A COMPARISON OF BIG *R* AND THE TSP MULTIVARIATE CORRELATION STATISTICS

William Conley
Professor of Business Administration (Statistics)
University of Wisconsin at Green Bay
Green Bay, Wisconsin 54311-7001, U.S.A.
Email: conleyw@uwgb.edu

KEYWORDS
TSP, CTSP, *R* comparisons, multivariate data

ABSTRACT

Presented here is a discussion of the computer age based CTSP and TSP measures for multivariate correlation. Three examples are reviewed and the multivariate big *R* (which currently dominates correlation analysis) will be compared with a CTSP or TSP analysis in each case. CTSP is effective in initial screening of large data sets.

The data featured here consist of a four variable linear relationship, a five variable "ellipsoid" relationship and lastly a seven variable "elliptic paraboloid."

INTRODUCTION

The classic traveling salesman problem (TSP) of finding the shortest route to connect *n* points in a closed loop tour can be solved (or at least approximated) for fairly large *n* using the simulation based multi stage Monte Carlo optimization (MSMCO) with a desktop PC. (See (Conley 1991) and (Conley 2000) for examples.) Notice that (Conley 2000) makes the jump from the usual two dimensional (or occasionally three dimensional) TSPs to six dimensional "shortest route" problems. Typically these problems are applied in operations research to reduce transportation costs and improve customer satisfaction.

However, the TSP approach with MSMCO can be applied to connecting data points in a "shortest route" to discover relationships between the variables that gave rise to the data. Let us look at three examples to see if our 21st century PCs can help us with some data analysis.

THE MSMCO TSP SIMULATION ALGORITHM

Multi stage Monte Carlo optimization (MSMCO) is a general purpose simulation based optimization technique for linear and nonlinear problems. It makes repeated "random" searches in an ever narrowing and repositioning search area (starting with the whole feasible solution space initially) which is continually recentered about the best answer obtained so far as the simulation proceeds. This MSMCO approach in a sense has a look around the feasible solution space and then just heads in the direction that the best answers are coming from and then closes in on and finds the exact solution or a near approximation.

It has been further adapted for the mathematically difficult traveling salesman problem (TSP) of finding the shortest route to connect *n* points in a closed loop. The coordinates of the *n* points are read into the MSMCO TSP simulation algorithm. Then the distances from each point to each other point are calculated and stored in an array. Then this array is column ranked from smallest to largest. Therefore, as an example, the first entry in column *i* (representing the *i*th point of the 1, 2, 3,…*n* points) would be the distance from point *i* to its nearest point. The second point in the *i*th column would be the distance from point *i* to its second nearest point. The third point down from the top in the *i*th column would represent the third closest point to point *i* and so on.

Then the MSMCO simulation would be done on the subscripts of this ranked array, always closing in on and repositioning as better and better shortest routes are found (connecting the *n* points in a closed loop). The simulation should be started about 3 or 4 subscripts down from the top of the array for each point for best results. This initially puts the MSMCO TSP search in the region where the best answer will be (points that are not too far away from the points). Also, an important feature is continually checking the arrays to see that no doubling back takes place (each point is used once).

Even though MSMCO TSP is a "random" simulation approach (Conley 1991) has shown superior results to competing TSP algorithms on the famous (in the mathematics literature) Problems number 30 and 32. Additionally, the MSMCO TSP algorithm is adjusted for four, five and seven dimensions for the four, five and seven variable correlation problems presented here. Also, the definition of correlation is being expanded to include the recognition of any type of pattern between the variables. Once this is established perhaps additional study with the standard multivariate *R* or the new DTSP distribution analysis (Conley 2005), TSP statistic will reveal more about the variables relationship.

Proceedings 20th European Conference on Modelling and Simulation
Wolfgang Borutzky, Alessandra Orsoni, Richard Zobel © ECMS, 2006
ISBN 0-9553018-0-7 / ISBN 0-9553018-1-5 (CD)

A FOUR VARIABLE PROBLEM

The following 70 lines of data were collected on four variables.

