Core/Non-Core Configuration within Big Data Analytics

Turkan K. Gardenier[1], John S. Gardenier[2]
Pragmatica Corp., 1733 Kirby Rd., Unit 1408, McLean, VA 22101
National Center for Health Statistics (Retired) 1733 Kirby Rd., Unit 1408, McLean, VA

Abstract

The authors expand on a Trinary or Trinomial-based approach presented in two preceding annual meetings of the Joint Statistical Meetings, tailoring the procedure to data streaming into Big Data. Core is defined as the central (0) region, with adjoining non-core (+) or (-) regions. Analytic issues related to streaming data, as contrasted with data analyzed as a whole, are summarized. Characteristicss of Trinary (+/0/-) three-category approach are presented and contrasted with two-category or Binomial formulations. Two applications are included (a) air quality monitoring for Nitrogen Dioxide spanning 20 years (1991-2010) in two cities in the US and (b) lung cancer mortality data.

Key Words: Trinary; Trinomial; Binomial; step-shifts; streaming ; Big-Data

1. Configuring Analytics for Records Streaming into Big Data

Gardenier and Gardenier (2014, 2013) discussed challenges related to recent increase in the size of databases, including non-uniformity in compilation and merging. Kannath (2009) distinguished between streaming data into Big Data from data analyzed as a whole. Into Big Data enter records from various sources with different sampling frames and different periodicities. Kannath (2013) referred to such data as semi-infinite time series, where the experimenter does not have a specific single population from which to sample. In applications of the Normal or Gaussian distribution and estimates of central tendency and variability, observations outside "x" number of standard deviations are often treated as outliers. In applications to streaming data the assumptions that the measure of central tendency has a certain degree of accuracy and is consistent over time are usually violated. There are huge numbers of applications of the method we are advocating in environmental analysis, genomics, energy, public health and education among others.

Sayed (2003) stressed the need for adaptive filters. Howe and Balazinska (2013) discussed the challenge for scalable data processing, the use of a reducer output cache as we interact with multiple, perhaps hundreds of computers in the cloud. Our contributions to streaming Big Data, as described in the present report, center upon (a) the use of Trinary (3-category) interval estimates in preference to point estimates of central tendency and (b) application of Trinary based search for identifying conjoint record identification which provides scalability in data processing which, in turn, leads to efficient management of vast databases, and (c) advocating time-varying random sampling schemes using the Trinomial distribution for checking observational stability.

1.1 Core/Non-Core Orientation

Figure 1 includes two representations from results of playing darts (Tibshirani, 2011 (a)). The option displayed at the top shows more accuracy in dart throwing because the darts landed nearer the central region, or Core. Viewed horizontally, the deviations from Core may be minus (-) if located to the left, and plus (+) if located to the right of Core. In essence, three ranges are depicted for the data: minus (-), Core (0),and (+).

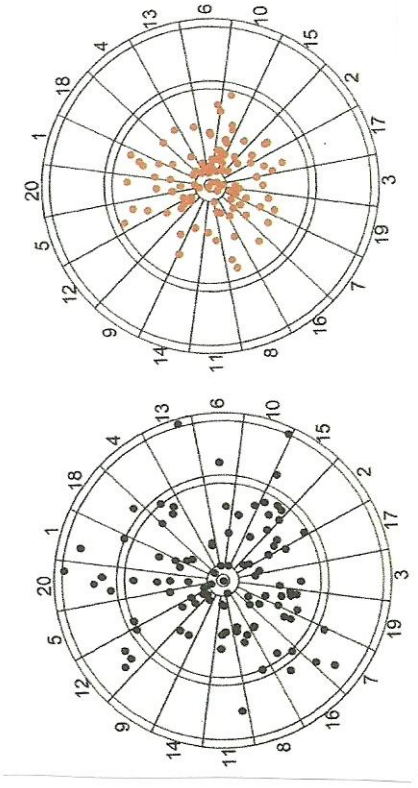


Figure 1: Illustration of Core as central region and Non-Core as regions outside of Core
Source: Tibshirani (2011) (a) p. 48.

1.2 A Historical Perspective

In the seventeenth century Antoine Gombaud, who assumed the name Chevalier de Mere, because of his innate curiosity typical of noblemen of that time, posed the following query to mathematician Pierre de Fermat: "If two players of equal competence are playing a game, and the game ends as a win if either player scores twice, what is the probability that the game will end in two (2) tries? " This initiated the development of the Binomial expansion.

Having no access to modern telecommunication systems, Fermat, in turn, wrote to his friend Blaise Pascal. The correspondence of Pascal and Fermat, whose pictures are shown in Figure 2, led to Pascal's Triangle as illustrated in Figure 3.

