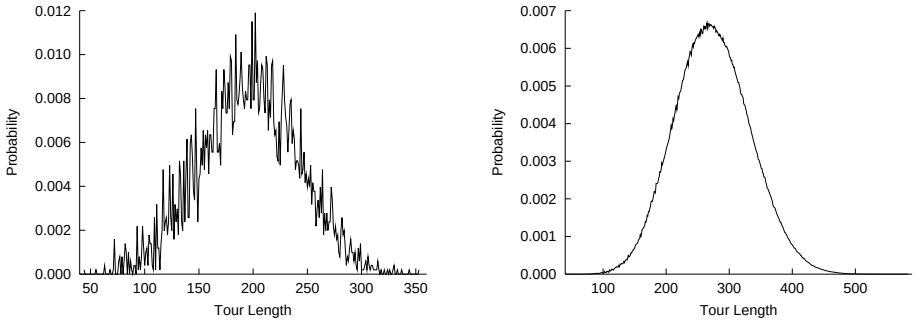
Instances with 200 vertices



Instances with 250 vertices

Figure 4: Frequency distribution of optimal tours to large random ESTSPs

values of n, it is practical to generate all tours, obtain their lengths, and construct the probability distribution of these lengths. We constructed the probability distributions for tour lengths of ESTSPs of sizes 8 through 12. For each size, we generated one hundred instances randomly. We observed that the distribution of tour lengths became smoother as the problem sizes increased. As an example, in Figure 5 we show the distribution of tour lengths for an instance of size 8 and an instance of size 12. The distribution of tour lengths for the instance with 8 vertices is jagged, while the distribution of tour lengths for the instance with 12 vertices is quite smooth. We expect that as the problem sizes increase, the distributions of tour lengths will become smoother.



Instances with 8 vertices Instances with 12 vertices

Figure 5: Distribution of tour lengths for small sized random ESTSPs

For each distribution of tour lengths that we generated, we computed the mean, the standard deviation, and the skewness values. These data are summarized in Table 2. In this table, for each problem size, we provide the mean, standard deviation, minimum and maximum values from the data on all the one hundred instances that we have of that size. Each column in the table refers to a parameter for the individual distributions, while we summarize the characteristics of each of the parameters in the rows of the table. Notice that the average of the lengths of tours is an approximately linear function of the problem size. Also note that the standard deviations in all the cases are quite large, compared to the means. Further, note that the skewness values that we obtain for all problem sizes are small, but not zero, so that the distributions of tour lengths are not symmetric for these problem sizes.

For large ESTSP instances, it is not practical to generate all tours and examine the parameters of their distributions. For these ESTSPs, we therefore use a sampling based procedure to obtain estimates of the parameters of the distributions of the tour lengths. We work with ESTSPs of sizes 50, 100, 150 and 200. One hundred instances of each of these problem sizes are generated. These are identical to the instances that were generated to test the distributions of lengths of optimal tours in Section 3.1. For each of these instances, we generated ten million random tours and noted the distribution of their lengths. This distribution is assumed to be an estimate of the distribution of all tour lengths for the instance. The distributions that we obtained were all found to be similar to the distribution in Figure 6, which is the distribution

Table 2: Parameters of the distribution of tour lengths for small random ESTSPs

		Distribution Parameter		
		Mean	Std. dev.	Skew
8 vertices	Average	417.32	49.66	-0.21
	Std. dev.	62.61	11.46	0.14
	Minimum	250.57	20.29	-0.68
	Maximum	558.57	78.00	0.04
9 vertices	Average	466.93	54.26	-0.21
	Std. dev.	59.20	10.30	0.12
	Minimum	290.75	23.97	-0.60
	Maximum	599.00	75.18	-0.01
10 vertices	Average	517.34	58.80	-0.22
	Std. dev.	61.71	10.41	0.12
	Minimum	650.67	79.18	-0.04
	Maximum	358.89	32.03	-0.67
11 vertices	Average	569.15	62.61	-0.21
	Std. dev.	63.61	10.01	0.10
	Minimum	420.00	32.34	-0.62
	Maximum	708.80	82.30	-0.06
12 vertices	Average	622.65	66.53	-0.20
	Std. dev.	65.61	10.06	0.09
	Minimum	484.73	44.05	-0.65
	Maximum	754.55	89.01	-0.07

of the lengths of ten million tours for an ESTSP instance of size 200. Notice that the distribution is quite smooth and unimodal. We modeled it as a GB distribution, and the maximum difference between the empirical and estimated distribution functions was 2.1×10^{-6}. On a diagram, the two distribution functions were indistinguishable.