Table 1: Four Variable Data

	X_1	X_2	X_3	X_4
1	52	100	38	58
2	1	60	39	27
3	71	100	69	76
4	75	81	19	56
5	78	90	46	69
6	28	77	1	29
7	52	7	94	56
8	54	7	75	50
9	70	81	63	70
10	36	37	6	25
11	3	32	55	27
12	74	56	84	73
13	56	54	27	45
14	87	25	91	74
15	46	19	9	26
16	31	20	47	34
17	45	35	12	31
18	17	62	12	25
19	75	69	46	63
20	10	50	100	50
21	24	92	21	38
22	55	9	80	53
23	69	14	81	60
24	76	28	1	38
25	68	31	89	66
26	100	85	12	65
27	92	45	61	70
28	66	91	53	67
29	22	11	75	38
30	35	32	79	49
31	84	85	71	79
32	7	6	75	30
33	24	33	79	45
34	30	56	65	48
35	68	1	18	35
36	9	0	23	11
37	90	35	26	54
38	29	95	96	67
39	50	15	51	42
40	59	35	94	65
41	5	4	80	31
42	64	30	22	41
43	16	83	40	39
44	21	13	99	46
45	59	75	24	50
46	38	98	35	50
47	35	0	95	47
48	51	92	91	74
49	100	12	88	75
50	47	61	47	50
51	55	21	44	43
52	77	55	52	63
53	73	42	13	44
54	61	41	5	36
55	1	30	65	30
56	9	48	99	49
57	95	20	96	78
58	33	100	96	70
59	94	63	51	71
60	48	50	72	56
61	67	56	39	54
62	100	33	16	55
63	13	95	34	39
64	65	88	49	64
65	76	96	98	88
66	74	97	18	59
67	43	43	79	55
68	13	26	37	24
69	83	28	43	56
70	53	75	74	65

The MSMCO approach (repeated ever more focused and narrowed simulation based solution attempts) is used to find a shortest route in four dimensional space for the data. This value is A=1516.947 (using the four dimensional generalization of the Pythagoreon Theorem for a distance measure). Then shortest routes were similarly found for four random sets of 70 four dimensional points in the same ranges as the original data (1-100 for X_1, 0-100 for X_2, 1-100 for X_3 and 11-88 for X_4).

These shortest routes were 2025.959, 2022.519, 1982.501 and 2058.072. We let CTSP=A/(median of the random shortest routes) or in this case CTSP=1516.947/((2025.959+2022.519)/2)=.7494.

The 4x3=12 ratios (CTSP's for the random data) are all in the range .9633 to 1.038 (example 1982.501/2058.072=.9633). Therefore, our CTSP=.7494 is statistically significantly less than this range where most of these ratios would be under the null hypothesis of no correlation between the four variables under consideration. Therefore, the null hypothesis can be rejected and the alternative hypothesis that the variables are correlated is accepted.

A calculation of the standard multivariate big *R* for this data also confirms a relationship (*R*=.999 in this case). It turns out that the relationship is linear and Equation (1) fits the data fairly well.

$$Y=X_4=.42X_1+.23X_2+.35X_3 \quad (1)$$

A FIVE VARIABLE PROBLEM

The following 109 lines of data were collected on five variables.

