Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 186 (2008) 111–119
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Discrete Optimization

On estimating the distribution of optimal traveling salesman tour lengths using heuristics

Vikrant Vig, Udatta S. Palekar *

Department of Business Administration, University of Illinois at Urbana-Champaign, 1206 S. Sixth Street, Champaign, IL 61820, United States

Received 13 January 2003; accepted 15 December 2006
Available online 1 March 2007

Abstract

The traveling salesman problem is an important combinatorial optimization problem due to its significance in academic research and its real world applications. The problem has been extensively studied and much is known about its polyhedral structure and algorithms for exact and heuristic solutions. While most work is concentrated on solving the deterministic version of the problem, there also has been some research on the stochastic TSP. Research on the stochastic TSP has concentrated on asymptotic properties and estimation of the TSP-constant. Not much is, however, known about the probability distribution of the optimal tour length. In this paper, we present some empirical results based on Monte Carlo simulations for the symmetric Euclidean and Rectilinear TSPs. We derive regression equations for predicting the first four moments of the distribution of estimated TSP tour lengths using heuristics. We then show that a Beta distribution gives excellent fits for small to moderate sized TSP problems. We derive regression equations for predicting the parameters of the Beta distribution. Finally we predict the TSP constant using two alternative approaches.

Keywords: Travelling Salesman Problem; Tour length; TSP constant

1. Introduction

The traveling salesman problem (TSP) is the problem of identifying the minimum cost tour of n pre-specified cities such that the salesman, starting from an origin city, visits each of the cities exactly once and then returns to the origin. TSP is an NP-Hard problem and has been extensively studied.

Most researchers have focused their efforts on obtaining optimal solutions to the TSP. In much of this research, the problem parameters are deterministically known. Comparatively little research has been done on stochastic versions of the problem. One area that has received very little attention is the probability distribution of optimal TSP tour lengths. Even where such work exists, the attention has been on the asymptotic case where the number of cities is infinitely large. Some efforts have also been expended in determining the value of the TSP constant.

Knowledge of the distribution of optimal tour lengths is important for a variety of reasons. We list below some cases in which such knowledge can be useful:

* Corresponding author. Fax: +1 217 244 6534.
E-mail address: palekar@uiuc.edu (U.S. Palekar).

0377-2217/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.ejor.2006.12.066

1. Service time estimation for traveling repairperson models: In such models, a traveling repairperson visits several customers. Depending on the application, the traveling repairperson may visit the cities on a first come first serve or, more commonly, will visit the customers according to a TSP tour to minimize travel time. Examples of such systems include equipment repairpersons, pizza delivery, WebVan like grocery delivery, social service providers etc. To determine service time distribution in such systems, it is important to know the amount of travel time that might be required.
2. Simulation of traveling repairperson systems: During simulation of traveling repair person systems, it is often necessary to solve TSP problems to know the amount of time that a repair person spends traveling. Knowledge of the distribution of the optimal tour length will significantly shorten the run times for such simulations.
3. Simulation and analysis of multi-modal Public Transportation systems: Many urban areas are experimenting with the idea of providing public transportation using a main arterial train/tram system which would be fed by a dial-a-ride taxi or van system. In this case the taxi tours are often modeled as TSP tours and knowledge of the tour lengths would greatly simplify the analysis of such systems.

In most of these instances the number of stops in the tour is a small. It is typical for traveling repairpersons to make between 15 and 20 stops. However, depending on the application the number of such tours may be as large as a few hundred.

For many of these applications, the city areas covered tend to be of irregular geometric shapes. However, service districts for traveling repairmen are often imposed on the cities and such zones are usually laid in the form of grids making it possible to consider square and or rectangular service areas. It may also be possible to break up the service area into square sections, so that the resulting TSP tour can be considered as the aggregation via a process such as sub-tour patching of a number of independent sub-tours. In this later case, the resulting tour is clearly a heuristic solution to the problem. This approach of decomposition into smaller square areas would also allow the modeling of non-homogenous regions in which the demand is not uniformly distributed over the region of interest. In this paper we concentrate on randomly generated homogenous unit square areas.

The primary purpose of the paper is to provide distributional results for small to medium traveling salesman problems drawn from a unit square. We give empirically fitted distributions for the optimal tour lengths estimated using heuristics and regression equations for the first four moments of the distribution. We also use our empirical studies to estimate the TSP constant. We consider both Euclidean and Rectilinear TSP problems.

