Step-Function Approach to Time-Series: An Air Quality Application

Turkan K. Gardenier, PhD
Director of Research, Pragmatica Corp.
115 Saint Andrews Drive, Vienna, VA 22180, USA.

Abstract

Estimating environmental exposure is a crucial component of health risk assessment. With the recent emphasis on personalized medicine and patient participation, data submitted to analysis and evaluation need to clear, retrievable and easy to merge. Patients, physicians, and public health officials may all be interested in individuals' exposure histories. Large databases are now available and accessible to the public including, patients and caregivers. Some environmental data are monitored under mandate and are available for many years, thus making it feasible to trace change over time. When merging data over time, it is also important to remember that there is uncertainty in data elements; for example, discretized continuous data, incorporate a region of uncertainty between adjacent categories. A step-function based 3-category approach (trinary) is presented and discussed with reference to data collected between 1990-2010 for three criteria pollutants: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Ozone (O3) in a rural and an urban city in northeastern United States. Correlation and regression analysis was applied, as well as a trace of successive data in a step-assignment of +1, 0 or -1 delineated using distance from the 20-year average for each city. Various distance allotments of bandwidth, were explored, including mean + or minus ½ standard deviation, 3rd minus 1st quartiles, 90th minus 10th percentiles and visual inspection. Switching patterns from +1 to 0 range and from 0 to -1 range were examined for CO, NO2 and O3 for the rural and urban areas with a view to quantifying a decision rule for stability of shifting pattern. This paper demonstrates how three-level categorizations of environmental exposure data (and potentially other measurement data as well) can simplify and clarify health related exposure histories. The analytic approach has broad applications in personalized medicine, epidemiology and public health policy.

Key Words: Step-Function; Environmental exposure; Air Quality; Regression analysis; Geographic Information Science (GIS); Trinary categories; Time-series; Uncertainty

1. Introduction

In parallel with rapid advances in computer technology, databases in multiple disciplines have increased in number and magnitude. Interactive features of retrieval now enable the user to query data for a specific geographic location, for a specific time interval, and for specific demographic or age groups. Yet this process of delving into such diverse databases exploring linkages among multiple variables which relate to health and environment raise challenges.

When new types of data are added to each record, e.g., genomic findings, the number of variables for each individual increases. When the same individual is evaluated over time, there are multiple measurements, some for the same attribute and some for the new variables added to the database over time. Similarly, for environmentally-oriented databases such as air or water quality, data are compiled differently, such that the measurement of magnitude of pollution differs in precision. For example, air quality measurements are obtained at monitoring stations which are often distant from the specific location of interest to the researcher who is trying to analyze the exposure level of a specific individual. Or, data may be analog, later to be discretized into "per-hour" or "per day" basis, later to enter into monthly averages. Thus associating data from different record types, although necessary and worthy of merit, introduces uncertainties which demand recognition and integration with analytical procedures.

The field of Geographic Information Science (GIS) uses the term "locational analytics" addressing the issue of multiple attributes for the same location, similar to that in personalized medicine, whereby analysis is directed to all available data for a given individual (Gardenier, 2011). For a specific person or record, data may be obtained at different times, with missing values interspersed. They may emanate from a single measurement or may be averages of multiple measurements.

They may arrive from a specific laboratory or different laboratories, creating other sources of error or uncertainty. Thus, although researchers strive for precision, the inherent lack of uniformity due to a variety of reasons, whether they be from the way in which data are measured, processed, and associated with each individual record or a location, needs to be recognized and reflected in the analytics.

