Optimal Data Analysis
Vol. 2, Release 2 (October 26, 2013), 98-105

2155-0182/13/$3.00

Statistically Significant Increases in Crude Mortality Rate of North Dakota Counties Occurring After Massive Environmental Usage of Toxic Chemicals and Biocides Began There in 1998: An Optimal *Static* Statistical Map

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

The use of optimal data analysis (ODA) in making a map reporting the findings of confirmatory statistical analyses is demonstrated by comparing the annual crude mortality rate in counties of North Dakota, before versus after large-scale commercial usage of toxic chemicals and biocides in the environment began there in 1998.

The *Imago Mundi* illustrating Babylon on the Euphrates River is the earliest known map, dating to the 6th to 9th century BCE.[1] Today maps are ubiquitous, and used to illustrate everything from the wealth of nations to different cuts of beef. Some maps, such as physical and topographic maps, display quite stable phenomena. In comparison, road or resource maps require more frequent modifications to remain accurate, because phenomena they portray change more quickly than geological phenomena change. At the other end of this continuum are maps of dynamic phenomena such as weather, spread of infectious disease, results of political polling, or acreage consumed by a forest fire, that require near-real-time updates to be able to effectively augment real-time decision-making.

Unless the nature of change is pre-determined, any phenomenon that changes its state or value is a random variable (unchanging or stable phenomena are constants). Maps are sometimes used to display findings of statistical analysis of random variables. *Statistical maps* usually give exploratory (two-tailed) results: between-group comparisons for a given point in time, or within-group comparisons for consecutive measurements. Statistical methods commonly used to construct such maps include 95% confidence intervals and *t*-tests.[2,3]

2155-0182/13/$3.00

The present paper is first to demonstrate use of UniODA[4] to create a statistical map. It is unknown if this is also the first statistical map to report one-tailed (confirmatory) findings.

Temporal Change in Annual Crude Mortality Rate in North Dakota Since 1998

Fueling the burgeoning American energy boom, extraction of shale oil and natural gas is dramatically improved by combining directional drilling and hydraulic fracturing. Widely known by the moniker *fracking*, horizontal slickwater fracturing breaks (fractures) rock by pressurized liquid.[5] Commercialized in 1998, fracking has inspired global boomtown growth.

However, there is concern over possible short- and long-term human[6] and animal[7] health effects of air and water contamination attributable to additives in the fracking liquid. Fracking a single well typically uses four million gallons of water containing 80 tons of toxic chemicals and biocides including agents such as benzyne and benzene derivatives, glycol-ethers, toluene, ethanol, naphthalene, and methylene chloride, some of which are known as carcinogenic.[8] A study of 353 common fracking chemicals (approximately one-third of the known fracking chemicals) found that exposure to 75% of the chemicals affect skin, eyes, and other sensory organs; 52% affect the nervous system; 40% affect the immune and kidney systems; and 46% affect the cardiovascular system and blood.[9]

Untreated *produced water* ("brine") that is brought to the surface along with the oil or gas has created massive and yet unresolved environmental problems in countries other than the US.[10] Brine includes fracking liquid injected into the well, and water naturally trapped underground that contains the chemical composition of the geologic formation, including naturally occurring radioactive material (NORM).[11,12] Following injection, approximately one quarter of the total produced water generally comes to surface within two months, and the balance in 15-20 years: this *flow-back* is then treated for subsequent reuse, or transported to deep wells for injection.

The largest volume byproduct and waste stream in the field of energy exploration and production, an estimated one million US oil/gas wells generate 2.4 *billion* gallons of produced water *daily*.[13] In addition to the composition (or "formula") of the fracking liquid, factors influencing physical and chemical properties of produced water include geographic location, geological formation, and the type of hydrocarbon product produced.[11]

Other public health hazards unrelated to large-scale commercial use of toxic chemicals and biocides also occur as a typical consequence of boomtown development: increases in social problems such as mental health problems, substance abuse, prostitution, traffic and other accidents, unreliable medical care, crime, and water, food and other resource shortages, are inevitable.[14] And, the drilling profession is inherently hazardous, with workplace accidents in the US resulting in 10 to 20 deaths annually for oil and gas extraction.[15]