Table 2: Five Variable Data

Point number	X_1	X_2	X_3	X_4	X_5
1	74	70	75	26	67
2	37	27	35	69	84
3	88	68	38	37	70
4	83	65	68	27	68
5	30	20	56	52	83
6	52	50	41	3	64
7	66	36	43	39	93
8	40	52	49	40	98
9	49	82	47	69	82
10	41	35	55	7	68
11	36	23	23	37	75
12	73	12	42	39	66
13	75	35	60	43	89
14	42	18	49	22	74
15	31	53	33	37	91
16	24	46	57	34	89
17	43	60	78	34	86
18	56	25	63	52	90
19	35	73	47	45	93
20	49	82	48	71	81
21	51	54	73	56	94
22	36	56	6	61	64
23	35	64	66	23	83
24	9	40	48	33	71
25	80	38	36	77	73
26	22	59	50	80	77
27	39	34	64	74	87
28	65	31	17	33	73
29	44	8	71	37	51
30	82	77	36	46	72
31	24	62	49	51	91
32	59	50	82	45	87
33	69	78	57	57	85
34	60	77	24	77	65
35	62	38	85	64	78
36	30	51	37	75	86
37	41	38	67	59	94
38	51	56	89	26	69
39	50	50	65	41	97
40	42	36	40	22	87
41	34	43	54	41	96
42	49	85	67	61	78
43	45	55	26	80	81
44	30	43	89	46	73
45	19	53	39	66	84
46	53	36	59	25	90
47	72	85	67	56	70
48	12	57	23	40	63
49	22	45	72	20	67
50	60	13	58	22	61
51	85	52	72	66	73
52	22	74	33	57	78
53	33	57	48	84	82
54	54	41	72	67	90
55	47	37	12	50	79
56	39	88	69	40	30
57	65	60	80	79	44
58	81	25	51	78	40
59	78	23	63	65	26
60	43	18	42	81	31
61	30	74	65	37	17
62	43	60	79	36	14
63	29	55	66	9	40
64	37	47	58	25	10
65	65	54	49	20	13
66	88	73	66	55	35
67	56	58	55	14	17
68	54	18	41	84	37
69	81	43	56	50	12
70	50	63	29	22	17
71	55	78	73	52	17
72	52	52	47	28	5
73	57	46	76	44	9
74	84	81	43	52	33
75	23	80	61	67	30
76	78	68	21	39	30
77	66	18	39	71	25
78	54	33	30	89	34
79	27	13	49	69	37
80	34	64	73	75	20
81	68	79	76	54	25
82	82	57	29	74	30
83	46	75	33	45	11
84	79	56	45	74	18
85	45	44	70	24	13
86	42	21	64	86	42
87	36	31	69	28	17
88	16	50	50	62	15
89	40	51	45	20	12
90	71	39	16	60	24
91	60	64	51	44	4
92	52	76	85	41	28
93	66	53	40	78	13
94	51	71	74	12	46
95	40	85	48	46	17
96	16	70	65	75	41
97	53	54	58	70	5
98	64	48	24	24	19
99	96	51	66	45	40
100	51	89	70	42	29
101	51	67	48	4	41
102	72	10	40	44	36
103	92	45	75	44	43
104	22	80	74	37	45
105	34	66	91	53	33
106	77	52	38	75	18
107	56	27	80	22	35
108	64	88	50	33	28
109	52	73	71	83	30

Again, the MSMCO TSP algorithm (modified for five dimensions) was used and produced a shortest route (connecting the data in a closed loop) of A=2818.906. (This used less than one minute of computer time on a 2003 model PC.)

Then shortest routes were similarly found for four random sets of 109 five dimensional points in the same ranges as the data under consideration (9-96 for X_1, 12-89 for X_2, 6-91 for X_3, 3-89 for X_4 and 4-98 for X_5). These shortest routes were 3505.722, 3591.132, 3557.813 and 3593.164. We use the median of these for B in our CTSP=A/B. This gives us CTSP=2818.906/((3591.132+3557.813)/2)=.7886.

The 4x3=12 ratios (CTSP's for the random data) are all in the range .9757 to 1.025. Therefore, CTSP=.7886 is statistically significant and the null hypothesis of no correlation can be rejected. The idea here is that points that are closer together (than random points in the same ranges) are related in some fashion. Note that this is a more general view of correlation that looks for any relationship.

The big *R* did not do so well on this problem yielding a value of *R*=.233 for our data. Therefore, perhaps the relationship is not linear.