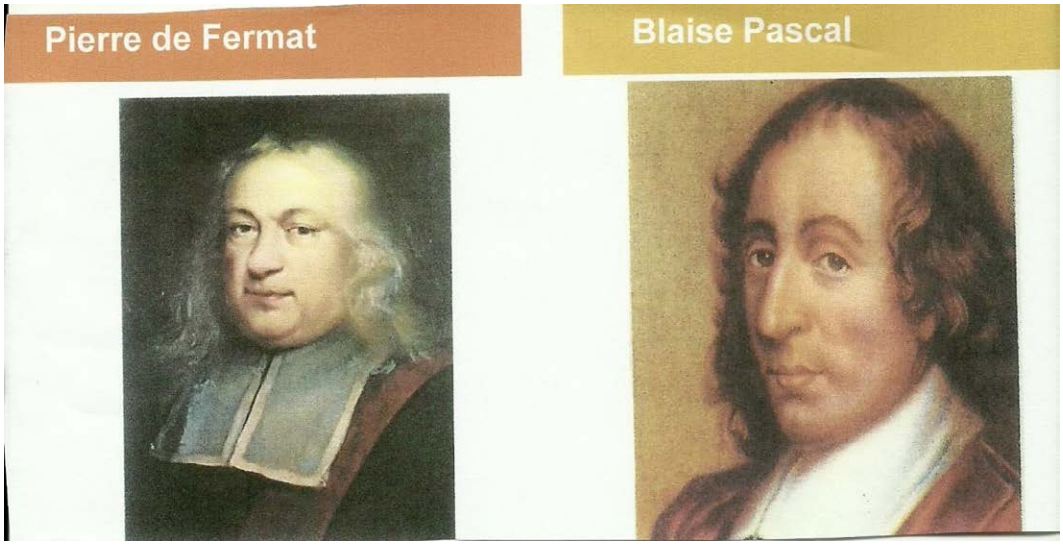


Figure 2: Pierre de Fermat and Blaise Pascal who originated Pascal's Triangle
Source: Wikipedia, 2014.

The assumptions are that (1) there are two possibilities (correct and not correct or Right or Wrong) and, in this case, two successive tries. The process can be expanded to more than two (2) tries as a branching process.

In this case, for either player, if the game is to end as a "win" in two successive Rights, or +'s, the probability is 0 .25 or ¼, because there are 4 possibilities: ++, +-, -+ and - -. The coefficients of the Binomial expansion are included in Pascal's Triangle.

