Delving into Megadata: Evolving Challenges

Turkan K. Gardenier[1] and John S. Gardenier[2]

[1] Pragmatica Corp. 115 Saint Andrews Drive NE, Vienna, VA 22180
[2] National Center for Health Statistics (Retired), 115 Saint Andrews Drive NE. Vienna, VA 22180

Abstract

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, during this process of delving into such "megadata" exploring linkages among multiple variables which relate to health and environment raise challenges. For example, analog data which are summarized into daily then to monthly measurements, averages being used to display one attribute versus ranks for another, unequal time intervals between successive data. Integrating displays from Greographic Information Science (GIS) oriented maps with tabular summary data also generate further challenges. Resolution-related issues which need to be recognized and implemented during the process of fine-tuning inferences and conclusions are addressed, along with a new approach using 3-category step-based approach.

Key Words: Megadata, interactive retrieval, Geographic Information Science (GIS), fuzziness in data, 3-category allocation

1. New Vistas Associated with Data Increase

Computer technology has taken many strides of advancement from the times of fixed diskettes for storage and access to virtual clouds. As databases increase in complexity and magnitude, their efficient use demands inputs from statistical "signaling" technology. In conjunction with working with contents of megadatasets as a maze through which we need to trace a path the following issues relating to data uncertainty need to be recognized:

1.1 Non-Uniformity in Compilation and Merging

For environmentally-oriented databases such as air or water quality data compilation procedures differ. Thus a direct association between a specific geographical location and the magnitude of pollution level is not precise. For example, air quality measurements are obtained for monitoring stations which are often distant from the specific place of interest to the researcher who is trying to analyze exposure level for a specific individual. Data may be analog, later to be discretized into per-hour or per-day basis creating a megadataset, later to enter into monthly or annual averages. Sometimes average values over different lengths of time, or peak values within these times are used. Water quality measurements are compiled into geographical water basin records. Associating data from different record types, although necessary and worthy of merit, introduces uncertainties which demand recognition.

1.2 Data Presence Issues

Not all data for each individual record may be available as records for multiple clinical trials, for example, are merged into a large database. For a specific geographical location morbidity and mortality data may be available from hospital records; but privacy issues may prevent disclosure of identity. For many environment- and health-oriented databases long-term data are retrievable and available to the public (Pickle, L. W.,Mugniole, M, et. al 1996; U. S. Environmental Protection Agency 2001).

2. Various Ways to Approach Megadata

2.1 Parallels between Location (GIS) based and Individual-based Records

In map-based displays, as illustrated in Figure 1, the legend below each map uses colors with 10 intervals ranging from highest (red) to lowest (dark blue). The shades get lighter and lighter and fade into white near the average for each group. The national averages differ, yet each group provides its own reference, which then gets trensferr3e to color coding. A similar approach can be useful with continuous data even without mapping and is illustrated below.

2.1.1 Database Used to Explore Time-Varying Trends in Environmental Data

Striving to capture time-varying patterns in environmental data, annual air pollution levels for each of 20 successive years were retrieved from the Air Quality Trends by City database of the Environmental Protection Agency.

2.1.2 Subsets and Time Intervals

The data were for 1990-2010 for a relatively small city (population of 126,000 in 2009) and a large city (population of 2,691,000 in 2009). The annual data were in parts per million (ppm) based on hourly measurements for Carbon Monoxide (CO) and Nitrogen Dioxide (NO2). The data which were submitted to analyses are in Table 1.

2.2 Forecasting Oriented Approach

Figure 2 shows graphical displays relating to plots over time: the top section relates for CO with data for each year and variation around the regression line with best fit; low population city on the left, city with high population on the right. Both show relatively similar slope, indicating that air quality is improving. The lower section of Figure 2 is for NO2, small city with practically no slope, larger city with negative slope. Yet, it is difficult to assume that thew observed trends will continue at the same rate. The regulatory mandate implies a lower bound established by air quality standards.

The cumulative distribution (CFD) plots revealed that the straight line fit was quite good near the center, but not elsewhere. Autocorrelation computations showed a decrease until a lag of 5 or 6, after which negative autocorrelations occurred. Yet the conclusion that a cyclical pattern exists cannot be reached because 20 successive values are too few in order to ascertain model stability.