Searching for data having direct bearing on the public health correlates of toxic fracking liquid was unproductive. However, data on the annual crude death rate are available for 1937-2005 in North Dakota.[16] For the purpose of this exposition these data were compared before versus after wide-scale use of toxic chemicals in the environment began there in 1998, testing the *a priori* hypothesis that the crude mortality rate increased after 1997. The class variable was year (0 for 1937-1997; 1 for 1998-2005), and the attribute was crude mortality rate. Data were analyzed by the following UniODA[4] code (control commands indicated in red; the *a priori* hypothesis of a higher mortality rate since 1998 is specified using the dir command):

```
open example.dat;
output example.out;
vars year rate;
class year;
```

2155-0182/13/$3.00

```
attr rate;
dir < 0 1;
mcarlo iter 25000;
loo;
go;
```

A reliable and elegantly parsimonious model emerged: if the annual crude death rate is *less than or equal* to 8.9%, then predict the year was 1997 or earlier; and if the annual crude death rate is *greater than* 8.9%, then predict the year was 1998 or more recent. The model made 65% of the possible improvement in accuracy beyond chance: 50% is a strong result for this ESS index. The model correctly predicted 77% of the years before 1998, and 88% afterward. Jackknife validity analysis indicated the model should generalize into the future: predictions of higher annual crude death rates are expected to be 75% accurate—consistent with the accuracy level achieved prior to 1998, and the model is expected to achieve a strong ESS of 52%. The identical analysis was individually repeated for each county, and results are given in Table 1.

Mortality rate cut-point, determined by UniODA, is the value which most accurately separates years before versus after 1998 (this value was 8.9% for the analysis of annual statewide crude mortality, discussed earlier). Cut-points with low values indicate the county has a low annual crude mortality rate, and cut-points with high values indicate the county has a high rate. For the US the current annual crude mortality rate—the number of deaths per 1,000 people—is 8.0, so as seen, 43 of 53 (81%) cut-points exceed the national rate.[17]

Table 1: Summary Findings of Separate UniODA Analyses Conducted for Every North Dakota County

County	Mortality Rate Cut-Point	*p*<	ESS	Jackknife *p*<	Jackknife ESS	Mean Mortality Rate Before 1998	Mean Mortality Rate After 1998	Mean Annual % Increase in Mortality Rate
Adams	11.8	0.0001	80	0.0005	66	9.6	14.3	6.0
Barnes	10.9	0.005	57	0.25	20	10.5	12.5	2.5
Benson	9.6	0.61	17	0.84	-11	10.1	9.7	-0.5
Billings	6.8	0.36	28	0.56	10	5.3	6.0	1.9
Bottineau	11.2	0.008	55	0.13	28	10.4	12.5	2.5
Bowman	12.9	0.0001	78	0.0002	65	10.1	13.9	4.8
Burke	8.0	0.97	2	0.92	-12	10.8	9.6	-1.4
Burleigh	7.6	0.0004	67	0.002	53	6.6	8.3	3.1
Cass	5.9	0.94	5	0.97	-9	7.6	6.6	-1.6
Cavalier	9.6	0.04	46	0.08	33	10.1	11.9	2.2
Dickey	12.8	0.02	52	0.05	38	10.6	13.5	3.4
Divide	15.0	0.001	64	0.003	51	11.2	16.0	5.4
Dunn	9.4	0.003	59	0.01	46	7.8	10.5	4.3
Eddy	15.6	0.002	62	0.005	49	11.3	15.9	5.1
Emmons	11.1	0.0006	68	0.03	42	8.6	12.6	5.7
Foster	15.4	0.0001	84	0.0001	70	10.5	15.4	5.8
Golden Valley	6.1	0.93	5	0.97	-9	9.6	8.7	-1.2
Grand Forks	10.4	0.74	12	0.22	11	7.2	6.8	-0.7
Grant	11.0	0.0001	87	0.0001	74	8.4	14.4	8.9
Griggs	12.0	0.002	62	0.02	50	11.1	15.2	4.6
Hettinger	10.5	0.0001	82	0.0004	68	8.6	12.6	5.8
Kidder	11.6	0.0001	83	0.0001	71	8.3	12.8	6.7