Long term data exist for many environmental- and health-oriented databases through publicly available databases of the Centers for Disease Control and Prevention (CDC), Environmental Protection Agency (EPA) and National Institutes of Health (NIH). Death rates from various causes, including many types of cancer may be traced back to several decades in maps (Pickle, L, Mugniole, M. et al 1996). The maps now have friendly user retrieval interface. Simultaneous inspection of maps enables the user to visualize the relationship among variables and conceptualize any possible association. If the colors or shading are similar, for example, if red, denotes "high" and blue "low" rates in a geographic location across all layers mapped, it becomes easier to identify "hot spots" for further exploration. Viewing joint occurrence of high or low observations across multiple layers, or comparisons along the time dimension are possible, as well. For example, one may classifying each 5-year interval as "high" or "low" based upon the national standards established for a specific pollutant, or as compared with prior observations in the specific geographical area. Then the analysis of the sequence of high and low coded observations over time provides the user with a heuristic profile of whether time-trends are being observed If one traces other attributes in a similar fashion, an impression becomes available as to whether or not values are getting higher or lower jointly. This type of comparison of profiles yields perhaps a simpler appearing, but concise overall picture of jointly occurring trends.

2. Application to 20-Year Air Pollution Data

An illustration of site-specific analyses, let us use annual summaries twenty successive years, 1990- 2010) for three primary air pollutants, Carbon Monoxide (CO), Nitrogen Dioxide (NO2) and Ozone (O3) monitored at the by the U.S. Environmental Protection Agency (*Latest Findings on National Air Quality 2000 Status and Trends* 2001). These data are available for public use for a number of cities in the U.S. For the present analyses annual city-level summaries were retrieved for 1990-2010 from www.epa.gov/airtrends/aqtrends.html. on 1/18/2013. Two cities were from Eastern US, the fist with a 2009 population of 126,122, and the second with a 2009 population of 2,690,886. They had complete data on three air contaminants, Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Ozone (O3) for the years 1990-2000. For purposes of the present analyses they will be referred to as rural and urban, respectively. Data were analyzed using correlation and regression-based approaches, presented in section 3, as well as a new method developed by the present author, presented in section 4.

3. Correlation and Regression Based Analyses

Pearson Product-Moment correlations and Spearman-Brown non-parametric rank correlation coefficients were computed between CO, NO2, and O3 annual levels for the two cities. The correlation between CO and O3 levels are high for both the rural and urban cities; Pearson r = 0.44 and 0.66 respectively and similarly Spearman r = .56 and .73. However, NO2 and O3 levels show a high positive correlation for the urban city (Pearson r = .53, Spearman r = .61, but not for the rural city (Pearson r = -.20, Spearman r = -.21). Similar results are obtained for the relationship between CO and O3 levels. Urban Pearson r = 0.78, Spearman r = .75; rural Pearson r = 0.15, Spearman r=- 0.04. This may denote high concentration of multiple pollutants in areas of high population.

The annual concentrations of each pollutant were plotted for the urban and rural area over the 20-years. Results of the plots, along with a linear regression fit to the data are shown in

Figure 1: Urban data are plotted on the right, rural data on the left.