2.3 Step-Based Approach

The challenge relating to the fact that the observed trend may not continue led to a search for a step-based approach, for switches upward or downward and for when relatively stable change occurs. The step itself became a challenge, for it needed sufficient resolution to identify change and the consistency thereof. Exploring various criteria included ways to create "bands" akin to control limits in Quality Control (QC) charts. The bands should not be so wide as to exclude only outliers, but sufficiently wide to detect earliest consistent decrease over time. Mean +/1 sd, upper and lower quartiles, deciles, etc were used, finding the most preferable to be mean +/- one half (.5) sd, leaving approximately one third of the data around the mean, and one-third at each end of the distribution.

While Table 1 merely lists the raw data, Figure 3 shows the apportionment of the data using the three-category (+/0/-) allotment of data elements using the criteria of allocation of mean plus (+) or minus (-) one half (½) or .5 sd. In darker lines, the first observed relatively stable shift to decreased pollution levels for CO and NO2 are marked.

2.3.1 Prospects for Personalized Medicine

Let us assume that an individual is interested in tracing possible exposure to contaminants in a specific area. Knowing the city's relative level and whether a relative increase or decrease was observed during periods of interest will be a useful supplement Let to other health-related data. The importance of patient-cerntered communication and its advantages have been addressed by Epstein R.M. et al (2005) and Bertakis, R.K. and Azari R (2011).

2.3.2 Wide-Reaching Advantages

The example demonstrated illustrates several features of the generic methodology being advocated:

(a) It is an exploratory way of looking at data regardless of the size/scale of the data sets;

(b) It is designed to offer insights to the researchers involved and perhaps tothe funders/clients as well;

(c) It is not designed to provide precise detailed analyses, but rather to help understand the nature of the data set(s) so that they can theu design the best methodological path(s) t o meet various study purposes;

(d) It focuses on the upper and lower regions of each data set, leaving out the fuzzy middle region;

(e) It is not a single cast-in-stone approach Judgment and subject matter expertise will adapt this basic concept for each problem at hand including, in the case of immeasurably large data sets, simple random sampling of one data point out of a million (or some other factor) may improve the tractability of the set while preserving all of the interactions relevant to the exploratory analysis;

(f) The approach is not limited to health, environment or social issues. It may be useful in astrostatistics, particle physics, optics, chemistry or any other applications.