2155-0182/13/$3.00

Lamoure	11.3	0.03	47	0.02	45	9.9	11.8	2.4
Logan	10.9	0.0001	71	0.003	57	8.5	12.4	5.8
McHenry	9.6	0.26	28	0.62	2	10.0	10.5	0.7
McIntosh	16.4	0.0002	71	0.02	45	10.9	17.5	7.5
McKenzie	10.6	0.07	42	0.05	29	8.4	9.7	1.9
McLean	11.2	0.0001	89	0.0001	76	9.3	13.4	5.5
Mercer	8.0	0.008	56	0.03	42	7.6	9.4	3.0
Morton	9.3	0.002	62	0.005	49	8.3	9.4	1.7
Mountrail	11.5	0.03	48	0.07	36	10.8	12.9	2.4
Nelson	16.7	0.003	63	0.009	50	13.4	17.6	3.8
Oliver	7.0	0.29	27	0.31	15	5.7	6.4	1.7
Pembina	8.8	0.18	33	0.26	19	10.6	11.1	0.6
Pierce	12.5	0.0001	87	0.0001	85	9.5	15.0	7.2
Ramsey	12.2	0.0005	67	0.005	54	10.2	12.6	3.0
Ransom	13.9	0.0001	75	0.001	63	11.8	15.4	3.8
Renville	10.7	0.09	40	0.35	13	9.6	10.7	1.4
Richland	7.4	0.62	16	0.74	2	8.9	8.6	-0.4
Rolette	8.6	0.33	25	0.69	-1	9.0	9.9	1.1
Sargent	7.5	0.91	5	0.84	-7	9.7	8.4	-1.6
Sheridan	9.9	0.09	38	0.17	24	8.5	9.5	1.4
Sioux	15.1	0.73	12	0.22	11	9.1	8.5	-0.9
Slope	9.1	0.38	33	0.58	7	6.5	7.4	1.8
Stark	8.6	0.0001	90	0.0001	78	7.6	9.4	2.9
Steele	12.9	0.55	18	0.48	6	9.9	9.6	-0.3
Stutsman	9.8	0.0001	82	0.0002	69	8.9	10.8	2.7
Towner	12.4	0.009	55	0.02	43	10.2	12.6	2.8
Traill	10.5	0.07	41	0.38	14	11.1	12.7	1.8
Walsh	11.5	0.006	58	0.03	44	10.4	12.4	2.4
Ward	7.6	0.002	62	0.02	50	7.4	8.5	1.8
Wells	13.3	0.002	62	0.007	48	10.6	14.2	4.2
Williams	9.0	0.02	53	0.15	26	8.8	9.8	1.5

Note: See text for discussion.

The third column (p<) is Type I error for the *a priori* hypothesis, for the UniODA model. Values in red are statistically significant under the most conservative criteria (experimentwise p <0.05). Otherwise, Type I error for the *a priori* hypothesis is considered statistically significant if generalized p<0.05, or statistically marginal if generalized p<0.10. Effects with p>0.10 are *not* statistically significant. As seen, 16 counties had statistically significant effects at the experimentwise, and 18 at the generalized criterion. Since 2.65 significant models are anticipated under the null hypothesis, these results represent a 13-fold greater effect than was expected by chance.

Simply because something rarely occurs by chance doesn't imply that if chance strikes and something happens—it happens well. An event may indeed be rare, and still not be very good in an absolute sense. The fourth column, ESS (effect strength for sensitivity), is a normed index of effect strength for any application, on which 0 represents the classification accuracy expected by chance, and 100 represents perfect, errorless classification. By rule-of-thumb: ESS values < 25 are considered relatively weak; <50 are moderate; <75 are strong; and $\geq$75 are very strong.[4] As seen, of 53 counties: 11 (21%) had very strong effects; 20 (38%) had strong effects; 13 (25%) had moderate effects; and 9 (17%) had weak effects.

Simply because something happens now is no guarantee that it will happen again in the

2155-0182/13/$3.00

future. Jackknife validity analysis, also called the "leave-one-out" method, is an upper-bound estimate of expected cross-generalizability of a model when applied to an independent random sample.[4] Here jackknife validity estimates the generalizability of the model across time—for predicting future annual crude mortality rates.

Column five (jackknife $p<$) gives Type I error of jackknife test of the *a priori* hypothesis: red values are statistically significant at experimentwise $p<0.05$; other effects are significant if generalized $p<0.05$, marginal if generalized $p<0.10$; and *not* statistically significant if $p>0.10$. Ten counties had statistically significant effects at the experimentwise, and 20 at the generalized criterion, representing an 11-fold greater effect than expected by chance.