2. Literature review

The Traveling Salesman problem has been extensively studied in the literature. An excellent survey of this research is the book by Gutin and Punnen (2002), which covers research on a variety of different aspects and variations of the problem. The problem of obtaining the expected length of the shortest Hamiltonian cycle through n random points in a two dimensional region has been considered by several researchers. Mahalanobis (1940) estimated the expected length of the shortest path through n points in any region to be $(\sqrt{n} - 1/\sqrt{n})$. Marks (1948) obtained a lower bound for the expected length as $\sqrt{A/2}(\sqrt{n} - 1/\sqrt{n})$, A being the area of the region from which the points are drawn. Ghosh (1951) gives an empirical result of $1.266\ \sqrt{A_n}$ for the expected length of the shortest path. Few (1955) gave an upper bound on the shortest distance connecting n points in a 1×1 square of $(\sqrt{2n} + 1.75)$. A bound of $(\sqrt{2n} + o(\sqrt{n}))$ was rediscovered in 1983 by Supowit et al. Karloff (1989) improved this bound to $0.984(\sqrt{2n} + 11)$.

The seminal paper in this area is the one by Beardwood et al. (1959). In this paper they present some asymptotic results on the expected value of the shortest distance among n points using tours that they constructed by hand for a 202 and 400 city instance. They estimated the expected length of the optimal tour found for a random uniform distribution of n points over a rectangular area R given by $L_e(n, R) = K\sqrt{nR}$. K is estimated to be 0.75. Here K is defined as $\beta_{TSP}\sqrt{2}$ where β_{TSP} is the TSP constant defined by $\beta_{TSP}(t) = \lim_{n\to\infty} L(n, t)/(n^{(t-1)/t} \cdot \sqrt{t})$ where n is number of points and t is the dimension. Since then many researches using varied methodologies, have given different estimates of the TSP constant. Bonomi and Lutton (1984) quote the value of $K = 0.749$ for large n. Based on computer experiments with more sophisticated heuristics, Stein (1997) estimated the constant to be 0.765. In recent years many researchers have realized that these figures are over-estimates. Ong and Huang (1989) reported that a version of 3-Opt yielded normalized tour lengths converging to 0.74. Krauth and Mizard (1989) use the cavity method to predict the TSP constant to be 0.7257. Fiechter (1994) using a "parallel tabu search" algorithm observed normalized tour lengths converging to 0.721. Lee and Choi (1994) use a "multicannonical annealing" algorithm to obtain normalized tour lengths converging to 0.721. Norman and Moscato (1995) in their research imply a value of 0.7148 based on the "MNPeano" constant which is related to a fractal space-filling curve. The most current estimate ($K \approx 0.7124$), based on computer simulation and $n = 1000$, is due to Johnson et al. (1996). Percus and Martin (1996) using an independence assumption give 0.7120 for the TSP constant which matches the value obtained by Johnson et al. (1996).

It is important to note that our work differs from the work on statistical inference to obtain point and interval estimates for the optimal solution (e.g. Golden and Alt, 1979; Los and Lardinois, 1982). In that stream of research, the idea is to find a confidence interval for the optimal solution of a given instance of the TSP. The approach there is to sample the $n!$ tours associated with the given instance to estimate the optimal solution. Our interest is in finding the distribution of the optimal solution for uniformly generated problems.

The paper is organized as follows. In Section 3 we discuss the methodology used in obtaining the empirical results for TSP tours. In Section 4 we present the results and some of the insights that can be drawn from them. In Section 4.1 we discuss the empirical model for the different moments and validate it. In Section 4.2 we report the different parameters of the hypothesized distribution and the goodness-of-fit results. In Section 4.3 we calculate the value of the TSP constant using two approaches. Finally in Section 5 we present our conclusions.

3. Methodology

Previous researchers have used different methods to come up with their estimate of the TSP constant. Some have used the Monte Carlo techniques while some others have used statistical mechanical arguments. In this research, we generate large numbers of random instances of the TSP problem. The coordinates for the cities are drawn randomly from a 1×1 square. We implemented the Marse and Roberts (1985) random number generator. This generator generates 100,000 random numbers for each given stream. For different values of n (number of cities) several random instances of the problem are generated. The length of both the Euclidean tour as well as the Rectilinear tour is approximated using Helsgaun's implementation of the Lin–Kernighan heuristic algorithm (Helsgaun, 2000). Johnson and McGeoch (2002) show that the solutions from this implementation are typically very close to optimal. For problems with 1000 cities they found the solution to be on average 0.9% from the optimal.

For the purpose of analysis we broke the TSP runs into 9 batches of 4000 TSP's, which totals 36,000 TSP's for each n. The maximum number of points that we considered were $n = 2000$ for which we solved only 1034 TSP's as we observed a marked reduction in the variance as n increased. This is in line with the observation of Johnson et al. (1996). Of the 9 batches we used one batch to estimate parameters for the hypothesized distribution tests and the other 8 batches for testing the goodness of fit and regression Table 1.

