

OPTIMIZATION AND DEVIATION WITH THE TRAVELING SALESMAN PROBLEM IN REVERSE

William Conley
University of Wisconsin at Green Bay
Business Administration and Statistics
2420 Nicolet Drive
Green Bay, Wisconsin 54311-7001 U.S.A.
email: conleyw@uwgb.edu

KEYWORDS

VTSPA statistic, multivariate deviation test, longest routes

ABSTRACT

Analysis of variance (if normality conditions hold) and the Kruskal Wallis test present ways to test multiple populations for equal averages. Presented here is a new approach to testing multiple populations for equal deviation or variability. Its VTSPA statistic calculations are based on finding the longest routes to connect the data points rather than the well known reverse of that concept, finding the shortest routes to connect points, the so called "traveling salesman problem" or TSP. The mathematically difficult TSP and its reverse can be handled with approximation techniques made viable by the fast computers currently available. Various versions of the traveling salesman problem (TSP) can do much more than route delivery trucks or sales people to customers in an optimal fashion. They can form the basis for the TSP class of statistics that help to analyze large multivariate data sets.

INTRODUCTION

A researcher, engineer, scientist or business person is frequently confronted with large data sets. When this happens the question that arises initially is, does this data mean anything? Can it be reduced to a useful piece of information? Has production increased? Has quality improved? Are the factors correlated? In fact, an old name for statistics was "data reduction." The idea was to reduce the data to one or two useful conclusions that could be advantageously acted upon.

The CTSP statistic discussed in (Conley 2003) showed how to use an adaptation of the TSP problem with multi stage Monte Carlo optimization (MSMCO) to see if k factors were correlated using n data points for the statistical test. Also, the TSP and MSMCO based MTSP statistics in (Conley 2003) presented a general purpose multivariate test for averages. Continuing with this theme, the VTSPA statistic tests k populations to see if they have equal deviation or variability. However, variability or deviation is concerned with how far apart data points are, so we will use the TSP in reverse and find the longest routes (instead

of the shortest routes) connecting the k dimensional data points.

Let us look at a five and a nine variable example from business and industry.

FIVE SALES TEAMS

A supervisor of five sales teams looks at the number of units of product ABXC sold in each of the last fifty-five weeks (Table 1) by the sales teams working for her. She will, of course, study the means and total sales production by each team separately and together. However, the supervisor wonders if some subtle or large differences in deviation or variability in sales production could also be tested for. Therefore, the null hypothesis of equal deviation of the five populations will be tested with the VTSPA=A/B statistic on the data (given below) using the farthest available point algorithm (MSMCO (Conley 1993) and (Conley 1994) could also be used adjusted for maximizing instead of minimizing).

Table 1: Five Sales Teams Data

Week Number	Team 1	Team 2	Team 3	Team 4	Team 5
1	3	27	2	42	31
2	83	49	50	66	95
3	37	40	36	49	50
4	37	27	88	70	15
5	13	48	100	42	57
6	60	21	55	70	39
7	38	28	100	57	6
8	90	38	100	58	40
9	57	35	32	65	32
10	51	45	86	70	8
11	6	24	21	70	84
12	34	50	50	49	26
13	93	52	7	57	31
14	25	27	3	41	95
15	34	37	96	66	95
16	32	39	52	63	39
17	35	24	42	51	17
18	7	30	27	64	69
19	90	46	41	70	29
20	68	24	34	67	46

21	87	42	99	54	75
22	83	49	53	70	95
23	53	39	58	70	21
24	21	47	16	46	70
25	37	33	98	69	43
26	18	41	20	47	50
27	85	30	89	44	12
28	33	33	35	47	21
29	23	50	52	53	78
30	9	50	73	68	23
31	36	40	31	46	43
32	65	40	74	56	27
33	40	38	55	43	91
34	28	51	9	70	57
35	19	34	99	60	77
36	49	39	88	66	67
37	46	29	34	51	91
38	42	48	19	48	70
39	3	28	17	66	80
40	18	31	43	46	34
41	21	21	69	63	78
42	89	49	8	45	23
43	74	21	19	67	82
44	0	52	92	52	14
45	50	31	35	70	18
46	5	24	22	44	82
47	61	47	20	61	74
48	1	44	54	45	21
49	77	32	96	43	33
50	86	48	15	70	28
51	7	31	41	46	60
52	71	34	98	60	90
53	25	47	78	66	60
54	92	46	15	59	77
55	73	22	41	69	23

Briefly, one of the 55 five dimensional points is selected at random. Then, this point's distance to the point furthest away from it (in five dimensional space) of the other 54 points is calculated and added to the A total. Then from this second point, the distance to the point furthest away from it, of the remaining 53 points is calculated and added to the A total. Then from this third point, the point (of the remaining 52 points) furthest away from it is identified and this distance is calculated and added to the A total. This process continues until all 55 points are connected with a total distance of $A=5010.41$ for the data given here.