Column six provides jackknife ESS for the *a priori* hypothesis. Of 53 counties: 3 (6%) had very strong effects; 15 (28%) had strong effects; 16 (30%) had moderate effects; 13 (25%) had weak effects; and models of 6 (11%) with *negative* ESS values are expected to yield accuracy in future predictions which is weakly lower than that anticipated by chance.

The seventh and eighth columns give the mean annual crude mortality before versus after 1998, respectively. Consistent with the findings regarding high cut-point values, mean crude mortality rate in 43 of 53 (81%) counties exceeded the current national average before 1998, as did 48 (91%) counties after 1997.

The final (ninth) column is the annual mean percentage increase in crude mortality rate for the eight years of data since 1997, that were available for analysis. The annual mean crude mortality increase *exceeds 5% per year* for 12 of 53 (23%) counties.

Constructing an Optimal Map

Figure 1 is a map illustrating findings in the third column of Table 1. The map indicates statistical significance via color: red, orange and yellow respectively indicating experimentwise, generalized and marginal significance, with no significance indicated by green. Also illustrated (not to scale) is the Bakken formation, indicated using a light, broken purple line (see the fourth county from left-hand-side on top of map, and third county from top of map on left-hand-side). Drawn by artist, waterways are not to scale.

As hypothesized, data clearly revealed a significantly higher annual crude death rate in North Dakota after commercialization of toxic fracking liquid occurred in 1998.

In the absence of data which correspond to the mortality rate series, it wasn't possible to adjust rates for age, or to rule-out an aging population as a competing hypothesis. Data indicate that since 2000 the population (more than 90% white) of North Dakota is aging, and illicit drug use has been increasing as well.[18] Interestingly, in some primary producing regions the median age *decreased* as young people pursued boomtown opportunities. Anecdotal reports credit influx of young people, combined with frontier culture (long hard work; surplus money, scarce, and inadequate resources—including health and medical care; and social as well as geographic isolation), as an explanation underlying noted increases in the numbers of assaults, homicides, and vehicular accidents.

For counties with a very strong UniODA model, 75% or more of the increase in mortality rate was explained by the *a priori* hypothesis. This level of classification performance leaves little residual opportunity (unexplained increase in mortality rate) remaining for additional attributes to work with, in an effort to enter the model. It is interesting that a cluster of counties having exceptionally strong effects is seen in the lower west and lower center areas of the state, where there is a west wind component for 9 out of 12 months, and a north wind component for 9 of 12 months. The prevailing wind is pointed directly toward these counties from the producing area, and the affected counties are also down river from the energy-producing region.[19]