We then divided the data into groups consisting of 500 points for each value of n and fit a regression line through them. The standard basic steps for distribution fitting were used. This gave us an estimate of the parameters of the line. Repeating this with the entire set of data gave us $E(\hat{\beta})$, which due to the unbiased property of the optimized least square (OLS) estimator is an estimate of β. A similar approach was used for all the moment equations – only the hypothesis was different.

For the distribution we hypothesized a Beta distribution based on our preliminary study and the use of Stat::fit software. We then estimated the parameters of the Beta distribution using the mean and the variance values obtained from our Monte

Table 1
Number of instances for each problem size

Number of points	Number of TSP instances
11–50	36,000
51–99 (every alternate)	36,000
100–150 (every 5th)	36,000
160–200 (every 10th)	36,000
1000	1768
1500	1344
2000	1034

Carlo simulation of the problem. A Beta distribution in the range $(0,1)$ is represented by $B(\alpha_1, \alpha_2)$ where α_1 and α_2 are the shape parameters of the distribution with α_1 and α_2 both greater than zero. The scaled mean and scaled variance of the Beta distribution is given by Eqs. (1) and (2), where μ and σ^2 are the mean and the variance, respectively.

$$\text{Mean: } \mu = \alpha_1/(\alpha_1 + \alpha_2), \tag{1}$$

$$\text{Variance: } \sigma^2 = \alpha_1\alpha_2/(\alpha_1 + \alpha_2)^2(1 + \alpha_1 + \alpha_2). \tag{2}$$

It is important to note that the parameters α_1 and α_2 of the Beta distribution are obtained using the scaled mean and the scaled variance. To scale the distances in the $(0,1)$ interval, knowledge of both the lower as well as the upper bound is required. The lower bound in all the cases has been assumed to be zero. The value for the upper bound required some more analysis. While Few (1955) gives an upper bound, this upper bound is clearly loose and using this bound will lead to a poor fit. For larger values of n (100 or more) our initial analysis had indicated that the Normal distribution was a reasonable fit. If the distribution had been a Normal distribution we would have been justified in using a bound of $\mu + 3\sigma$. However, for smaller values the Normal does not give a good fit. We interpolated between $\mu + 3\sigma$ and Few's bound to find a value for the upper bound that gave the best fit for our observed data. The upper bound was thus given by $b_{\text{few}} - k(b_{\text{few}} - \mu - 3\sigma)$. The selection of a value for k is discussed in the results. Solving Eqs. (1) and (2) gives estimates for α_1 and α_2.

The distributions were then tested using standard tests namely the chi square test, Kolmogorov–Smirnov (K–S) test and the Andersen Darling test to check for the validity of the hypothesized distribution. Chi-square tests are the oldest goodness-of-fit tests, and may be thought as a more formal comparison of a histogram or a line graph with the fitted density or mass function. The K–S tests, on the other hand, compare an empirical distribution function with the distribution function of the hypothesized distribution. This test generally seems to be more powerful than the chi-square test against many alternative distributions. One possible draw back of the K–S tests is that they give the same weight to the difference $|F_n(x) - \widehat{F}(x)|$ for every value of x, where $F_n(x)$ and $\widehat{F}(x)$ are the empirically observed and theoretical distributions, respectively. The Anderson–Darling (A–D) test is designed to detect the discrepancies in the tails and has higher power than the K–S test against many alternative distributions.

A regression was also fitted with $\hat{\alpha}_1$ and $\hat{\alpha}_2$ as the dependant variables and n as the independent variable. The estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are used to predict the value of the TSP constant. The results of the analysis are discussed in Section 4.

4. Results

In this section the results of the analysis are presented and discussed. This section is further subdivided into Sections 4.1, 4.2 and 4.3. In Section 4.1 we present the empirical results of the four moment equations based on regression. In Section 4.2 the results of the distribution tests as well as the regression equations for the distribution parameters are presented and in Section 4.3 we calculate the value of the TSP constant using different methodologies.

4.1. Moment equations

We derived regression equations for the first four moments of the distributions as a function of n. The guiding intuition behind the hypothesized function was the upper bound equation derived by Few (1955). The preliminary hypothesis for the different moments is shown in Table 2. In this table, d is the optimal tour distance obtained from the heuristic. The coefficient terms that were statistically significant were retained and the rest were discarded. The final hypothesis is shown in Table 2. Also the hypothesis H_{02} is presented, which is the hypothesis having no intercept term. The validation was done using the standard procedures involving the behavior of residuals.