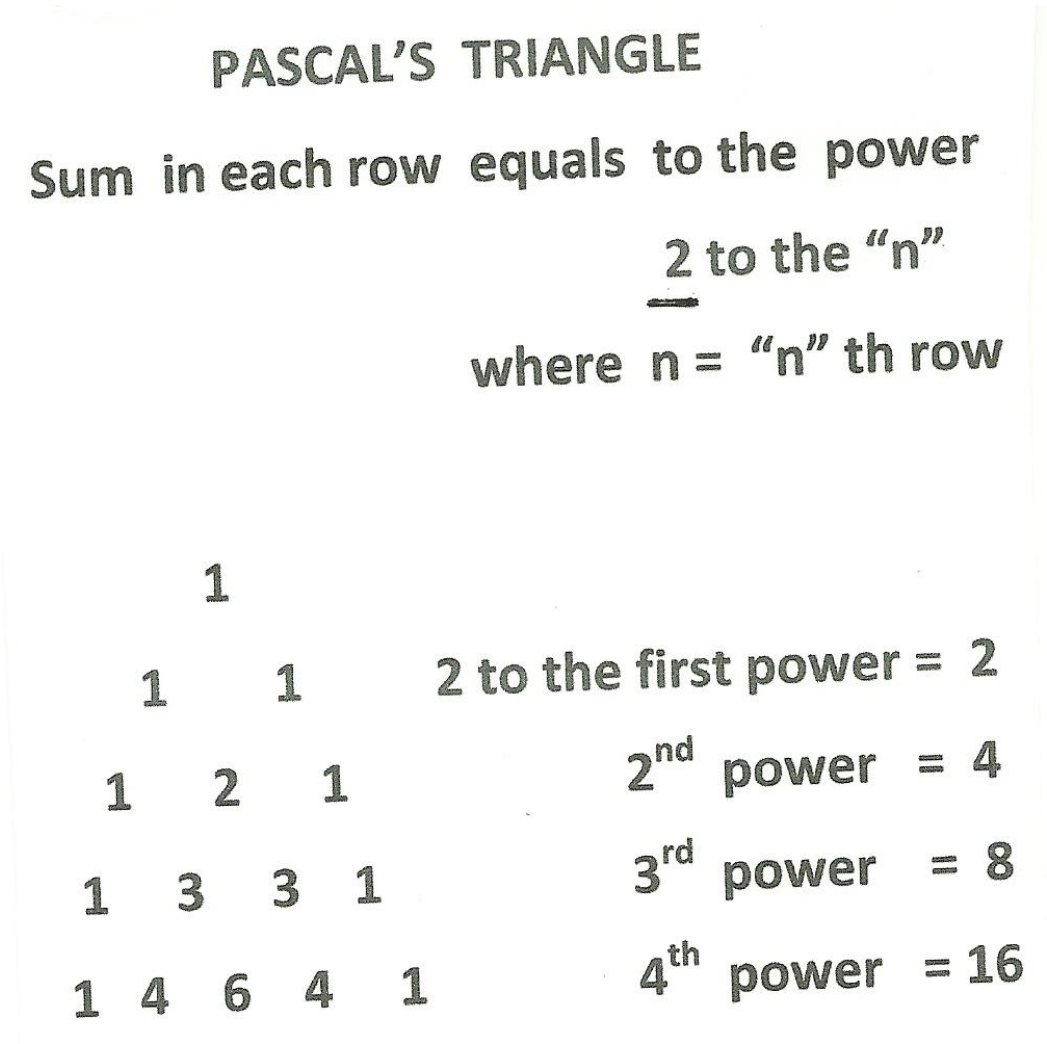


Figure 3: Illustration of entries in Pascal's Triangle

An important issue is that the players have equal competence, i.e., ½ probability of being Right. The histogram created from Pascal's Triangle, as shown in Figure 4, is therefore symmetrical. As the number of tries increases, the rectangles shown in Figure 4 become narrower and narrower, thus transforming the Binomial-based rectangles into a continuous distribution. They approximate the Normal, often called "standard Normal" distribution. If the two probabilities were unequal, the distribution would be skewed.

.

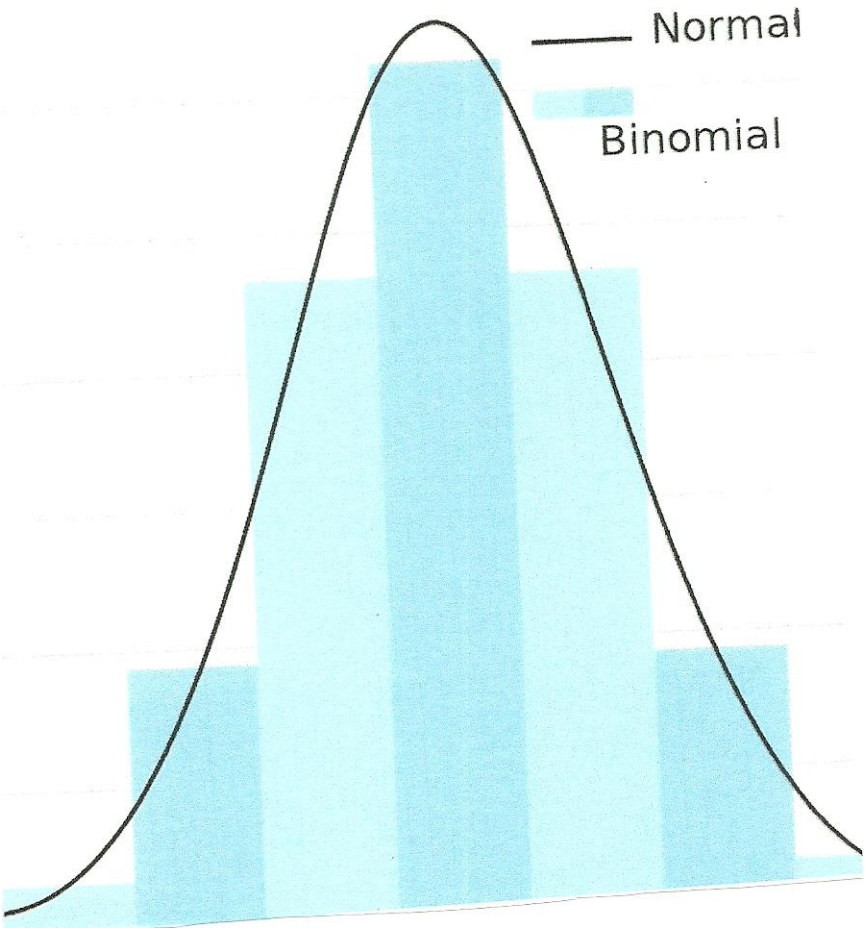


Figure 4: Normal Distribution approximation to Binomial p=1/2
Source: Google retrieval for "Normal distribution"

It is important to reiterate that the distribution is created from two players with two outcomes (+ and – or Right and Wrong) and the probability of + and – is equal. Yet, Sandholm (2015, p. 123) stressed that "games with more than two players. . . deserve further attention." Sheng et al (2015) in the same issue of Science entitled "Popping into the Third Dimension" show three-dimensional architectures for nanomaterials.