2155-0182/13/$3.00

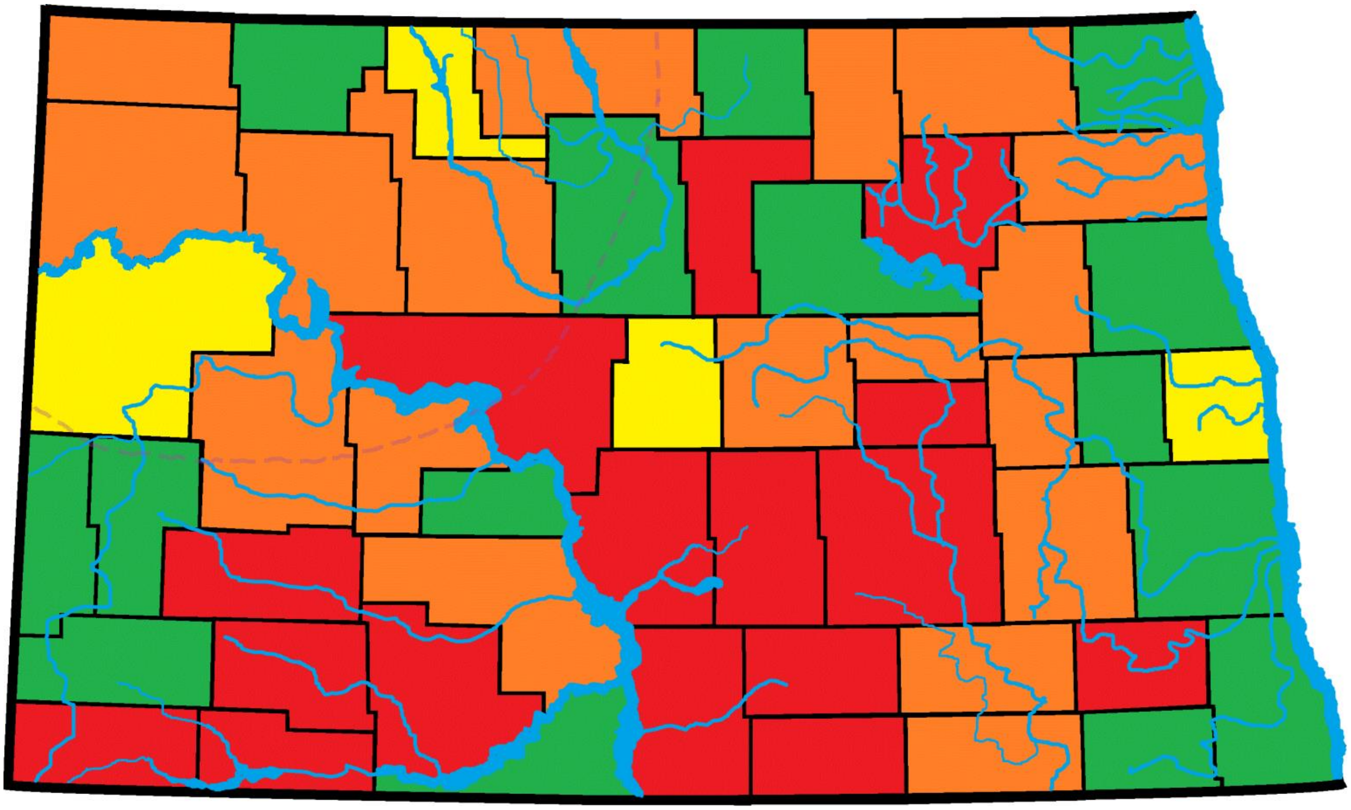

Figure 1: Statistical Reliability of Increase in Annual Crude Mortality After 1997

Improving Statistical Maps

Although confirmatory and exploratory optimal (maximum-accuracy) analyses may be used to create maps capable of addressing hypotheses exactly, powerful improvements of the present map are possible.

First, by presenting the results of one statistical analysis the present map is static, like a snapshot. However, the findings of successive analyses can be integrated and illustrate changes occurring across time, creating a dynamic map like a motion picture (serial overlay presentations are already available[20]). This may be accomplished using a forward-stepping little jiffy procedure[21] with time-ordered series and binary class variable: if the value is higher on current versus prior measurement then class=1, otherwise class=0 (optimal analyses may be weighted by absolute magnitude of the change[4]). First use statistical power analysis[22] to identify the appropriate sample size, N*. For an application with an ordered series of T measurements, a little jiffy has T-N*+1 steps. In step 1 select measurements 1 though N*, run analysis, and code the map for analysis 1. In step 2 select measurements 2 through N*+1, run analysis, and code the map for analysis 2. In step T-N*+1 select the last N* observations in the series, run analysis, and code the map for analysis number T-N* +1. The resulting ordered collection of coded maps may then be presented either serially or dynamically. A demonstration of this methodology is presently under development in the ODA laboratory.

Second, maps published in an eJournal can be made interactive, with pointers used to select a county or counties, and operate menu-

2155-0182/10/$3.00

driven information systems for purposes of information acquisition. If additional serial data streams are available for different attributes, they should be easily selectable for UniODA (illustrated presently) and CTA analysis. Between-group as well as within-subject (illustrated presently) analyses should be available.

The information content within the map can be increased if multiple attributes are available for analysis, because the accuracy of statistical models typically increases if multiple predictors of the class variable are identified and CTA is the analytic engine. Irrespective of whether a total sample (static) or a little jiffy (dynamic) analysis is conducted, in each map every county is simultaneously coded on one dimension for each attribute loading in the model, rather than on only one dimension (presently, outcome of the test of the *a priori* hypothesis). A dynamic multivariate display would help to identify and understand interactions of variables and/or time. Unique discrete states could be indicated using color (with the creative use of primary colors, interactions could be visualized as color shifts) and/or 3D displays. Overlays could be used to display simultaneous phenomena such as storm motion, economic development, and so forth.