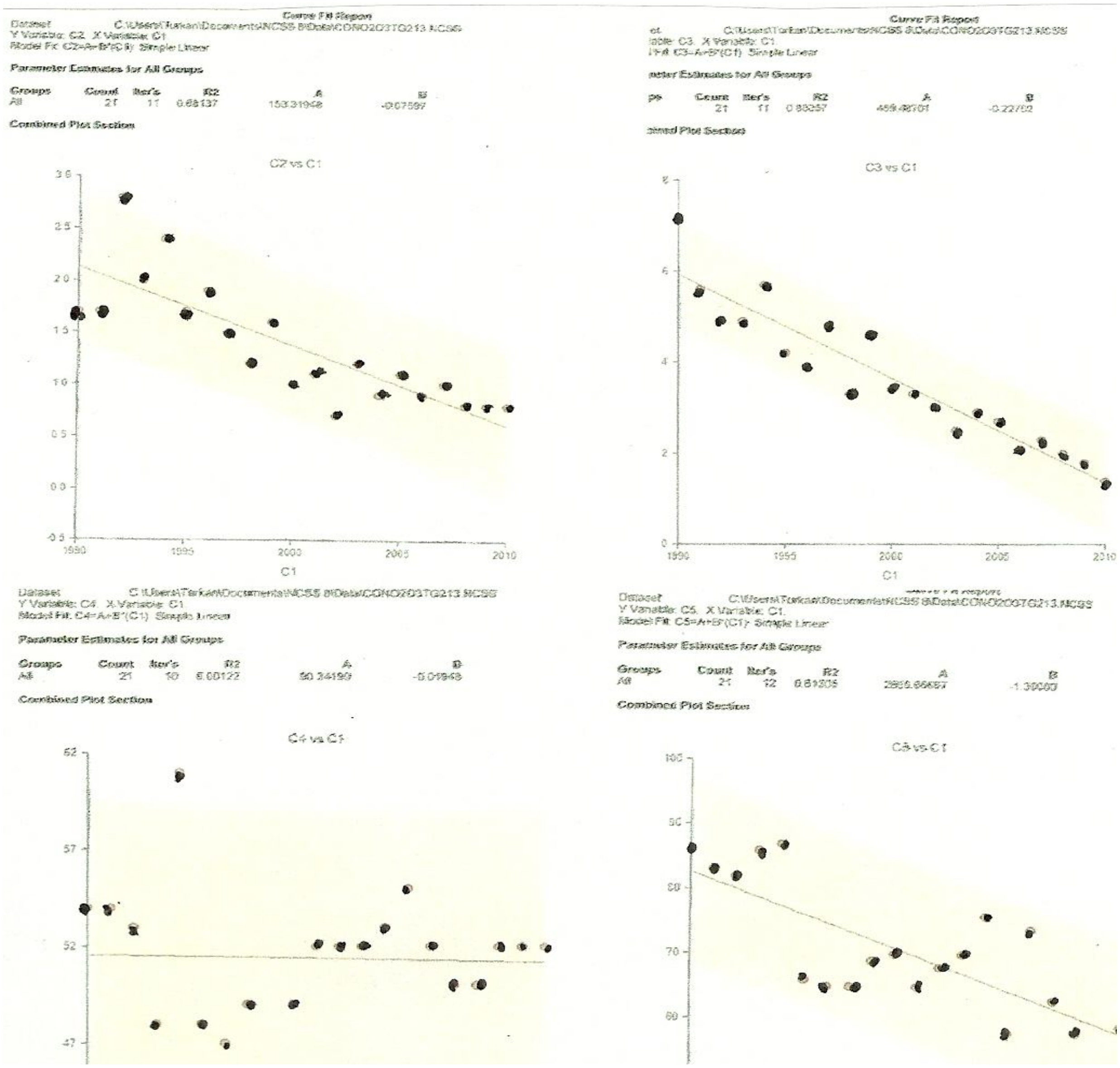


Figure 1. 20-Year Carbon Monoxide (upper two plots) and Nitrogen Dioxide (lower two plots). It is apparent that a downward trend is observed over the 20 years for all except for NO2 rural data. Yet these trends cannot continue indefinitely, because they will be truncated by the regulatory standard. This challenge prompted the necessity for further analytic exploration, the challenge to search for a step-based approach using data available, for switches upward or downward and for relatively when change occurs. The step needs sufficient resolution for identifying change and consistency thereof.