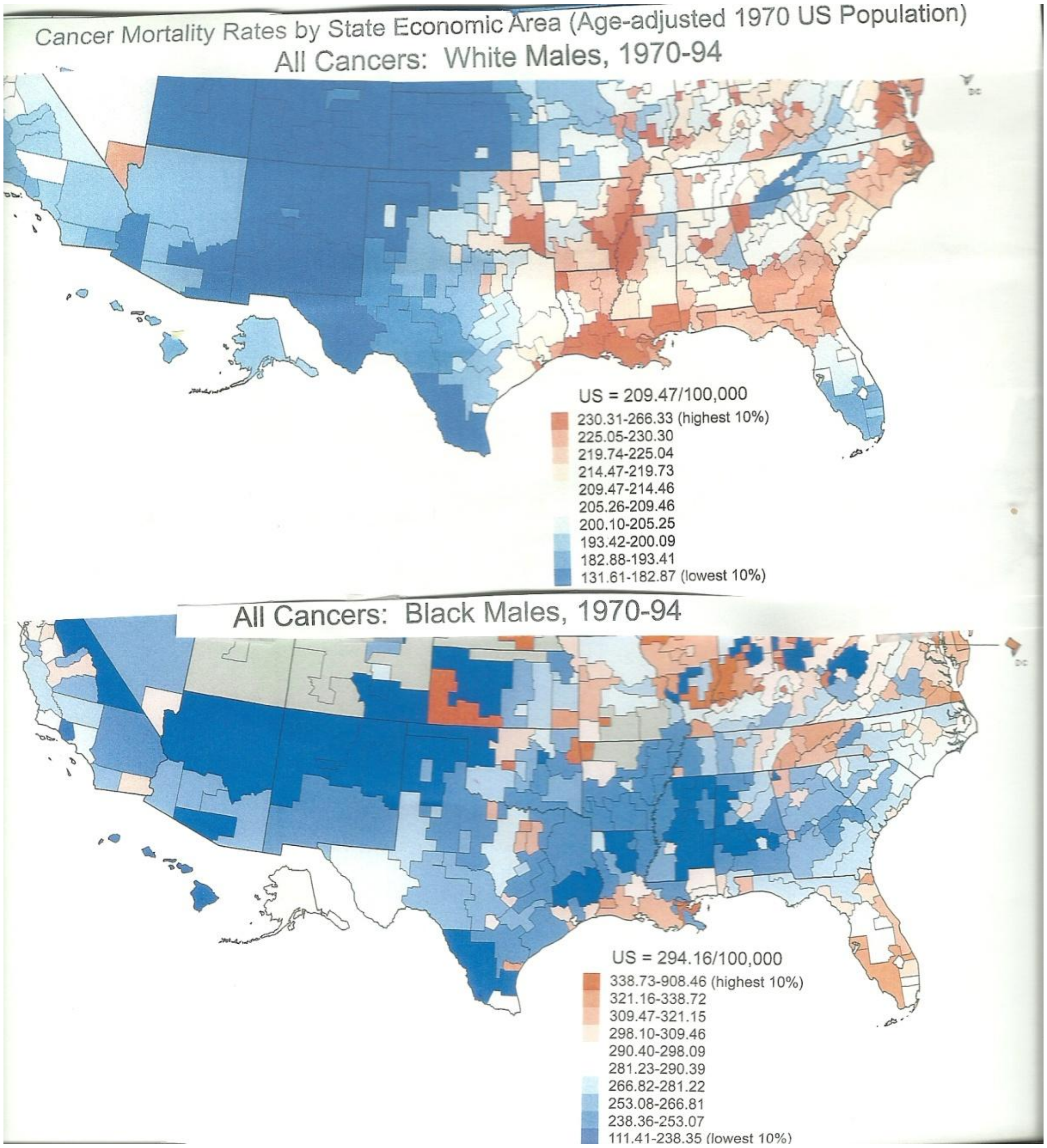
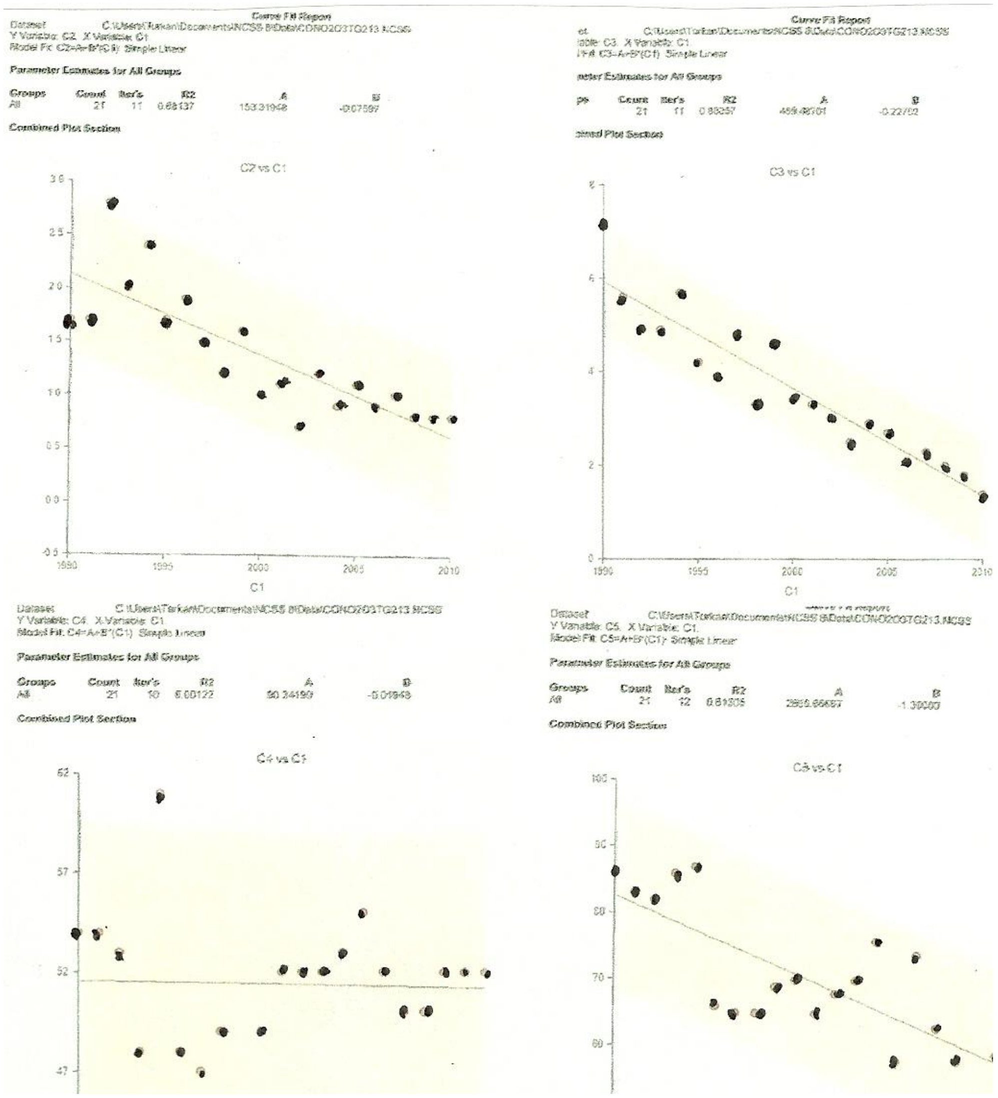