The distributions of the sample tours from each of the instances allow us to estimate the parameters of the distributions of all the tour lengths of these instances. By our assumption, the average of the sample of tours that we generate is an unbiased estimate of the average of the tour lengths of all tours. We obtained the sample standard deviation of the tours in our sample, and used it as an unbiased estimate of the standard deviation of the tour lengths in the distribution of all tours. Given the lengths x_i of the tours in the sample, their mean$\bar{x}$, and n = 10,000,000, we considered the following as an unbiased estimate of the skewness of the distribution of the lengths of all tours:

$$\frac{n\sqrt{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-2)\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}.$$

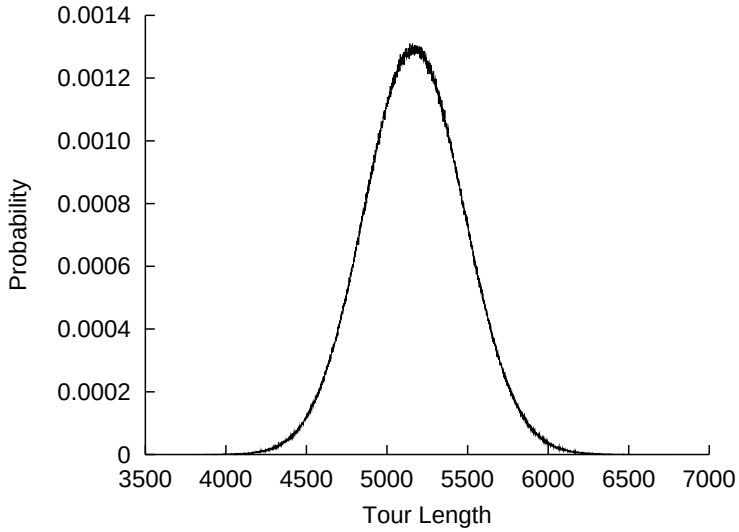


Figure 6: Distribution of tour lengths of 10 million tours in a random ESTSP with 200 vertices

The properties of the estimated parameters for the distributions of the lengths of all tours is shown in Table 3. The explanation of the figures in this table is identical to those of the figures in Table 2. Notice that here too, the averages of the lengths of all tours are approximately linearly related with problem size. The values of the standard deviations here too are quite large. This implies that randomly generated ESTSP instances are quite different from each other. The skewness values are smaller than the values presented in Table 2, thereby showing that the distribution of tour lengths for larger ESTSPs are more symmetric than those of smaller ESTSPs. We conjecture that for still larger ESTSPs, the distribution of tour lengths would be very nearly symmetric. The symmetric distributions of the tour lengths for random ESTSPs suggest that for randomly generated ESTSP instances, the fraction of solutions that are near optimal solutions are reasonably small, and good bounding mechanisms can fathom a large proportion of the solutions quite early in the branch and bound execution. Thus, although they are varied in nature, they are unlikely to be among the more difficult problems for branch and bound algorithms.

4 Summary and Discussion

In this paper, we have empirically examined some properties of randomly generated Euclidean symmetric traveling salesman problem (ESTSP) instances, which are often used to test the performance of heuristics and exact algorithms for the symmetric traveling salesman problem. Our experiments were on random ESTSP instances generated by distributing vertices randomly on a square of side 100 units. They concerned three aspects of ESTSP instances, the lengths of the edges in such problems,

Table 3: Estimates of the parameters of the distribution of tour lengths for large random ESTSPs

		Distribution Parameter		
		Mean	Std. dev.	Skew
50 vertices	Average	1308.76	155.22	0.06
	Std. dev.	62.68	8.28	0.01
	Minimum	1130.62	131.74	0.04
	Maximum	1477.30	181.18	0.07
100 vertices	Average	2603.67	220.05	0.04
	Std. dev.	86.48	8.08	0.00
	Minimum	2395.64	198.07	0.03
	Maximum	2766.96	236.57	0.05
150 vertices	Average	3883.48	268.41	0.03
	Std. dev.	111.50	8.15	0.00
	Minimum	3546.03	248.43	0.03
	Maximum	4150.62	285.60	0.04
200 vertices	Average	5201.81	311.99	0.03
	Std. dev.	117.10	7.55	0.00
	Minimum	4933.15	293.99	0.03
	Maximum	5411.57	324.94	0.03

the distribution of the lengths of optimal tours in such problem instances, and the distributions of lengths of all tours in any given instance.