1.3 Third Dimension Introduced by the Trinomial or Trinary Orieentation

The Trinomial embodies a third, additional category, the Core. While Pascal's Triangle is based on Binomial coefficients, Pascal's Tetrahedron, the layers of which are shown in Figure 5 separated by solid lines show Trinomial coefficients. Between each line, which depicts a layer, is a triangle. The sum in each layer equals 3 to the "n"th power. For example, 3 to the 2^{nd} power is 9, 3 to the 3^{rd}

power is 27. Within each layer the edges are associated with Pascal's rows. For example, for 2 to the 3rd power, 8, the coefficients in Pascal's Triangle are 1, 3, 3, 1 which are also edges for layer 3 of Pascal's Tetrahedron. These are mathematically as well as visually intriguing.

The Trinomial is part of the d-nomial or multi-nomial distributions. Pascal generated Pascal's Tetrahedron as well as his 2-category based Triangle. We are proposing the use of three categories, Core, Non-core minus (-) and Non-core plus (+) in modeling for Big Data analytics. This approach is advantageous because the Core provides a central region or interval for focusing, while the Non-core (+) and (-) regions provide a glimpse of the range in data.

The Binomial, with two possibilities, uses to compute N things taken *m* at a time, the formula: N factorial divided by (N-*m*) factorial. In the Trinomial, partitioning into three categories, we would be using (N – *m*) factorial, (*m* – *k*) factorial and *k* factorial. Thus we would have an additional term *k* for further partitioning and (*m* – *k)*. A third option depicted in three categories, Core, Non core (+) and Non core (–) are more natural partitions for Big Data based streaming data.

Now the query is: How do we apply this concept? Stagewise sampling is one way. Use past results about Core, and Non core (+) and (–) to use as baseline for different scenarios. Then, while data are streaming, select N number of successive samples randomly and over time. Then compare observed with expected as deviations from the past observations or across different scenarios.

2. Applications to Univariate Monitoring and Bivariate Mapping

2.1 Air Quality Monitoring Spanning 20 Years

This application is described in Gardenier and Gardenier (2013) and Gardenier 2014, 2013, 1984). Data related to 20-year annual Nitrogen Dioxide levels spanning the years 1991-2010 in two US cities, one with low population, and another densely populated. Using 20-year annual summaries for each city, data for each year were coded as zero core (0), minus (-) or plus (+).

As shown in Figure 6, we detected different patterns in step-based shifts in onset and stability over time. For the city with high population a shift to a lower step was observed in 1996, stability until 2003, followed by another shift to a lower step in 2006 continuing through 2010. This was not apparent for the city with low population. Thus our approach displayed the profile of relative differences over time.