Cancer Mortality Rates by State Economic Area (Age-adjusted 1970 US Population)
All Cancers: White Males, 1970-94
US = 209.47/100,000
230.31-266.33 (highest 10%)
225.05-230.30
219.74-225.04
214.47-219.73
209.47-214.46
205.26-209.46
200.10-205.25
193.42-200.09
182.88-193.41
131.61-182.87 (lowest 10%)
All Cancers: Black Males, 1970-94
US = 294.16/100,000
338.73-908.46 (highest 10%)
321.16-338.72
309.47-321.15
298.10-309.46
290.40-298.09
281.23-290.39
266.82-281.22
253.08-266.81
238.36-253.07
111.41-238.35 (lowest 10%)

Figure 1: Geographic Information Science (GIS) based comparison of cancer death rates over two decades Source Devesa S. S., 1999 *Atlas of Cancer Mortality in the United States 1950-1994* NIH Publication 99-4564 (National Institutes of Health, National Cancer Institute, Bethesda, MD.



```
                                   Curve Fit Report
Dataset          C:\Users\Turkan\Documents\NCSS 8\Data\CONO2O3TG213.NCSS
Y Variable: C2  X Variable: C1
Model Fit: C2=A+B*(C1)  Simple Linear

Parameter Estimates for All Groups

Groups      Count  Iter's       R2            A           B
All            21      11  0.68137    153.31948    -0.07597

Combined Plot Section
```

```
                                   Curve Fit Report
et               C:\Users\Turkan\Documents\NCSS 8\Data\CONO2O3TG213.NCSS
iable: C3  X Variable: C1
l Fit: C3=A+B*(C1)  Simple Linear

neter Estimates for All Groups

ps          Count  Iter's       R2            A           B
               21      11  0.88257    459.48701    -0.22752

bined Plot Section
```

```
Dataset          C:\Users\Turkan\Documents\NCSS 8\Data\CONO2O3TG213.NCSS
Y Variable: C4  X Variable: C1
Model Fit: C4=A+B*(C1)  Simple Linear

Parameter Estimates for All Groups

Groups      Count  Iter's       R2            A           B
All            21      10  0.00122     90.34190    -0.01948

Combined Plot Section
```