Using the regression equations generated from different samples we produced an estimate of the

Table 2
Hypothesized regression equations for moments

	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
H_{01}: (initial)	$\beta_0 + \beta_1\sqrt{n}$	$\beta_0 + \beta_1\sqrt{n} + \beta_2 n$	$\beta_0 + \beta_1\sqrt{n} + \beta_2 n + \beta_3 n^{3/2}$	$\beta_0 + \beta_1\sqrt{n} + \beta_2 n + \beta_3 n^{3/2} + \beta_4 n^2$
H_{01}: (final)	$\beta_0 + \beta_1\sqrt{n}$	$\beta_0 + \beta_1 n$	$\beta_0 + \beta_1 n^{3/2}$	$\beta_0 + \beta_1 n^2$
H_{02}:	$\beta_1\sqrt{n}$	$\beta_1 n$	$\beta_1 n^{3/2}$	$\beta_1 n^2$

parameters of the line. Averaging this with the entire set of data gives $E(\hat{\beta})$, which due to the unbiased property of the optimized least square estimator is an estimate of β. A similar approach was used for all the moment equations. Finally, t-tests and F-tests were used to verify the hypothesis. Results were also obtained with the additional constraint of setting the intercept equal to zero. These results are important as they help us estimate the value of the TSP constant. The regression equations were also obtained for Rectilinear distances as well as for the Held–Karp (HK) lower bound.

The results of the analysis are presented in the following order. Tables 3 and 4 show the results for the TSP tour lengths based on runs for both the Euclidean as well as the Rectilinear tour lengths, respectively. Regression results based on the H_{01} and H_{02} are shown in Tables 3 and 4. The 95% confidence interval for each coefficient is also reported.

Table 3(a)
Regression for Euclidean TSP tour length

TSP-Euclidean	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0.66268	4.179435	28.33828	186.2417
β_1:	0.710301	0.55803	0.432637	0.344446
β_0-Lower 95%	0.661689	4.172579	28.27833	185.6856
β_0-Upper 95%	0.66	4.19	28.4	186.80
β_1-Lower 95%	0.71017	0.55795	0.43257	0.34439
β_1-Upper 95%	0.710428	0.558115	0.432707	0.344506

Table 3(b)
Regression for Rectilinear TSP tour length

TSP-Rectilinear	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0.764621	5.856592	61.26064	438.7929
β_1:	0.892139	0.876018	0.848252	0.813122
β_0-Lower 95%	0.762918	5.841595	60.85072	436.8275
β_0-Upper 95%	0.766324	5.871589	61.67057	440.7583
β_1-Lower 95%	0.891373	0.87529	0.847524	0.812382
β_1-Upper 95%	0.892904	0.876745	0.84898	0.813862

Table 4(a)
Regression for Euclidean TSP tour length – zero intercept

TSP-Euclidean	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0	0	0	0
β_1:	0.788236	0.6007	0.455424	0.344004
β_1-Lower 95%	0.78818	0.60064	0.45537	0.34395
β_1-Upper 95%	0.788289	0.60076	0.455481	0.344054

Table 4(b)
Regression for Rectilinear TSP tour length – zero intercept

TSP-Rectilinear	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{02}	H_{02}	H_{02}	H_{02}
β_0:	0	0	0	0
β_1:	0.982307	0.938436	0.890462	0.840627
β_1-Lower 95%	0.98182	0.937841	0.883821	0.839948
β_1-Upper 95%	0.982794	0.93903	0.891104	0.841305

The R-square value for all the regressions is very high with values of 99% for $E(d)$, 98.5% for $E(d^2)$, 98% for $E(d^3)$ and around 97% for $E(d^4)$. The F statistic value and the t-statistic values confirm the validity of the regression equation as well as the coefficients of the various terms.

The results show that the confidence intervals for the slope co-efficient are small for all cases. On constraining the constant term of the regression to be zero, we see a marked increase in the value of the slope. We expect this to fall gradually with increase in the value of n and asymptotically approach the TSP constant.

We also report the results obtained by regressing the HK lower bound. The results clearly follow the same pattern. It can be seen from the lower bound results that the heuristic on average gives good solutions. This validates the results of the heuristic. The average lower bound results are presented in Table 5.