4. An Alternative Approach Using Step-Function Shifts across Monitoring Bands

4.1 Establishing the Monitoring Bands and Application to NO2 Annual Data

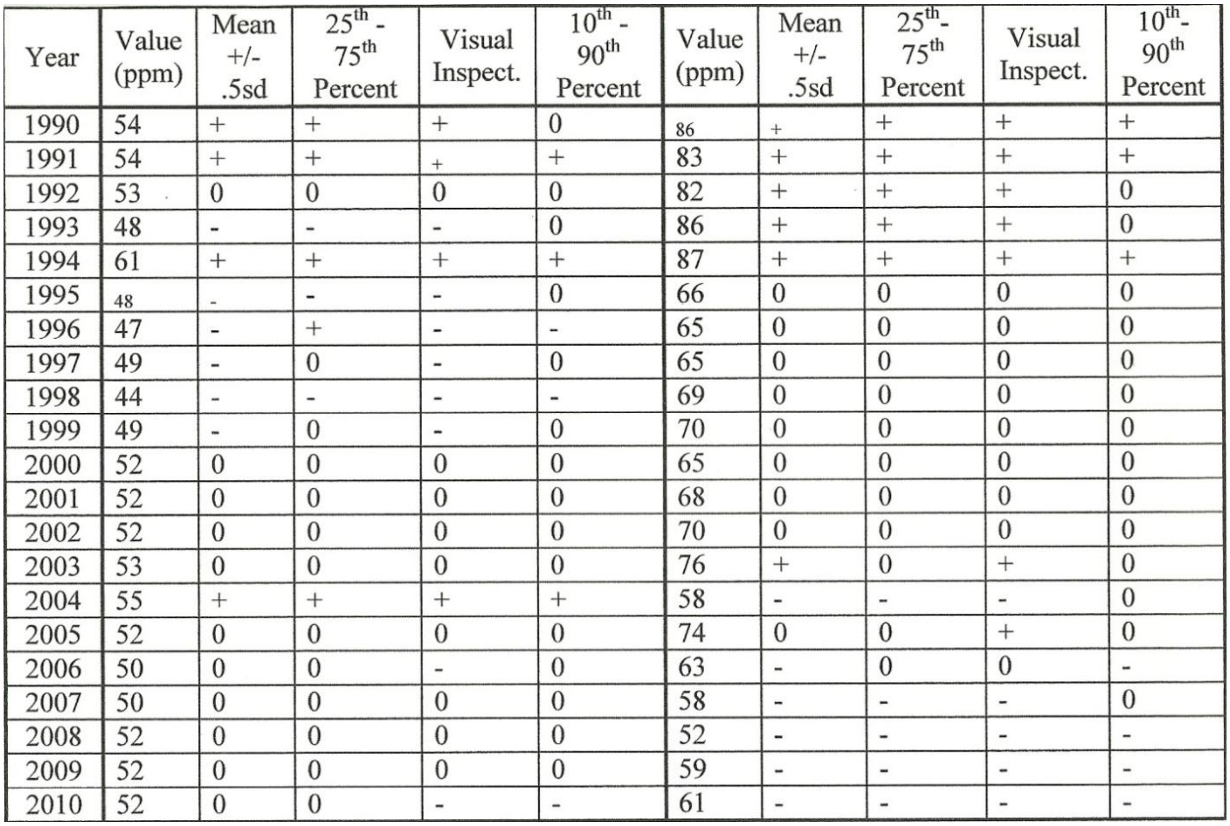
Proceeding along this inquiry, the cumulative distributions and results of tests for Normality were evaluated, confirming Normality assumptions. We then proceeded to determine if it would be feasible to ascertain ranges or bands around the mean, akin to those in quality control (QC) charts through the use of which one can track successive observations as to whether a step-change could be detected during the 20-year time interval. In this effort, four types of alternative ranges or bands were established, as illustrated for the Nitrogen Dioxide data in Table 1.

The first, shown in the left column, uses, as the lower limit, 0.5 standard deviation (½ sd) below then mean, and as the higher limit 0 .5 standard deviation above the mean. The second, shown in the next column to the right, uses the range between the 25th and 75th percentiles, often displayed in Box Plots. The third, shown in the next right column, uses visual inspection of the range in data—averaging the first three and the last three observations and inspecting the data within this range for "natural breaks," as employed in mapping with Geographic Information Science (GIS). The fourth establishes the bands based upon the 10th and 90th percentiles. The data observed are in Columns 2 and 7; the {+/0/-} allocation based on the alternative allocation schemes are indicated as entries in columns 3-6 (for rural) and columns 8-11 (for urban) locations.