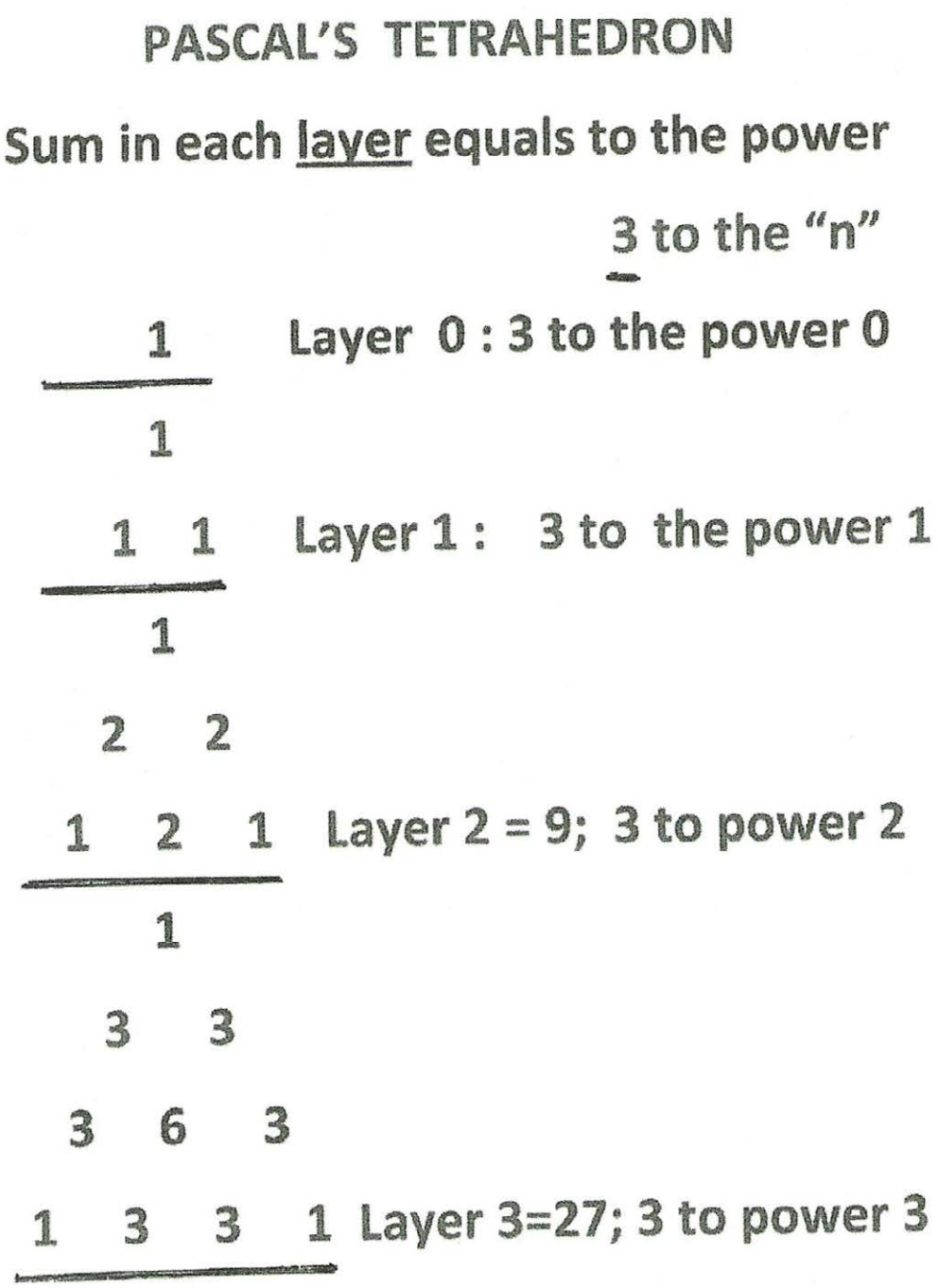


Figure 5: Illustration of Pascal's Tetrahedron

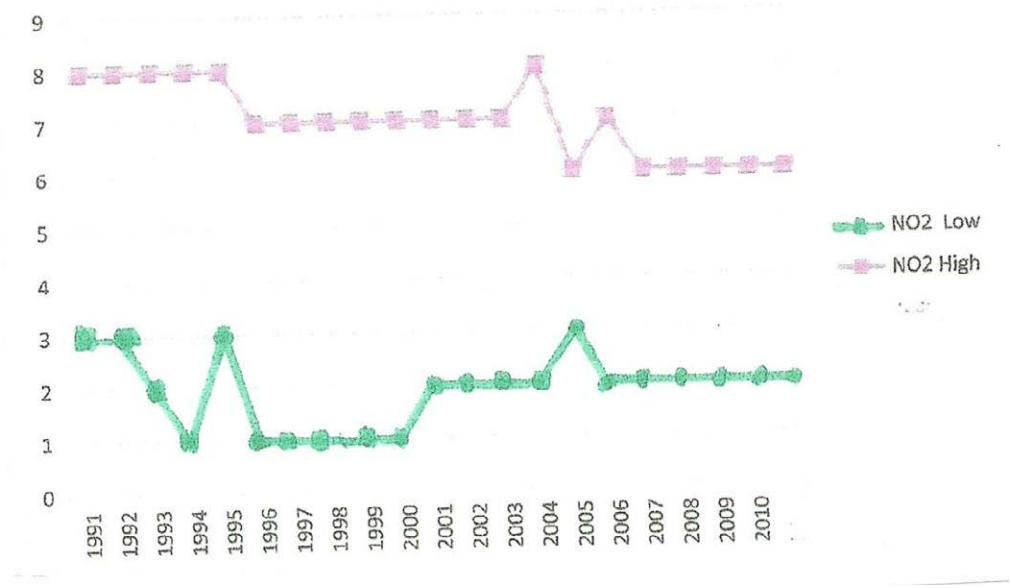
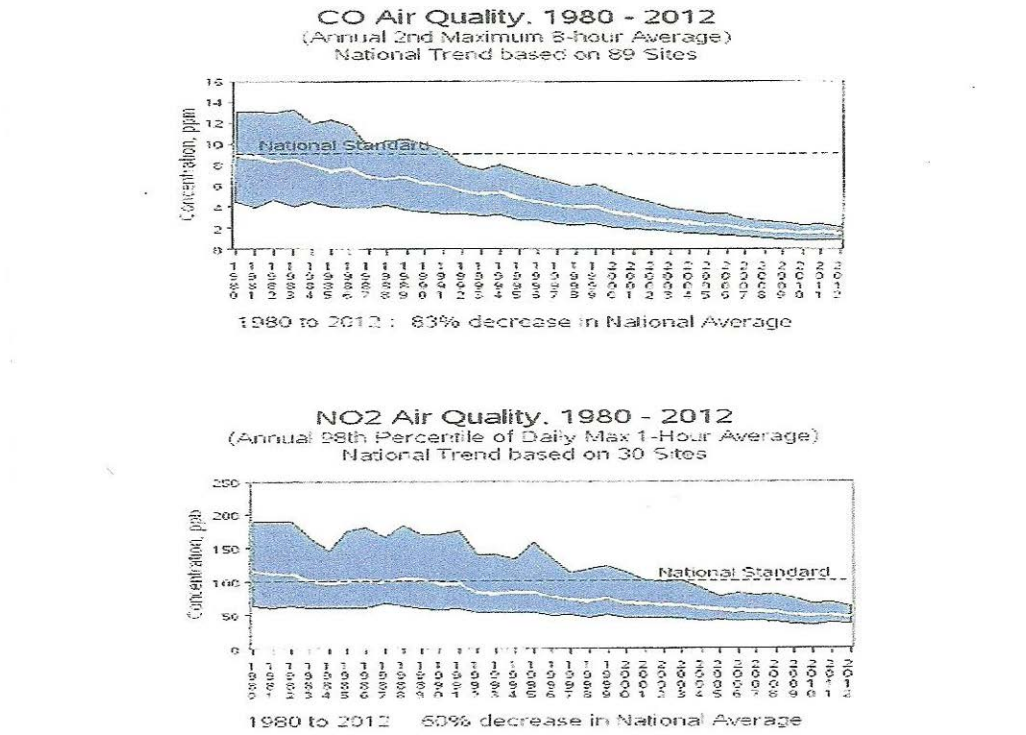