4.2. Distribution fitting

Our preliminary study of the distribution with the help of Stat::Fit® yielded four competing distributions namely Beta, Weibull, Normal and Lognormal. In almost all cases the Beta distribution was ranked as the best fitting distribution. The Lognormal seemed to be a good fit for smaller values of n while the Normal distribution was a good fit for larger values. Therefore we hypothesized a Beta

Table 5(a)
Regression for Euclidean TSP lower bound

TSP-Euclidean	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0.69026	4.28946	28.926	209.0146
β_1:	0.701594	0.547869	0.420944	0.320253
β_0-Lower 95%	0.689286	4.282677	28.86671	208.432
β_0-Upper 95%	0.69	4.3	28.99	209.60
β_1-Lower 95%	0.70147	0.54778	0.42088	0.32020
β_1-Upper 95%	0.701716	0.547953	0.42101	0.320307

Table 5(b)
Regression for Rectilinear TSP lower bound

TSP-Rectilinear	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0.796068	5.838576	49.63161	436.6565
β_1:	0.882094	0.861779	0.827842	0.789298
β_0-Lower 95%	0.794341	5.823986	49.47149	434.7043
β_0-Upper 95%	0.797795	5.853166	49.79173	438.6087
β_1-Lower 95%	0.881878	0.861054	0.827122	0.789098
β_1-Upper 95%	0.882309	0.862504	0.828562	0.789498

Table 5(c)
Regression for Euclidean TSP lower bound – zero intercept

TSP-Euclidean	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{01}	H_{01}	H_{01}	H_{01}
β_0:	0	0	0	0
β_1:	0.782162	0.591662	0.443509	0.331921
β_1-Lower 95%	0.78211	0.59160	0.44345	0.33187
β_1-Upper 95%	0.782215	0.591722	0.443564	0.331968

Table 5(d)
Regression for Rectilinear TSP lower bound – zero intercept

TSP-Rectilinear	$E(d)$	$E(d^2)$	$E(d^3)$	$E(d^4)$
Hypothesis	H_{02}	H_{02}	H_{02}	H_{02}
β_0:	0	0	0	0
β_1:	0.970203	0.927078	0.869531	0.81631
β_1-Lower 95%	0.969702	0.926485	0.868889	0.816461
β_1-Upper 95%	0.970704	0.927671	0.870172	0.816802

Table 6(a)
Number of rejections Euclidean TSP

Euclidean				
k	K–S		A–D	
	Confidence level		Confidence level	
	0.1	0.05	0.1	0.05
.400	11.375	5.625	11.5	5.5
.425	10.125	5.375	10.625	5.125
.450	9.375	5.375	9.625	4.125
.475	8.75	4.75	9.125	3.5
.500	8.75	4.625	8.25	3.375
.525	**8.25**	**4.625**	**7.25**	**3.125**
.550	8.625	4.75	7	3
.575	7.75	4.25	6.375	2.875
.600	7.875	4.125	6.125	2.75

Table 6(b)
Number of rejections Rectilinear TSP

Rectilinear				
k	K–S		A–D	
	Confidence level		Confidence level	
	0.1	0.05	0.1	0.05
.400	6.75	3	6.5	2.75
.425	6.75	2.75	5.75	2.5
.450	7	3	5.75	2
.475	7.25	3.25	5.5	2
.500	7.75	3.5	5.25	2
.525	**7.25**	**4**	**5**	**2**
.550	7.5	3.75	5.25	2
.575	8	4	5.5	2.5
.600	8.5	4	5.25	2

distribution for the TSP distances. As discussed earlier we used scaled values to fit Beta distributions in the range (0,1). We systematically searched the interval between $\mu + 3\sigma$ and Few's bound for the value that gave the best fit. Partial results for this process are given in Table 5. Surprisingly, we found that the same value of 0.525 gave the best results for both the Rectilinear and the Euclidean cases.

The results of the K–S tests and the A–D tests are reported in Table 6. The average number of errors for each set of data is reported. The table lists the results for both alpha values of 0.05 and 0.1.

On further breaking down the number of rejections into different runs we obtain the following result (Figs. 1a and 1b).

Figs. 1a and 1b report the number of rejections for the K–S and A–D test for different values of alpha on the ordinate axis and the sample number on the abscissa. The number of different points that were considered was 81 for the Euclidean case and 80 for the Rectilinear case. It can be seen from Fig. 1a that E (4) and E (5) had considerably high rejections as compared to the others. However, the addition of more sample points led to a considerable reduction in the number of rejections. Fig. 1b on the other hand is more consistent. It can be seen that more runs were done for the Euclidean than Rectilinear owing to the time intensive nature of each TSP run (Table 7).

Once the adjustment factor has been defined, the next step is to define the various parameters of the Beta distribution. Our aim is to come up with a function for the shape parameters in terms of n. To do this we hypothesize the shape parameters to be linear functions of n. We test the hypothesis by performing a linear regression on the different values of α_1 and α_2. The results are shown in Table 6.