Reviewing the monitoring bands under the four alternatives across all three air pollutants showed that using the mean +/-0 .5 standard deviation was preferable to the others in pinpointing and isolating site-specific trend-oriented step shifts. For example, prior to 1995 data for the urban site, data were in the {+} range; then a switch to the {0} range is observed, then another shift to {-} or lower values after another 10 years. For the rural location there is a downward shift after the first five years, but no consistent change thereafter. 90th – 10th percentile and Box-Plot oriented 75th – 50th percentiles resulted in wider bands, thus leaving more observations outside the monitoring band. They pinpointed outliers from grouped data and were not designed to isolate trend –oriented step shifts for the individual locations.

Table 1: Step-Shifts in Trinary (+/0/-) Categories During Successive 20 Years (1990-2010) For Nitrogen Dioxide (NO2) Annual Levels fur a Northeastern Rural and Urban Area

Rural Urban



Year	Value (ppm)	Mean +/- .5sd	25th - 75th Percent	Visual Inspect.	10th - 90th Percent	Value (ppm)	Mean +/- .5sd	25th- 75th Percent	Visual Inspect.	10th- 90th Percent
1990	54	+	+	+	0	86	+	+	+	+
1991	54	+	+	+	+	83	+	+	+	+
1992	53	0	0	0	0	82	+	+	+	0
1993	48	-	-	-	0	86	+	+	+	0
1994	61	+	+	+	+	87	+	+	+	+
1995	48	-	-	-	0	66	0	0	0	0
1996	47	-	+	-	-	65	0	0	0	0
1997	49	-	0	-	0	65	0	0	0	0
1998	44	-	-	-	-	69	0	0	0	0
1999	49	-	0	-	0	70	0	0	0	0
2000	52	0	0	0	0	65	0	0	0	0
2001	52	0	0	0	0	68	0	0	0	0
2002	52	0	0	0	0	70	0	0	0	0
2003	53	0	0	0	0	76	+	0	+	0
2004	55	+	+	+	+	58	-	-	-	0
2005	52	0	0	0	0	74	0	0	+	0
2006	50	0	0	-	0	63	-	0	0	-
2007	50	0	0	0	0	58	-	-	-	0
2008	52	0	0	0	0	52	-	-	-	-
2009	52	0	0	0	0	59	-	-	-	-
2010	52	0	0	-	-	61	-	-	-	-

4.2 Integration with On-Line Trend Reporting

Interactive computerized access to environmental data is widely available, as shown inFigure 2, plots retrieved during 9/29/ 2013 from on-line EPA air quality trends website. The reporting shows, for the group of 89 sites for Co and 30 for NO2, the average for each year, The national standard, and the 80% bracketed between the upper and lower 90th and 10th percentiles. Thus, If an individual is interested in tracing possible exposure in a certain city, knowing the city's relative standing and whether relative increase or decrease appeared during periods of interest will provide a useful supplement to other health-related data.