Then in the range of the data (0-100 in this case) five sets of 55 points were drawn at random and their longest route distances were calculated to be 6386.74, 6152.23, 6694.64, 6410.67 and 6527.27. Taking the median of these values for B, we get $VTSPA=A/B=5010.41/6410.67=.7816$. The $5 \times 4 = 20$ A/B quotients from the five random sets of data are all greater than $6152.23/6694.64=.91898$. Therefore, the null hypothesis of equal deviation can be confidently rejected, because the $VTSPA=.7816$ did not occur by chance. It happened because the data was much more compact in some of the five dimensions than in the others.

Therefore, the supervisor can look into this for keys to improving or steadying the performances of her sales teams.

Geometrically, the idea is, if for example, data were collected on three factors of equal variability, the graph of these points in three dimensions could be covered by a cardboard box of fairly equal dimensions in each direction (like a cube). However, if three factors were such that one factor's points were in the range of the length of the cube, but the other two were much more compact (less deviation) then their graph could be covered by an object with different dimensions in the three directions, such as a cricket bat. Therefore, the longest route connecting the points covered by a cricket bat shape will be much shorter than the longest route connecting the points covered by the cube. VTSPA exploits this difference to test for deviations.

Let us look at another example.

NINE INVESTMENT ADVISORS

The world of investments sometimes looks at deviation or variability as a measure of risk. The idea being if one would earn a six percent rate of return, which is guaranteed on an investment, the risk is virtually zero. However, other investments might be estimated to earn six percent, but any amount of actual return is possible. Then the risk is considered greater.

A manager of a large investment house recorded the yield on investments that nine financial advisors (working for him) obtained for each of sixty individual clients (Table 2) that the manager assigned to them. The data is given below in yields of tenths of a percent. The manager wants to know if there is equal or unequal deviation (or risk) between the nine financial advisors working for him. The idea being that a top performer with low risk might be in the future assigned to the most important projects or clients and rewarded more.

Table 2: Nine Investment Advisors Data

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
1	21	81	16	63	34	31	44	42	34
2	24	37	93	57	85	62	53	52	13
3	36	24	36	70	40	80	45	51	4
4	40	25	24	67	37	40	30	42	51
5	10	24	96	47	95	65	31	62	74
6	45	32	100	61	49	58	49	45	85
7	89	27	82	67	39	53	31	51	31
8	85	41	85	41	90	75	44	64	56
9	65	27	29	58	42	73	51	55	20
10	88	47	67	48	66	61	40	43	77
11	29	22	79	68	74	49	44	55	35
12	81	20	47	58	94	20	58	51	43
13	44	20	96	52	88	80	35	50	100
14	15	23	17	43	17	72	56	54	85
15	84	49	41	48	83	67	41	54	86

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

The same previously described algorithm (always connecting points furthest away, this time in nine dimensional space) is used. The numerator of VTSPA was calculated to be $A=6642.09$ for the data collected by the manager. Then five sets of random data ($n=60$ points in nine dimensions) in the same range as the original data yielded longest routes of 9164.53, 9385.34, 9296.84, 9155.97 and 9071.01. Therefore, $VTSPA=A/B=6642.09/9164.53=.72476$ which is considerably outside the area of $9071.01/9385.34=.96651$ and $1/.96651=1.03465$ where the $5 \times 4 = 20$ quotients from the random data are. Therefore, the hypothesis of equal risk can be confidently rejected. The manager can then use

this information to perhaps improve customer and employee satisfaction. The goal of reducing variability is usually acceptable to all.

FTSP FORECASTING EXTENSION OF CTSP

Given a large multivariate data set, testing it for means (MTSP), variance or deviation (VTSP), an underlying distribution (with the DTSP statistic), and possible correlation among variables (with CTSP) are four important areas of statistical data analysis. However, let us say that a future CTSP statistical test for correlation among the variables represented by the data in question, is successful in showing that a correlation exists. That may be sufficient in some cases. However, in selected applications, the researcher may desire to fit an equation to the highly correlated data. A variation of the CTSP approach for correlation could also be used for curve fitting or so called "forecasting." Let us call this the FTSP approach (F for forecasting with a TSP algorithm).