From the table it can be seen that the R-square values are very high and the null hypothesis is clearly accepted. We also report the confidence

Table 7
Regression equations for the parameters of the Beta distribution

Shape parameters	Euclidean		Rectilinear	
	α_1 (n)	α_2 (n)	α_1 (n)	α_2 (n)
H_0:	$\beta_0+\beta_1 n$	$\beta_0+\beta_1 n$	$\beta_0+\beta_1 n$	$\beta_0+\beta_1 n$
β_0:	−27.47423	−4.266975	−7.48185	5.34572
β_1:	4.304877	2.222080	2.968166	0.95796
β_0-Lower 95%	−29.1875	−4.89928	−8.531407	4.819218
β_0-Upper 95%	−25.761	−3.63467	−6.432288	5.872224
β_1-Lower 95%	4.283617	2.214234	2.9547002	0.951209
β_1-Upper 95%	4.326139	2.229927	2.9816322	0.964720
$\beta_0(t_{\text{stat}})$	−31.9189	−13.4321	−14.1918	20.21
$\beta_1(t_{\text{stat}})$	403.0229	563.67	438.821	282.33
F	162427.5	317729.2	192564.2	79709.5
R square	0.9995	0.9995	0.9995	0.9995
Adjusted R square	0.9995	0.9995	0.9995	0.9995

intervals for each coefficient. Due to the sensitivity of the coefficients we have reported certain coefficients up to five decimal places. The equations show that there is a clear relationship between the shape parameters and n.

The regression equations give the parameters for a scaled Beta distribution in the range $[0,1]$. To recover the actual distribution we have to scale this distribution by calculating the upper bound. Recall the discussion in the previous section where we calculated the value of the upper bound as

$$Upperbound = b_{\text{Few}} - 0.525*(b_{\text{Few}} - \mu - 3\sigma),$$

where $b_{\text{Few}} = \sqrt{2n} + 1.75$ and μ and σ^2 can be estimated by $E(d)$ and $E(d^2) - (E(d))^2$ respectively.

The variance keeps decreasing as n increases and $E(d^2) - (E(d))^2$ occasionally gives a small negative value as n becomes very large. In such cases, one can simply assume that the variance is 0. This is consistent with the finding of Johnson et al. (1996). It is important to note this is due to the fact that we are forecasting these values based on regression and the regression fit is probably not very accurate for very large value of n.

Example:. Consider $n = 26$. The scaled Beta distribution parameters are given by

$$\alpha_1 = -27.47423 + 4.304877*26 = 84.45257$$

and

$$\alpha_2 = -4.266975 + 2.22208*26 = 53.50711.$$

Moreover,

$$E(d) = 0.66268 + 0.710301\sqrt{26} = 4.284519,$$

and

$$E(d^2) = 4.179435 + 0.55803*26 = 18.68822.$$

This gives an estimate for the variance equal to 0.3311. Few's bound for this cases is $b_{\text{Few}} = \sqrt{2\times 26} + 1.75 = 8.9611$. The upper bound for the Beta distribution is therefore equal to $Upperbound = b_{\text{Few}} - 0.525*(b_{\text{Few}} - \mu - 3\sigma) = 7.0274$.

To estimate the optimal length of a random Euclidean tour of 26 cities on a 1×1 square, we generate a Beta (84.45257, 53.50711) random variate and multiply it by 7.0274.

4.3. The TSP constant

In this section, we present two different methods for calculating the value of the TSP constant. In the first method we calculate the $(E(d)/\sqrt{n})$, where $E(d)$ is based on the results of the Monte Carlo simulation. The value for $n = 2000$ comes out to be 0.725341, which is expected to reduce further with increase in the value of n. A similar calculation

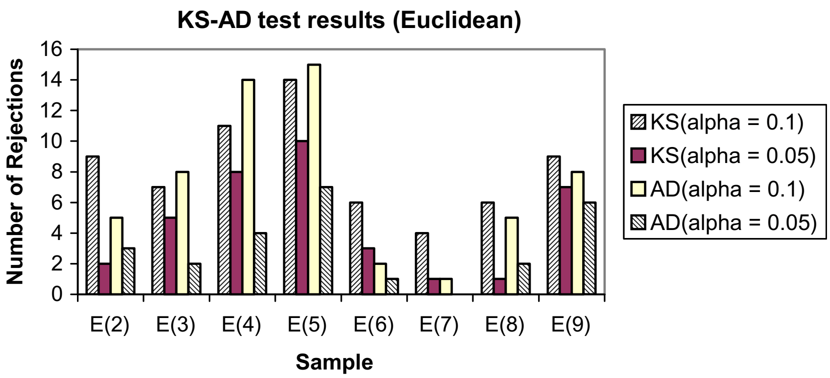

Fig. 1a. Distribution fitting results for Euclidean TSP.