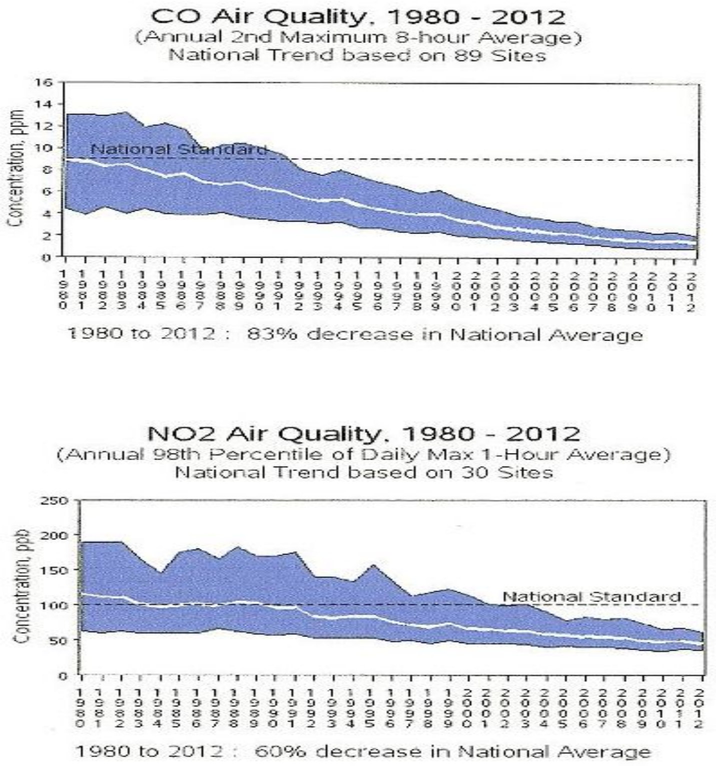


Figure 2.: Air Quality Trends 1982-1992 for Nitrogen Dioxide (NO2) and Carbon Monoxide (CO): National Average, National Standard, 90th-10th Percentiles : 30 Sites for NO2 and 89 Sites for CO

Source: *http:www.epa.gov/airtrends Retrieved*: 9/29/2013

5. *Discussion*

Einstein is said to have once remarked that not everything that can be counted counts and not everything that counts can be counted. In trend estimation, particularly if estimates are to be used in forecasting, it is important to be note step changes in data, as demonstrated by the application of monitoring bands in Table 1. A regression line, linear or non-linear, implies that the function will continue beyond the present. Bands of this type do not assume steady state and visually smooth uncertainties implicit in the data elements, as stated in the introductory section of this paper. Observing successive data within a specific band over time indicates stability prior to a switch downward or upward. The stock bands of Energy Information Administration (Hall, D., Gardenier, T., Slavic A. 1984) included seasonality adjustment as well (Durbin, J. and Koopmans, S. J. 2001). Tracing status and change in a 3-category or Trinary framework provides a succinct tabular representation of when change is elicited. In this paper, we have provided guidelines for what level of 3-way categorization is most likely to identify emerging trends in data and concluded that use of the mean +/- .5 standard deviation appeared to yield the best detection.

Patient engagement has taken an important role in personalized medicine (Bertakis, K. D., and Azari, R. 2011; Epstein, R. M., Franks, C. G. et al 2005). Not only are patients are encouraged to participate in their own care, but also programs such as *Patients-Like-Me* seek motivated patients to connect, research and integrate relevant data, much of which is available on-line. Reference guidelines for interpreting data are crucial for such evaluations. Furthermore, individuals are likely to inquire about the possible environmental exposures from where they resided or worked. Not only data for geographical locations, as addressed by Dangermond (2012), but also time-series data for individual locations, along with any summarized interpretation provided at the time of data retrieval, are essential. Individuals are likely to remember the city where they lived or worked; it is important to integrate city-level time trends with information on the national standard, national summary statistics and indices such as trinary {+/0/-} bands based on mean +/- .5 standard deviation, as discussed in this paper. In parallel, access to ongoing research as presented by Christensen and White (2011) will provide additional references.

The approach presented here is relevant not only to health and environmental data. In a report relating to climate change Washington (2011) reported annual global mean temperature 1880-2010 not only in the form of a line graph but also as a deviation from the overall mean during 1901-2000 (shown as a 0.00 reference point). It was clear from this perspective how temperature readings were consistently below the mean during earlier years and started rising with time. In this paper a similar approach is being advocated, using location-specific successive readings for each geographic location.