In Section 2 we found out the distribution of edge lengths for a random ESTSP. This distribution is difficult to compute theoretically, since it is not just the Euclidean distance between the vertices, but the greatest integer below the (floating point) value of the Euclidean distances. Hence we constructed this distribution empirically. Our construction was based on ten million randomly generated ESTSP instances. We observed that the edge lengths followed a unimodal positively skewed distribution (see Figure 1) and could be modeled using a Generalized Beta (GB) distribution (see Figure 2). Since the ESTSP instances were generated randomly, some pairs of vertices in them could be close to each other, and a number of edges could have zero lengths. Hence the estimated distribution of the edge lengths needed to have the left side of the support at -1 instead of 0, even though negative edge lengths do not have any physical significance. We noted that all the edges in a randomly generated ESTSP came from identical distributions, and that the edge lengths did not vary with varying problem sizes.

In Section 3, we examined the distributions of the lengths of optimal solutions, as well as those of the set of all solutions to random ESTSPs. We saw in Section 3.1 that the optimal tours in randomly generated ESTSPs are distributed over a significant range, and that the distributions did not stabilize as problem sizes increased (see

Figure 4). In Section 3.2 we looked at the distribution of the lengths of all tours in randomly generated ESTSP instances. For small instances (of size 12 and less), it was feasible to generate all the tours, compute their lengths, and obtain the distributions of the tour lengths (see Figure 5). For larger instances, generating all tours was impractical, and hence we generated the distribution based on random samples of ten million tours. (We expect the distribution of all tours in an instance to be similar to the distribution of the lengths of tours generated in this manner.) The distributions of the lengths of tours in our samples for each of the larger instances were found to be relatively smooth and unimodal (see Figure 6), and were seen to be amenable to modeling using GB distributions. The properties of the actual distributions of the tour lengths for smaller instances were presented in Table 2, and those of the estimated distributions of tour lengths for larger instances were presented in Table 3. In both cases, we saw that the average of the tour lengths varied approximately linearly with changing problem sizes, and that the spread of the distributions were quite significant. This showed that random ESTSPs are good candidates for testing algorithms and heuristics for ESTSPs. The skewness of the distributions were seen to decrease with increasing problem sizes, especially for the larger instances, but the distributions were not symmetric even for random ESTSPs of size 200. We conjecture that as the problem size increases, the skewness would approach zero. The fact that the skewness values are not high implies that there are a relatively small number of near-optimal solutions. So randomly generated ESTSP instances are not likely to be very hard for branch and bound algorithms for STSPs.

The fact that the distribution of lengths of tours of ESTSP instances are amenable to modeling using GB distributions is a particularly interesting result for future research. If the distribution of lengths of tours can indeed be modeled using GB distributions, then, given a relatively small set of randomly generated tours, it should be possible to estimate the parameters of the distribution that they come from. The left end of the support of this distribution is nothing but the length of an optimal tour to the ESTSP instance. Hence this could be a novel method of solving the evaluation version of the ESTSP. It may also be worthwhile to see if such a method could be used on other hard optimization problems.

Acknowledgement: The second author was supported in this work by IIMA R&P Grant 1105810.

References

[1] G. Gutin and A.P. Punnen (eds.), The Traveling Salesman Problem and Its Variations, Kluwer Academic Publishers, Dordrecht, 2002.

[2] N.L. Johnson, S. Kotz, and N. Balakrishnan, Continuous Univariate Distributions, Volume 2, Wiley Series in Probability and Mathematical Statistics, 210–275, 1995.

[3] R.M. Karp, Reducibility among combinatorial problems, in R.E. Miller and J.W. Thatcher (eds.), Complexity of Computer Computations, Plenum Press, New York, 85-103, 1972.

[4] J. Czyzyk, M. Mesnier, and J. Moré, The Network-Enabled Optimization System (NEOS) Server for Optimization, available at `http://neos.mcs.anl.gov/`

[5] G. Reinelt, TSPLIB — A Traveling Salesman Problem Library, ORSA Journal of Computing 3, 376-384, 1991.