Methods illustrated and discussed in this paper can be used to better understand and more lucidly present findings in a host of applications. The series of T measurements may be treated in a confirmatory manner in terms of class variable definition (e.g., measurement before versus after the phenomenon being evaluated), and/or in terms of hypothesized direction of effect (e.g., higher versus lower attribute values). If there are no *a priori* hypotheses to be made, data may be handled in an exploratory manner. In addition, the class variable may be ordered (e.g., in time), or it may be a random variable, and static and/or dynamic maps may be constructed. Examples of applications which might be fruitfully addressed by these methods include political polls; crime surveys; weather phenomena including storms, temperature, and precipitation; agricultural and mining inventories; spread of infectious disease; power consumption; sports team performance; and numerous economic series—money market inflows, consumer debt, unemployment, number of jobs created, consumer sentiment, retail sales, durable goods orders, manufacturing/purchasing indices, inflation, all manners of market indices, and leading indicators, among a cornucopia of others. There remains a great deal to learn about creating and utilizing optimal maps.

References

[1]Raaflaub KA, Talbert RJA (2009). *Geography and ethnography: Perceptions of the world in pre-modern societies*, Wiley.

[2]nces.ed.gov/pubs2009/2009081.pdf

[3]Colizza V, Vespignani A, Hardy EF (2007). *Impact of air travel on global spread of infectious diseases*. Indiana University.

[4]Yarnold, P.R., Soltysik, R.C. (2004). *Optimal Data Analysis: A Guidebook with Software for Windows*, APA Books.

[5]Charlez PA (1997). *Rock mechanics: petroleum applications*. Paris: Editions Technip.

[6]Colborn T, Schultz K, Herrick L, Kwiatkowski C (In Press). An exploratory study of air quality near natural gas operations. *Human and Ecological Risk Assessment: An International Journal*, *17*, 1039-1056.

[7]Bamberger M, Oswald RE (2012). Impacts of gas drilling on human and animal health. *New Solutions: A Journal of Environmental and Occupational Health Policy*, *22*, 51-77.

[8]*Chemicals Used in Hydrolic Fracturing* (April 18, 2011). Committee on Energy and Commerce, US House of Representatives.

[9]Colborn T, Kwiatkowski C, Schultz K, Bachran M (2011). Natural gas operations from public health perspective. *Human and Ecological Risk Assessment: An International Journal*, *17*, 1039-1056.

[10]http://chevrontoxico.com/about/environmental-impacts/produced-water

[11]http://www.netl.doe.gov/technologies/pwmis/intropw/

[12]http://rt.com/op-edge/fracking-radioactive-uranium-danger-ecology-057/

[13]Clark CE, Veil JA (2009). Produced Water Volumes and Management Practices in the United States, ANL/EVS/R-09/1, Environmental Science Division, Argonne National Laboratory.

[14]Jacquet J (2009). Energy boomtowns and natural gas: Implications for Marcellus Shale local governments and rural communities. NERCRD Rural Development Paper No. 43, The Pennsylvania State University.

[15]*Injuries, illnesses, and fatalities in the coal mining industry* (2010). US Bureau of Labor Statistics.

[16]http://ndhealth.gov/vital/stats.htm

[17]http://data.worldbank.org/indicator/SP.DYN.CDRT.IN

[18]http://www.aoa.gov/AoARoot/AoA_Programs/HPW/Behavioral/docs2/North%20Dakota.pdf

[19]http://www.npwrc.usgs.gov/resource/habitat/climate/wind.htm

[20]http://williamahuston.blogspot.com/2013/10/bradford-county-pa-place-with-sweetest.html

[21]Yarnold PR (1992). Statistical analysis for single-case designs. In: FB Bryant, L Heath, E Posavac, J Edwards, E Henderson, Y Suarez-Balcazar, S Tindale (Eds.), *Social Psychological Applications to Social Issues, Volume 2: Methodological Issues in Applied Social Re-search***.** Plenum (pp. 177-197).

[22]Soltysik RC, Yarnold PR (2013). Statistical power of optimal discrimination with one attribute and two classes: One-tailed hypotheses. *Optimal Data Analysis*, *2*, 26-30.

Author Notes

E-mail: Journal@OptimalDataAnalysis.com.

Mail: Optimal Data Analysis, LLC
1220 Rosecrans St., #330
San Diego, CA 92106

ODA Blog: http://odajournal.wordpress.com/