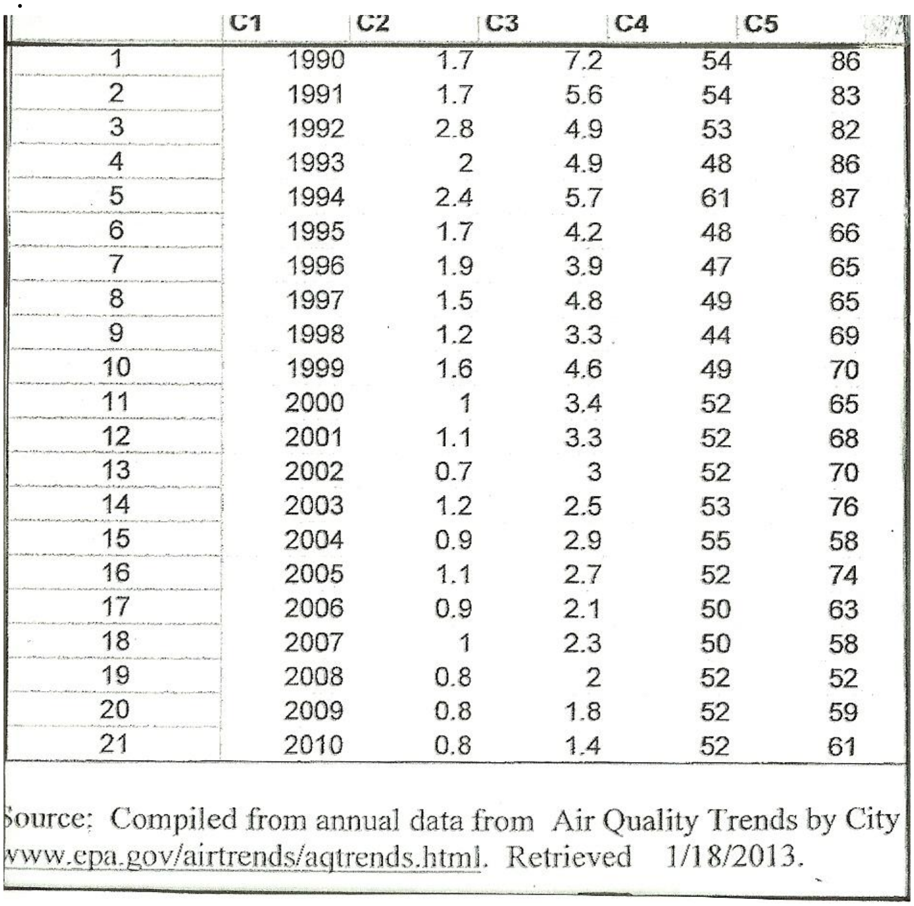
```
                                   Curve Fit Report
Dataset          C:\Users\Turkan\Documents\NCSS 8\Data\CONO2O3TG213.NCSS
Y Variable: C5  X Variable: C1
Model Fit: C5=A+B*(C1)  Simple Linear

Parameter Estimates for All Groups

Groups      Count  Iter's       R2            A           B
All            21      12  0.61305   2859.66667    -1.39000

Combined Plot Section
```

Figure 2. 20-year Carbon Monoxiude (upper two figures) and Nitrogen Dioxide Levels (lower two figures) for cities with low population (on left) and high population (on right)

Table 1. Data Retrieved for Analysis of 20-Year Changes in Carbon Monoxide (CO) and Nitrogen Dioxide (NO2) in Two Cities

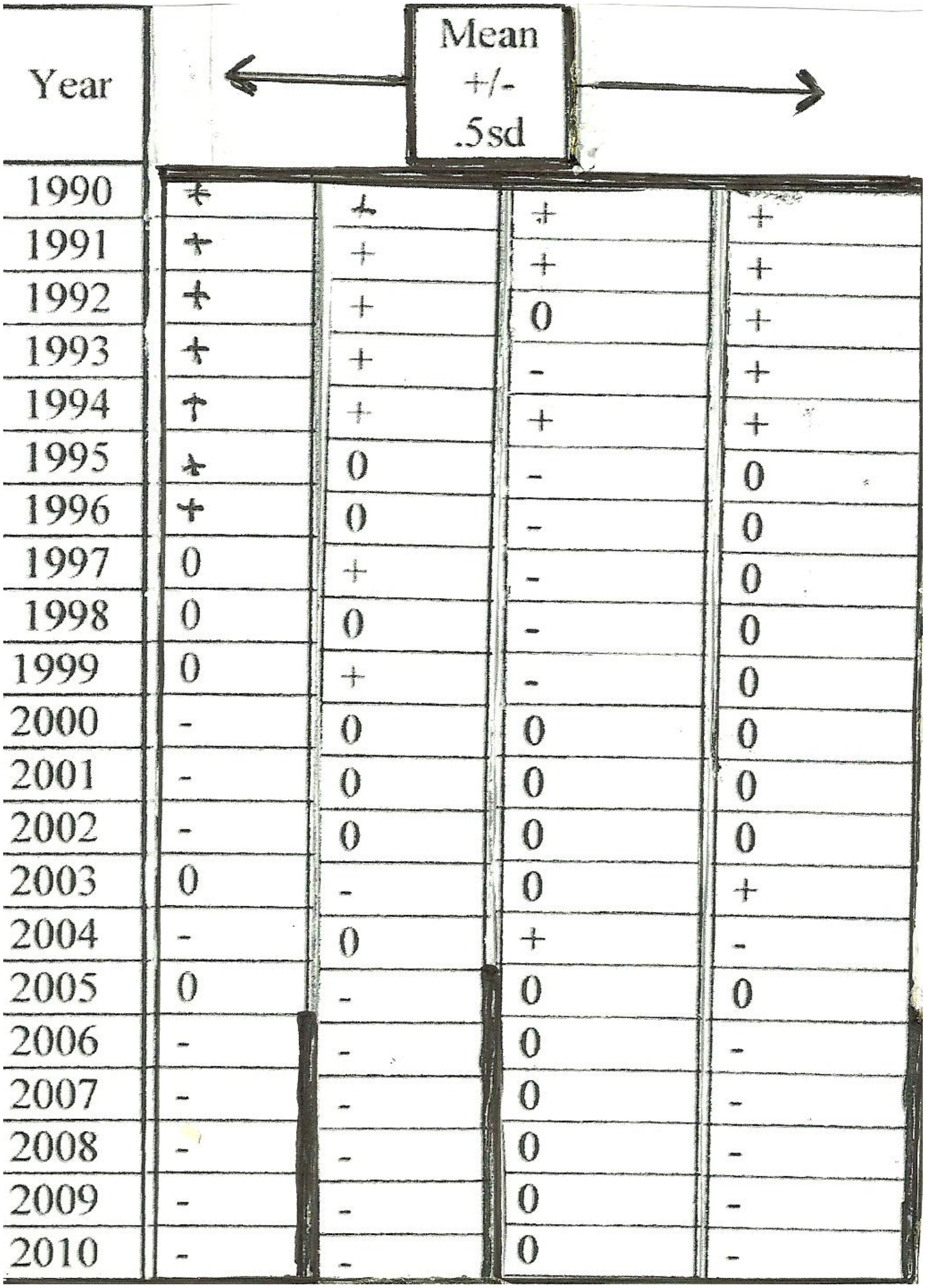
C1: Year; C2: Carbon Monoxide (Low Population City);
C3: “ “ (High Population City)
C4NitrogenDioxide(LowPopulationCity)
C5 “ “ (High Population City)