It turns out that Equation (2) fits the data well. This is an "ellipsoid" (rugby ball shape) generalized to five dimensions.

$$\begin{aligned} &.952X_1^2 - 95.24X_1 + X_2^2 - 100X_2 + .952X_3^2 - \\ &95.24X_3 + .952X_4^2 - 95.24X_4 + .952X_5^2 - \\ &95.24X_5 + 9235.80 = 0 \end{aligned} \qquad (2)$$

A POLLUTION REDUCTION MODEL

A power plant has come under government pressure to reduce its emissions of a certain chemical compound. Its environmental and chemical engineers have tried extensive experiments over 80 days of adding varying amounts (in kilograms) of six chemicals that should reduce the emissions of the troublesome compound.

The data is given below with X_1 through X_6 being the varying amounts (in kilograms) of the six abatement chemicals and $Y=X_7$ is the percentage amount of the dangerous compound still going into the atmosphere each day when compared with current daily unacceptable 100%. Therefore, if $Y=X_7=33$, for example, the polluting chemical compound daily emission is about one third of its previous value. A value of $Y=X_7=0$ would indicate complete success in removing the pollutant.

Table 3: Seven Variable Data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	100	80	94	50	86	66	46
2	67	41	63	4	84	52	25
3	83	43	32	17	21	75	27
4	65	7	83	16	86	52	38
5	100	66	100	9	88	26	60
6	39	85	55	89	40	64	21
7	99	43	100	24	55	5	51
8	15	51	63	24	97	54	29
9	72	67	69	19	23	69	21
10	81	1	14	73	98	32	52
11	93	33	97	69	67	96	47
12	91	8	72	41	16	36	36
13	10	22	40	43	2	40	33
14	17	9	5	56	74	46	36
15	70	64	68	40	51	46	7
16	80	95	63	46	21	27	30
17	74	21	87	17	54	68	28
18	59	21	17	16	51	52	21
19	3	95	17	87	47	23	50
20	75	80	13	87	89	66	40
21	27	23	99	73	69	71	33
22	50	1	86	80	21	41	37
23	30	54	53	43	96	82	24
24	42	81	100	84	21	100	54
25	78	61	85	61	27	26	22
26	90	54	36	46	55	24	17
27	80	82	91	16	5	100	62
28	7	83	10	49	100	51	47
29	95	99	52	50	41	2	45
30	34	69	0	12	64	67	34
31	29	60	98	45	18	10	37
32	47	17	99	16	48	37	32
33	83	83	62	48	46	48	16
34	81	26	61	24	16	88	33
35	2	83	58	20	17	26	40
36	93	5	34	51	100	46	44
37	54	90	48	0	100	65	46
38	52	36	64	58	86	96	26
39	13	25	29	49	32	64	20
40	82	80	59	55	6	62	27
41	86	73	88	78	99	97	58
42	2	66	55	51	19	75	28
43	49	93	11	66	83	13	41
44	51	53	68	39	38	91	15
45	60	40	9	11	87	100	49
46	69	10	72	27	34	68	24
47	91	62	80	22	44	61	24
48	87	90	74	56	89	27	38
49	95	62	41	0	47	71	35
50	95	47	60	42	25	82	26
51	27	58	100	77	84	18	40
52	30	43	93	63	61	9	29
53	100	17	37	64	52	50	26
54	1	56	12	90	7	74	53
55	11	50	98	32	44	77	33
56	7	95	78	34	26	59	37
57	100	19	38	55	89	41	35
58	61	28	43	2	45	38	21
59	38	70	79	48	100	76	30
60	10	20	27	86	57	94	42
61	92	88	13	21	1	32	54
62	48	34	92	31	90	50	27
63	62	16	61	9	34	14	31
64	9	61	73	100	48	18	39
65	92	10	67	100	20	20	53
66	6	9	0	27	52	66	46

67	34	96	53	89	53	44	26
68	91	70	22	53	73	66	23
69	31	30	8	86	26	5	43
70	19	9	85	9	26	68	43
71	28	45	13	91	6	60	37
72	0	33	92	38	30	49	34
73	17	95	100	64	6	92	63
74	55	51	13	45	42	55	10
75	64	66	32	16	14	45	22
76	97	95	4	16	54	73	54
77	89	44	94	62	53	71	27
78	92	83	92	14	97	29	57
79	85	33	62	72	81	48	21
80	61	39	2	80	39	39	25