Figure 6: Step-based time oriented shifts traced by application of Trinary monitoring

Implications arise for not only tracing the status of individual locations, but also for integrating our results with summaries provided in large datasets available to the public. Figure 7 displays a retrieval from Airtrends, the status and trend reports of the Environmental Protection Agency (2001). Annual levels of Carbon Monoxide and Nitrogen Dioxide are shown over time, with upper and lower bands bracketing the 90^{th} and 10^{th} percentiles in the blue center, giving a clearer glimpse into the ebb and flow of data over time. Thus it is possible for an individual location, e.g., city or a person who has lived in a specific location to compare possible exposure levels over time with national trends obtained from relatively consistent sampling frames. This transitional orientation from point estimates to Trinary intervals with a set-aside reduced output cache provides a Bayesian learning framework for use with streaming data. Not only are triggers provided for where and when changes occur, but personalized medicine can benefit from the approach by identifying probable locations of exposure.



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

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

Figure 7: National trends for Carbon Monoxide and Nitrogen Dioxide

There is potential for step-shifts derived from a Trinomial or Trinary perspective for uses as supplements to other statistical analyses.

2.2 Geographical Information Science (GIS) Data on Lung Cancer Mortality

Exposure to elevated levels of Carbon Dioxide and Nitrogen Dioxide has adverse pulmonary effects and is associated with respiratory ailments. To take the inquiry a step further, Geographic Information Science (GIS) based exploration are useful We used a region within eastern U.S. from the *Atlas of United States Mortality* (Pickle et al, 1996) showing lung cancer mortality rates at the geocoding level of Health Service Areas (HSA). A sample map is shown in Figure 8.

HSA's are groups of counties representing similarity in terms of health service utilization. There were 860 HSAs; the first 166, located primarily in New England, Middle Atlantic and North Central U.S. were used in our analyses. Each HSA had data for white males, white females, black males and black females. We used this subgrouping to illustrate the importance of mining the database prior to generalizing our conclusions from the full dataset.