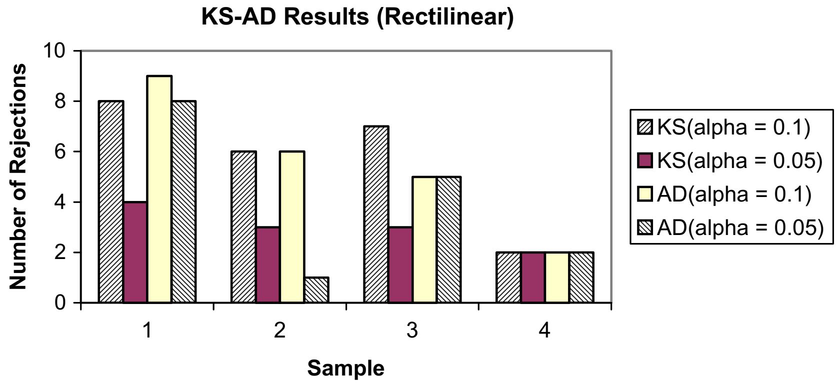

Fig. 1b. Distribution fitting results for Rectilinear TSP.

using the lower bound results gives the value of $(E(d)/\sqrt{n})$ to be 0.720192.

The actual values of the lower bound and the TSP tour lengths for $n = 2000$, based on the simulations results are 32.20798 and 32.43825, respectively. The corresponding values obtained using the regression equations are 32.42831 and 32.0665. Therefore we see that the regression equation predicts the actual data within 1% of accuracy for both the lower bound as well as the TSP tour lengths.

The second method to calculate the TSP constant involves the use of the regression equation for $E(d)$. For the Euclidean TSP, we have $\lim_{n\to\infty} E(d)/\sqrt{n} = \lim_{n\to\infty}(0.66268 + 0.710301 * \sqrt{n})/\sqrt{n} = 0.710301$. This is very close to the value of 0.7124 ± 0.002 estimated by Johnson et al. (1996). Indeed, the 95% confidence interval for our TSP constant is [0.71017, 0.710428] which overlaps with that obtained by Johnson et al. Our calculation is somewhat smaller mainly because the regression equations are biased towards giving a good fit for smaller values of n and the intercept of the regression equation plays a big role in this. As is apparent from Table 4(a) the value for β_1, which gives the TSP constant, is much larger if we use a zero-intercept. The corresponding constant for the lower bound is 0.7016. For the Rectilinear tour lengths the corresponding values are 0.8921 and 0.882, respectively. Figs. 1 and 2 plot the values of $(E(d)/\sqrt{n})$ for the Euclidean and Rectilinear cases, respectively Fig. 3.

It should be noted here that the difference between the lower bound values of the TSP tour lengths and the tour lengths as obtained by the heu-

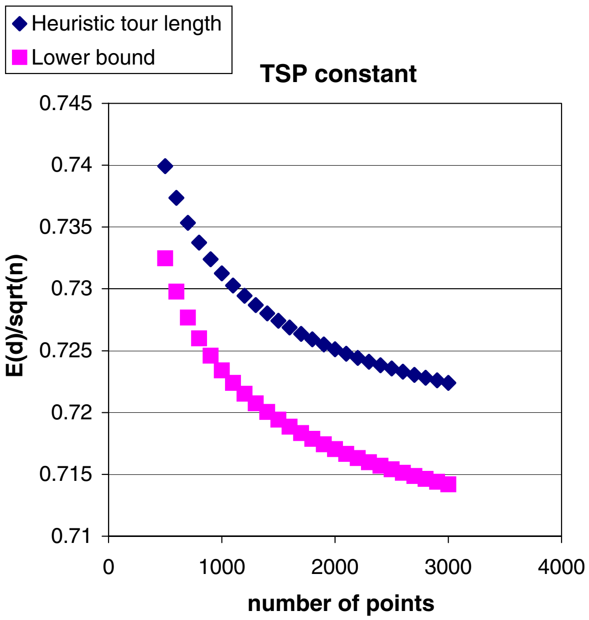

Fig. 2. TSP Constant for Euclidean TSP.