Therefore, the MSMCO TSP algorithm is adjusted for seven dimensions and n=80 sets of data points. About one minute of PC run time reveals a shortest route of TSP=4094.999. Then in the ranges of the data under consideration (0-100 for X_1, 1-90 for X_2, 1-100 for X_3, 0-100 for X_4, 1-100 for X_5, 2-100 for X_6 and 7-63 for X_7) ten sets of n=80 random seven dimensional points were selected and their shortest route TSP values were calculated to be 4245.925, 4139.702, 4412.964, 4256.414, 4302.403, 4269.953, 4410.044, 4415.784, 4265.486 and 4461.941. Notice that our TSP=4094.999 is less than all of these. However, because the values are close, we will look at the distribution of the TSPs more directly.

If the hypothesis of no correlation (in our pollution abatement study) is correct, then it would be a fifty-fifty (H_0: p=.5) or even chance that any of the shortest routes, were more or less than our data standard of 4094.999.

However, our 4094.999 was less than all ten random tries; therefore, $.5^{10}$=.000977. So there is less than a chance in one thousand of being incorrect if we reject the hypothesis of no correlation, and conclude that our six chemicals do influence the abatement of the dangerous chemical.

Again, R=.218 (for our data) was not too helpful because of the nonlinear nature of the process.

$$Y = X_7 = .00667\sum_{i=1}^{6}(X_i - 50)^2 \qquad (3)$$

Equation (3) is a pretty good fit for the data. This is an "elliptical paraboloid" (cereal bowl shape) extended to seven dimensions.

The problem is further complicated by the fact that the company's accountants want to keep the daily cost (of this abatement process) under 1000 dollars a day. The cost per kilogram of the six input chemicals are 3, 12, 6, 8, 4, and 2 dollars respectively. Also, the engineers know that the power generation process can not efficiently sustain more than 200 kilograms of these chemicals per day.

Therefore, a short MSMCO program to minimize Equation (4).

$$Y = .00667\sum_{i=1}^{6}(X_i - 50)^2 \qquad (4)$$

subject to

$X_i \geq 0$, i=1, 2, 3, 4, 5, 6

$3X_1 + 12X_2 + 6X_3 + 8X_4 + 4X_5 + 2X_6 \leq 1000$ dollars

and $X_1 + X_2 + X_3 + X_4 + X_5 + X_6 \leq 200$ kilograms

yields

X_1=40 X_2=18 X_3=33

X_4=29 X_5=37 X_6=43

With Y=13.813 percent of the harmful pollutant.

Note that the MSMCO technique is a completely general optimization algorithm. In addition to minimizing in this problem, maximizing a function with multi stage Monte Carlo optimization (if that is desired in say a profit equation) is also possible.

AN EXPANDED VIEW OF CORRELATION

One of the classic statistics books to deal with the multivariate correlation coefficient R (Anderson 2003) presents much of the subsequent analysis of R assuming that samples were taken from multivariate normal distributions. The new TSP and CTSP correlation statistics do no dispute any of this classical theoretical analysis. However, CTSP (correlation using the traveling salesman problem) is expanding the idea of correlation to include the detection of any pattern between the k variables represented by k columns of n rows of numbers.

Think of n points on a circle. The shortest route connecting those points (approximately 3.14159 times the diameter) will be surely shorter than the shortest route connecting n random points (no pattern) in the same ranges. Think of three dimensional points on a pyramid. The shortest route connecting them will surely be less than the shortest route connecting the same number of random points in the same ranges.

Shorter shortest routes indicate a "topologically" reduced space where the points are landing because they are following a pattern. Once this is established then the classical multivariate R coefficient analysis could be pursued. Additionally, the new DTSP multivariate distribution tests (Conley 2005) or the chi square goodness of fit tests could help to establish which distributions the data may have come from.