Briefly, given k columns of n rows of data (representing n samples of k variables), they would first be tested to see if a correlation exists. A shortest route connecting the n points in k dimensional space would be calculated with an appropriate TSP algorithm (MSMCO or other suitable ones). Then repeated random samples of size n in the ranges of the k columns of the original data, would have their shortest routes also calculated. Then if the shortest route from the original data was statistically significantly less than these random data sets' shortest routes, that would indicate the original data was more compact and hence correlated in some way. Keep in mind that this correlation could be functional or not represented by a function. Therefore, in some cases proceeding with a curve fit would be a good idea (but not always).

However, let us say that it is thought that variable one (represented by the first column of data) is dependent on the other $k - 1$ variables. Then for whatever functional form is to be used in the forecast or curve fit, the FTSP approach would be as follows. Use multi stage Monte Carlo optimization (MSMCO) to vary the betas (or constants) in the functional form and close in on the best fit for the data. For each set of betas under consideration (in the MSMCO simulation), random sample an n by $k - 1$ array in the ranges of the $k - 1$ independent variables and evaluate the function to get the new Y values (for column one). Then attach this new n by k array to the n by k original data and find the shortest route connecting these $2n$ k dimensional points. If this distance is less than the comparable distance of the "best answer so far" store these betas. Then proceed centering the rest of the simulation around these betas until a better yet answer occurs. Then store that one and keep proceeding with the MSMCO simulation until the betas are fixed after dozens of subsequent MSMCO stages (hence multi stage optimization).

This FTSP approach should allow the fitting of any type of function (linear or nonlinear) and even some relations. It will require a lot of calculation, but our computer age provides us with powerful, inexpensive desktop computers.

Computers today are used to efficiently carry out the massive calculations necessary to apply the traditionally accepted mathematics solution techniques that have been developed over the last several hundred years. This is obviously an important area of computing. However, let us also develop new statistics and new solution techniques that were unavailable and not possible before the computer age became a reality. Examples of these might be MSMCO and the TSP statistics.

CONCLUSION

Desktop computers are so powerful and available in the 21st century that new statistics can be used to as they say “mine large data sets” effectively and quickly. The TSP class of statistics is an entry into this area of so called data mining. The VTSPA statistic to test for deviation in multiple populations was featured here. An additional seven dimensional example is in (Conley 2003). Also, the new DTSP statistic (Conley 2005) will test multivariate data to see which distribution it may have come from. It makes heavy use of the multistage Monte Carlo simulation shortest route (TSP) adjusted for k dimensions and the standard normal approximation to the runs test for randomness.

Our computer age creates the possibility of the world being overwhelmed by large masses of data. However, it also creates new opportunities for us to “mine it” so to speak to find “precious” pieces of useful information, using the new FTSP, CTSP, MTSP, VTSP and DTSP multivariate simulation based TSP statistics for forecasting, correlation, means, variances and distributions.

REFERENCES

- Conley, W.C. 1993 “Multi Stage Monte Carlo Optimization Applied to a Six Hundred Point Traveling Salesman Problem.” Int J. of Systems Sci., Taylor & Francis, Vol. 24, 609-626.
- Conley, W.C. 1994 “Multi Stage Monte Carlo and Nonlinear Test Problems” Int J of Systems Sci, Taylor and Francis, Vol. 25, 155-171.
- Conley, W.C. 2003 “The TSP Statistics and Total Quality Management and Shipping.” Proceedings of the International Workshop on Harbour, Maritime and Multimodel Logistics Modelling and Simulation, HMS 2003 (September 18-20, 2003, Riga, Latvia) Riga Technical University, 53-57.
- Conley, W.C. 2005 “A New Multivariate Distribution Statistic Applied to Health Care Research.” Proceedings of the 2005 International Conference on Health Sciences Simulation HSS 2005, (January 2005, New Orleans, LA), SCS San Diego, 55-60.

AUTHOR BIOGRAPHY

William C. Conley was born in Lansing, Michigan, U.S.A. and went to Albion College where he studied mathematics and obtained a degree with honors in 1970. He then received 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 statistical optimization 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 ($k = 3, 4, 5, 6, \dots$). He is a member of the American Chemical Society and a senior member of SCS since 1994. He is a fellow in 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), 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 Athletic Hall of Fame in 1995 and was recently named to the 2005 Edition of Marquis's Who's Who in America.