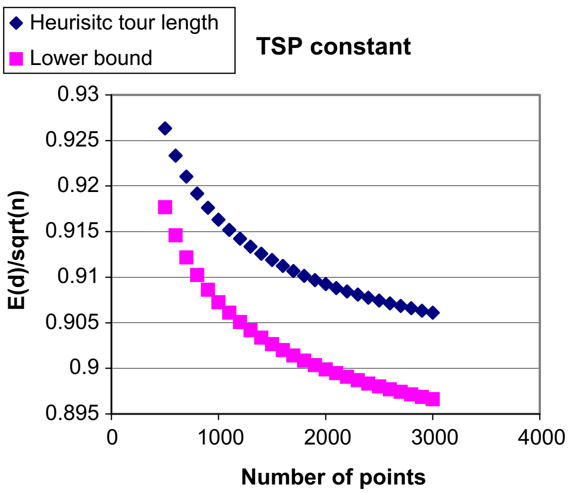

Fig. 3. TSP constant for Rectilinear TSP.

ristic, is very small showing that good solutions are obtained by the heuristic.

5. Conclusions

In this paper we derive empirical probability distributions for optimal traveling salesman tour lengths. The city locations are assumed to be drawn from a uniform 1×1 grid. We give regression equations to estimate the first four moments of the distribution as well regression equations to estimate the parameters of the Beta distribution. Our results indicate excellent fit of the proposed Beta distributions as well as all regression equations. These results can be used to speed up simulations involving traveling repair persons and to provide first and second moments for the calculation of expected service times in traveling repair person models. To our knowledge, this is the first comprehensive study of the distribution of optimal tour lengths.

Acknowledgements

The authors would like to thank two anonymous referees whose suggestions greatly improved the presentation in this paper.

References

Beardwood, J., Halton, J.H., Hammersley, J.M., 1959. The shortest path through many points. Proceedings of the Cambridge Philological Society 55, 299–327.

Bonomi, E., Lutton, J.L., 1984. The N-city traveling salesman problem and the Metropolis algorithm. SIAM Review 26, 551–568.

Few, L., 1955. The shortest path and the shortest road through *n* points. Mathematika 2, 141–144.

Fiechter, C.-N., 1994. A parallel tabu search algorithm for large traveling salesman problems. Discrete Applied Mathematics 51 (3), 243–267.

Ghosh, B., 1951. Random distances within a rectangle and between two rectangles. Bulletin of the Calcutta Mathematical Society 43, 17–24.

Golden, B.L., Alt, F.B., 1979. Interval estimation of a global optimum for large combinatorial problems. Naval Research Logistics Quarterly 26, 69–77.

Gutin, G., Punnen, A.P. (Eds.), 2002. The Traveling Salesman Problems and Its Variations. Kluwer Acdemic Publishers, pp. 369–443.

Helsgaun, K., 2000. An effective implementation of the Lin–Kernighan traveling salesman heuristic. European Journal of Operational Research 126 (1), 106–130.

Johnson, D.S., McGeoch, L.A., 2002. Experimental analysis of heuristic for the STSP. In: Gutin, G., Punnen, A.P. (Eds.), The Traveling Salesman Problems and Its Variations. Kluwer Acdemic Publishers, pp. 369–443.

Johnson, D.S., McGeoch, L.A., Rothberg, E.E., 1996. Asymptotic experimental analysis for the Held–Karp traveling salesman bound, Proceedings of the seventh ACM–SIAM Symposium on Discrete Algorithms.

Karloff, H.J., 1989. How long can a Euclidean traveling salesman tour be. SIAM Journal on Discrete Mathematics 2, 91–99.

Krauth, W., Mizard, M., 1989. The cavity method and the traveling-salesman problem. European Physics Letters 8, 213–218.

Los, M., Lardinois, C., 1982. Combinatorial programming, statistical optimization and the optimal transportation network problem. Transgenic Research Part B 16, 89–124.

Lee, J., Choi, M.Y., 1994. Optimization by multi-canonical annealing and the traveling salesman problem. Physics Review E 50, R651–R654.

Mahalanobis, P.C., 1940. A sample survey of the acreage under jute in Bengal. Sankhya Series B 4, 511–531.

Marks, E.S., 1948. A lower bound for the expected travel among *m* random points. The Annals of Mathematical Statistics 19, 419–422.

Marse, K., Roberts, S.D., 1985. Implementing a portable pseudo-random generator. Applied Statistics 34, 198–200.

Norman, M.G., Moscato, P., 1995. The Euclidean traveling salesman problem and a space-filling curve. Chaos, Solitons and Fractals 6, 389–397.

Ong, H.L., Huang, H.C., 1989. Asymptotic expected performance of some TSP heuristics: An experimental evaluation. European Journal of Operational Research 43, 231–238.

Percus, A.G., Martin, O.C., 1996. Finite size and dimensional dependence in the Euclidean traveling salesman problem. Physics Review Letters 76, 1188–1191.

Stein, D., 1997. Scheduling Dial-a-Ride Transportation Systems: An Asymptotic Approach. Ph.D. Dissertation, Harvard University, Cambridge, MA.