:



	C1	C2	C3	C4	C5
1	1990	1.7	7.2	54	86
2	1991	1.7	5.6	54	83
3	1992	2.8	4.9	53	82
4	1993	2	4.9	48	86
5	1994	2.4	5.7	61	87
6	1995	1.7	4.2	48	66
7	1996	1.9	3.9	47	65
8	1997	1.5	4.8	49	65
9	1998	1.2	3.3	44	69
10	1999	1.6	4.6	49	70
11	2000	1	3.4	52	65
12	2001	1.1	3.3	52	68
13	2002	0.7	3	52	70
14	2003	1.2	2.5	53	76
15	2004	0.9	2.9	55	58
16	2005	1.1	2.7	52	74
17	2006	0.9	2.1	50	63
18	2007	1	2.3	50	58
19	2008	0.8	2	52	52
20	2009	0.8	1.8	52	59
21	2010	0.8	1.4	52	61

Source: Compiled from annual data from Air Quality Trends by City
www.epa.gov/airtrends/aqtrends.html. Retrieved 1/18/2013.



Year	Mean +/- .5sd			
1990	+	+	+	+
1991	+	+	+	+
1992	+	+	0	+
1993	+	+	-	+
1994	+	+	+	+
1995	+	0	-	0
1996	+	0	-	0
1997	0	+	-	0
1998	0	0	-	0
1999	0	+	-	0
2000	-	0	0	0
2001	-	0	0	0
2002	-	0	0	0
2003	0	-	0	+
2004	-	0	+	-
2005	0	-	0	0
2006	-	-	0	-
2007	-	-	0	-
2008	-	-	0	-
2009	-	-	0	-
2010	-	-	0	-

Figure 3. 20-year air pollution data for Carbon Monoxide (2nd and 3rd columns) and Nitrogen Dioxide (4th and 5th columns) sub-categorized into 3 categories (+/0/-). 2nd and 4th columns refer to small city and 3rd and 5th columns to large city.

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.

Devesa, S, S, et al (1999) *Atlas of Cancer Mortality in the United States 1950-94. NIH Publication 99-4564*. Bethesda, MD. National Institutes of Health, National Cancer Institute.

Epstein, R. M., Franks, C. G., Shields, S. C..et al, "Patient-Centered Communication and Diagnostic Testing," *Annals of Family Medicine,* 3(5), 415-421.

Pickle, L. W., Mugniole, M., Jones, G. K. And White, A. A. (1996). *Atlas of United States Mortaliuty:* 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.