During analyses this trend in thought also indicated a theoretical query which needs to be explored further. Attempting to apply a test of significance to the sequence of +1, 0, or -1 coded successive observations, use of the non-parametric Runs Test as described in Freedman, Pisani and Purves (2007), Netter, Wasserman and Whitmore (1978, p. 194-198) was considered. The number of switches in status, denoted by runs up and down, is used in the test statistic. If the observation is higher than the preceding observation, it is coded +, if it is lower, it is coded as -, and if two successive observation are tied, the code of the preceding evaluation prevails. Using a buffer zone as in the present case eliminates ties, because each successive observation belongs into one of 3 categories or bands: +1, 0 or -1. It is the succession of the codes in the three options which determines either maintenance in each step or switch from one step into another. Thus the non-parametric Runs Test was not applicable to the 3-category delineation. The test statistic is based upon a 2-category delineation for evaluating successive observations, thus pointing to the necessity for further methodological research. CUSUMS(Cumulative Sums) and MOSUMS (Moving Sums) where the sums of successive observations are used to detect slowly emerging trends over time; within the context of our queries early detection is of utmost concern and importance.

In conclusion, this paper demonstrates how 3-level categorizations of environmental exposure data (and potentially other measurement data as well) can simplify and clarify health-related exposure histories. The analytic approach has broad applications in personalized medicine, epidemiology, and public health policy.

Acknowledgement

I wish to thank John S. Gardenier for editorial comments and for encouraging the illustration and application of the 3-category index which I developed.

References

Bertakis, K. D. and Azari, R. (2011), “Patient Centered Care Is Associated with Decreased Health Care Utilization,” *Journal of the American Board of Family Medicine*, 24 (3), 229-239.

Christensen, K. L. Y. and White, P. (2011), “A Methodological Approach to Assessing the Health Impact of Environmental Chemical Mixtures: PCBs and Hypertension in the National Health and Nutrition Examination Survey,” *International Journal of Environmental Research and Public Health,* 8, 4220-4237, *doi:10.3390/ijerph8114220.*

Dangermond, J., “Geography: A Platform for Understanding” (2012), *Arc News*, 34 (3), 1.

Durbin, J, and Koopmans, S.J. (2001), *Time-Series Analysis by State Space Methods,* Oxford, England: Oxford University Press.

Epstein, R. M., Franks, C. G., Shields, S. C. et al (2005), “Patient-Centered Communication and

Diagnostic Testing,” *Annals of Family Medicine*, 3 (5), 415-421.

Freedman, D., Pisani, R. and Purves R. (2007), *Statistics,* New York, NY: Norton.

Gardenier, T. K. and Gardenier, J.S. (2013), “Delving into Megadata: Evolving Challenges,” *Proceedings of Statistical Learning and Data Mining Section, 2013 Allied Statistical Meetings,* Montreal, CA.

Gardenier, T.K. (2011), “Personalized Medicine Featured in JSM and AAAS,” *AMSTAT News http//magazine.amstat.org/?cat*=17, 37.

Gardenier, T.K. (2005), *Trinomial Neural Networks for GIS Analysis,* Bethesda, MD.

Hall, D., Gardenier, T. K., and Slavic, A. (1984), *Intervention Adjustment of Data of the Joint Petroleum Reporting System.* Final Report of U.S. Department of Energy Contract *DE-AC06-76RLO-1830.* Richland, WA: Battelle Pacific Northwest Labs.

Netter, J., Wasserman, W. and Whitmore, G. A. (1978), *Applied Statistics (2nd ed.),* Boston, MA: Allyn and Bacon.

Pickle, L. W., Mugniole, M., Jones, G. K. and White, A.A. (1996), *Atlas of United States Mortality.* DHHS Publication PHS97-1915. Hyattsville, MD: National Center for Health Statistics.

U.S. Environmental Protection Agency (2001) , *Latest Findings on National Air Quality 2000 Status and Trends.* EPA 454/K-01-002 .Washington, D.C.