STATISTICAL OPTIMIZATION

Computers are being used today to carry out the calculations (for large data sets) for the classical statistics that have been developed over the last few hundred years. This, of course, saves a lot of time and money in applications in engineering, science and business.

However, the new TSP class of statistics (especially CTSP and DTSP for correlation and distribution identification) are an attempt to develop a new set of sophisticated and powerful multivariate statistics for work on the most complex statistics problems. The solution technique used to carry out the TSP statistics calculations (when screening large data sets) has been multi stage Monte Carlo optimization (MSMCO). This computer simulation based optimization technique (and others like it) are really working in a new developing field of statistics (made possible by computer simulation) that could be called statistical optimization which complements classical statistics. It is using statistical techniques and fast desktop computers to quickly cross the feasible solution space of any multivariate optimization problem to its minimum (minimum cost or pollution or error, etc.) or to its maximum (maximum profit or yield, etc.)

The presentation here of the CTSP multivariate correlation examples is really using both areas of statistics. It is using the new statistical optimization techniques to solve difficult TSP problems with the goal of using those results to make decisions (classical statistics) about whether correlations or patterns exist between variables in an initial screening of large multivariate data sets.

Then if it appears as though a pattern or correlation exists, subsets of the variables could also be tested and some of the classical statistical tests (R, r and least squares, for examples) could be used to further identify the complex relationships that may be present between the variables.

CONCLUSION

The challenge in statistics for hundreds of years was to standard normalize as many formulas and statistics as possible to take advantage of the power of the central limit theorem. This is still important. The fact that Equation (5) is approximately standard normally distributed (for large n) remains very useful today.

$$Z = (\overline{X} - \mu) / S / \sqrt{n} \qquad (5)$$

However, our computer age (and new simulation based optimization algorithms such as MSMCO) afford us the opportunity to create new statistics to try to shed light on the most complex multivariate relationships. The CTSP and TSP correlation statistics hopefully will complement big R in that quest.

REFERENCES

Anderson, T.W. 2003 *An Introduction to Multivariate Statistical Analysis*, 3rd edition, Wiley and Sons, New York.

Conley, W.C. 1991 "Programming an Automated Punch or Drill." International Journal of Systems Science, Vol. 22, No. 11, 2039-2056.

Conley, W.C. 2000 "Statistical Simulation and a Six Dimensional Shortest Route Problem." Proceedings of the Japanese Society for Simulation Technology, JSST 2000, Tokyo, 90-93.

Conley, W.C. 2005 "A New Multivariate Distribution Test Applied to Health Care Research." Proceedings of the 2005 International Conference on Health Science Simulation (New Orleans) SCS San Diego, 55-60.

AUTHOR BIOGRAPHY

William C. Conley was born in Lansing, Michigan, USA and went to Albion College where he studied mathematics and received a degree with honors in 1970. He then earned a masters degree in mathematics from Western Michigan University in 1971 and an M.Sc. and Phd in mathematics and computer statistics from the University of Windsor in 1973 and 1976. He joined the faculty of the University of Wisconsin at Green Bay in 1977 and is now professor of business administration and statistics. He teaches introduction and advanced business statistics courses and does simulation and computer statistics research. The developer of multi stage Monte Carlo optimization and the TSP class of statistics, he is the author of five books and more than 185 publications worldwide. His favorite research problems are the shortest route or traveling salesman problems (TSPs) especially the ones adapted to higher dimensions (k = 3, 4, 5, 6, 7, …). He is a member of the American Chemical Society and a senior member of SCS since 1994. He is a fellow of the Institution of Electronic and Telecommunication Engineers of India, a member of Phi Beta Kappa (national academic honorary), Omicron Delta Kappa (national leadership honorary), Omicron Delta Epsilon (national economics honorary), Kappa Mu Epsilon (national mathematics honorary), Sigma Beta Delta (international honor society in business, management and administration) and a Michigan Scholar in College teaching. He was elected to the Albion College Athletics Hall of Fame in 1995 (soccer football) and again in 2005 (golf). He was named to Marquis *Who's Who in America* 2005 and *Who's Who in the World* in 2006.