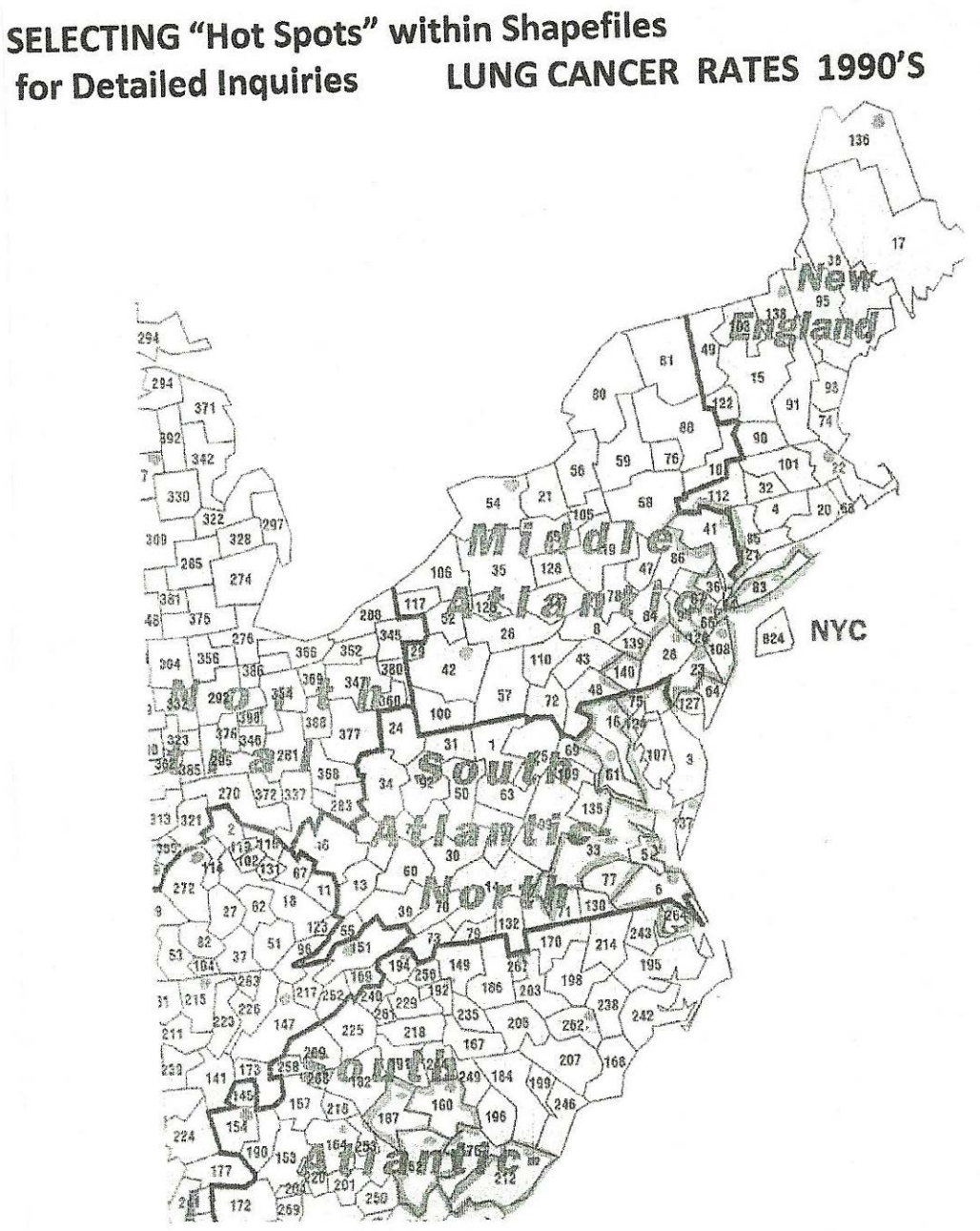


Figure 8: Health Service Area (HSA) Based representation for GIS analysess
Source: Pickle et al (1996)

Within the full dataset we observed lack of statistically significant correlations across race-gender subgroups, as shown below in Table 1.

Table 1: Pearson Product Moment and Spearman Rank r results -between black

SUBGROUPS	PEARSON PRODUCT MOMENT CORRELATION	SPEARMAN RANK CORRELARION
WHITE MALES WITH WHITE FEMALES	0.43	0.41
BLACK MALES WITH BLACK FEMALES	0.29	0.36
WHITE MALES WITH BLACK MALES	0.25	0.29
WHITE FEMALES WITH BLACK FEMALES	-0.01	0.08

The correlations between black and white females were low or close to zero. In contrast, the highest correlation was r=0.43 Pearson Product Moment r between the rates for white males and white females. This prompted a search for an additional, alternative approach to mining the database using Trinary orientation: Core and Non-core in the (+) and (-) ranges and using them to identify "conjoint" {++} observations in pairs or tuplets of gender/race subgroups.

The four race/gender subgroups were scanned and assigned a Trinary code of (+), 0 or (-). Then each of the 161 records in each gender/race subgroup in each HSA were scanned to isolate those pairs which were {++} across pairwise tuplets. For example, in comparing the rate for White Males with the White Female rate in HSA # 1, was the classification (+) for both? Results are shown in Table 2, identifying the geographic location codes of conjoint {++} coded HSAs, the total number of conjoint records identified and percentage of total comparisons.

Table 2: Trinary (+/0/-) orientation applied to isolating conjoint subsets

Records Identified for Tuplets of SUBGROUPS

.. Conjoint Records Identified for Tuplets of Race/Gender Subgroups: White Males with White Females (Row 2), White Males with Black Males (Row 3), Black Males with Black Females (Row 4) and White Females with Black Females (Row 5)

SUBGROUPS	Conjoint +/+	# Records	% of Total
White Males with White Females	2, 7, 11, 13, 16, 18, 37, 46, 55 , 60 67, 77, 89, 96, 116, 123, 137, 155, 158 161	20	12.43
White Males with Black Males	2, 13, 27, 29, 37, 45, 51, 62, 82, 96, 114, 115, 148, 154	14	8.69
Black Males with Black Females	24, 29, 35, 44, 47, 52, 63, 73, 105, 115 119	11	6.83
White Females with Black Females	7, 18, 119, 124	4	2.48

Efficiencies are implicit in this approach. We were able to isolate approximately 3-12% of the observations for further study. For example, might environment-related factors or similar propensisties across race/gender groups yield insights?

Thus, we are advocating an iterative stepwise search within Big Data for identification of causal elements or drivers as stressed by Karabel (2014, 2013). In this endeavor it is advantageous to use proxy or surrogate variables existing in large databases available through the U.S. Bureau of the Census, Environmental Protection Agency, National Cancer Institute and other sources.

3. Summary

This paper adapts our prior advocacy of Trinomial (Trinary), as opposed to Binary analyses, to be more specifically applicable to data streaming into Big datasets. Our method avoids unsupportable assumptions such as stationarity and consistency over time. We emphasize that this is an exploratory method useful in focusing on those elements and combinations within a Big dataset that are most

likely to reveal specific factors and relationships that are most relevant to a critical issue, such as environmental pollution hotspots and potential precursors of disease in demographic subgroups. This is but one example of a vast array of significant issues which can benefit from this exploratory approach.

References

Gardenier, Turkan K. Trinary (+/0/-) Categorization for tracing step-based shifts over time. Presented at *Memorial Workshop for George E. P. Box,* George Washington University, Washington, D.C. May, 2014.

Gardenier, Turkan K. Step-function approach to time series: an air quality application. *International Journal of Applied Science and Technology,* 3(8): 15-20, 2013.

Gardenier, Turkan K. and John S. Gardenier. Delving into megadata: evolving challenges. *Proceedings of the Statistical Learning and Data Mining Section, 2013 Joint Statistical Meetings,* 324-333, 2013.

Gardenier, Turkan K. Depicting time-dependent changes in environmental data for prospective uses in personalized medicine. *Proceedings of the 2013 Research Confeernce of the Federal Committee on Statistical Methodology,* Washington, D.C. November, 2013.

Gardenier, Turkan K. CTSS/CUSUM Applications to intervention analysis. *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings.* Alexandria, VA: American Statistical Association. 777-781, 1984.

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

Howe, Bill and Magdalena Balazinska. Beyond MapReduce: New Requirements for Scalable Data Processing. In *Data Intensive Computing. Architecture, Algorithms and Applications.* Ian Ian Gordon and Deborah K. Garcia (ed). New York, NY: Cambridge University Press, 180-234, 2013.

Kannath, Chandrika. Dimension Reduction for Streaming Data. In *Data Intensive Computing. Architecture, Algorithms and Applications.* Ian Gordon and Deborah K. Garcia (ed). New York, NY: Cambridge University Press, 124-156, 2013.

Kannath, Chandrika. *Scientific Data Mining.* Philadelphi, PA: Society for Industrial and Applied Mathematics (SIAM), 2009.

Karabel, Zachary. *The Leading Indicators: A Short History of the Numbers that Rule Our World.* Simon and Schuster, 2014.

Karabel, Zachary. (Mis)leading Indicators: Why our Economic Numbers Distort Reality. *Foreign Affairs,* 93 (2), 90-101, 2013.

Miller, Renee, Douglas Hale, Turkan K. Gardenier, Virginia Walker. Detecting differences between data series. In *An Assessment of the Quality of Data Series of the Energy Information Administration.* Washington, D.C: DOE/EIA-0292, 1983.

Pickle, Linda W., M Mugniole, G. K. Jones and A. A. White. *Atlas of United States Mortality.* Hyattsville, MD. National Center for Health Statistics, Centers for Disease Control. PHS97-1915, 1996.

Sayed, A. *Fundamentals of Adaptive Filtering.* New York: John Wiley and Sons, 2003.

Sandholm, Thomas.. Solving imperfect information games. *Science,* 347 (6218), 122-123, 9 January, 2015.

Tibshirani, Ryan (a). Don't try for the triple 20: Where to aim if you are bad at darts. *Significance,* 8 (1): 46-48, March 2011.

Tibshirani, Ryan, J. Price, and J. Taylor (b). A statistician plays darts. *Journal of the Royal Statistical Society,* Series 174 (1), 213-226, 2011.

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

Xu, Sheng, Zheng Yan, Kyung-In Jang, Wen Huang, Haoran Fu et. al. Assembly of micro-nanomaterials into complex, three-dimensional architectures by compressive buckling. *Science,* 347 (6218), 154-159, 9 January, 2015.