ABSTRACT

Exploratory data analysis methods usually rely on parametric estimates with the possibility of bias given certain types of data. An alternative, non-parametric approach to characterize multivariate data yields unbiased equivalents to dispersion and peakedness measures of the data. The primary objective is to apply this moment-free method, originated from probabilistic geometry in the past 15 years, to multivariate county-level North Carolina infant mortality data following with an extension to exploration of infant mortality differentials, with the expectation that this method can be applied to a wide range of other types of data.

INTRODUCTION

This paper outlines the application of an alternative method to exploratory data analysis for multi-dimensional data. The method, originating from probabilistic geometry in the past fifteen years, when applied to infant mortality (IM) data, provides information regarding distributional changes over four groups of years. The aim of the proposed method is an attempt through exploratory data analysis techniques to supply a unique perspective on changes that exist in multivariate IM distributions over time, including race and education. Verification of any changes occurs through a moment-free method based on a statistical depth function termed simplicial depth (Liu 1990). Specifically, the exploratory analysis based on depths produces centrality, dispersion and Kurtosis measures from depth measures. Although inferences are made from the distributional changes regarding differentials in infant mortality and race, the aim of this paper is to not to quantify an infant mortality differential between racial groups, as would a parametric analysis.

Depths are measures assigned to each point in a data set indicating their closeness to the most central point of the data. There are varieties of depths and simplicial depths are useful since they require fewer distributional assumptions than another commonly used depth, the Mahalanobis depth. In fact, knowing any resemblance to a distribution is not necessary. In particular, the use of simplicial depths (SD) offer less biased estimates of data characteristics, assuring conclusions that are more accurate.

One motivation behind SD use is its robust estimates even with asymmetric data. Often exploratory analysis fails with data that do not fit the symmetric assumption. The univariate infant mortality data used from North Carolina from 1989-2000 show asymmetry (not shown) with standardized normal curves fitted to the histogram of data. Additionally, using depths to obtain data characteristics allows examination of multivariate data, in our case allowing the examination of IMR, race and educational measures simultaneously instead of examining characteristics individually. The possibility to analyze multivariate data also leads to the inferences regarding IM differentials that follow in the discussion.

A description of the data follows in the methods section, which also includes a description of simplicial depth formation as well as depth applications. After the methods, results from three depth applications used with North Carolina infant mortality data come next, and lastly, there is a discussion of the results and conclusion.

METHODS

Depth Background and Formation

Depths apply to data in all dimensions, but three dimensions will be the maximum number of dimensions in this paper.

A simplicial depth produces “center-outward ordering” of each of the county points using their geometric distributional properties. In other words, “the simplicial depth is defined to measure the relative position of a point w.r.t. a distribution and thus to capture the underlying probabilistic geometry” (Liu 1999:792). With simplicial depths, one can be assured of determining outlying points while still respecting the probabilistic nature of the data -- even with asymmetric data. Most importantly, simplicial depth (SD) yields almost identical results as a parametric method with symmetric data, but outperforms its parametric counterpart in best identifying outlying points in an asymmetric distribution such as a bivariate exponential distribution (Liu 1999). Other methods, such as Mahalanobis depth determine the quadratic distance with respect to one point, the mean, which may not consider the structure of the data if certain distributional assumptions are not met.

In empirical terms, Figure 1 demonstrates the fit of SD values to the asymmetric data in the two-dimensional plot. There are three contours, each one representing a quartile of data points according to their SD value. The innermost contour encloses the 25 most central points as determined by the SD. Those points outside of the outermost ring are the 25 percent least central points. The most central point of that distribution is an ‘O’ shape. The contours in Figure 1 have irregular shapes owing to the nature of simplicial depths and the properties of the data. Drawing contours according to a depth with symmetric data assumptions (not shown) yields more symmetric shapes and with these data, the contours, representing centrality of the data, would not conform well to the oblong shape of the data also yielding misleading information regarding the status of a point relative to its center.

Given the suitability of this method, a brief description of the depth formation, from a three-dimensional data cloud as shown in Figure 2 and otherwise referred to as a collection of data points, is in order. The first step is to count the number of unique tetrahedrons from the cloud that encompass one point in the cloud. This number, when divided by the total number of unique tetrahedrons formed within this data cloud, equals the empirical simplicial depth for that one point. The center of the data cloud in this scheme is the point(s) having the most enclosures. Applying this approach to one-dimensional data would produce a median. In this particular application the following variables comprise the axes for the three dimensional space: percent black infants, number infant deaths per 1000 live births, and percent not completing high school (Figure 2). It’s also important to note that while the data is of three dimensions, the depth assigned to one point yields a one dimensional statistic.

In a more formal sense and directly relating to this applied situation, simplicial depth (Liu 90) at point x with respect to continuous unknown distribution F has the following definition:

$$\mathrm{SD}(F_n;x)=\binom{n}{4}^{-1}\sum_{1\le i<j<k<m\le n} I(x\in\Delta(X_i,X_j,X_k,X_m)) \qquad (1)$$

Where SD(F_n;x) indicates the probability point x is in a randomly chosen simplex with four vertices in the data cloud, unknown distribution F, with n points. In this application, n equals 66. Formula 1 is the empirical version of SD, equaling the proportion of all possible simplices from the sample points that enclose the point in question. The numerator being all possible choices of tetrahedrons and the numerator, a sum of I(.), which is an indicator function signifying whether or not the simplex encloses the point: '1' if yes, '0' if not. In this case, the simplex is a tetrahedron with four vertices in 3-dimensional real space (Figure 2). Another example would be two-dimensional space in which the simplex would be a triangle with three vertices. Thus, for three dimensions all possible tetrahedrons will be evaluated point by point to determine if they include that point within its boundaries.

Programs formed from SAS/BASE© software generated simplicial depth values through algorithms designed to measure and compare volumes of tetrahedrons (Von Holle 2003).

Source of Data

Linked infant birth/infant death files for 1989-2000 from the North Carolina State Center for Health Statistics serve as the basis for analyses (Odum Institute 2002). Data reporting was uniform from year to year with one exception: infant race reporting was based on that of the mother starting in 1990 instead of being based on race of both mother and father.

The objective in forming the multivariate distribution of interest was to obtain three county-level estimates. Infant mortality was the first estimate of interest. Secondly, among the non-Hispanic population, percentage of black mothers, also considered the percent of black births, was another estimate of interest. Third and last, the percent not completing high school provides a proxy estimate of the socio-economic status of a county, adding another dimension to the analysis. Characterizing the change of kurtosis and scale of a multivariate distribution by time of this multivariate distribution assists in determining direction of differential changes.

Pooling of birth and infant death counts occurred at the county level, and any county with less than 1,000 births per year was pooled with an adjacent county. The result was four sets of three successive years of pooled data totaling 1,000+ live births. There are 100 counties in North Carolina and 66 county units after pooling for each of the four groups of three years taken from 1989 to 2000. Having 1,000+ live births for the denominator of each county unit ensured more precision in the resulting proportions. Several more criteria applied for inclusion: the record had to be a singleton black or white non-Hispanic birth to narrow the scope of analysis and ensure consistency. Inclusion of the variable indicating completion of high school served as a proxy for socio-economic status in the analysis since prior studies have indicated its influence on infant mortality (Hummer 1999) and this variable is a reliable one (Buescher 1992).

The North Carolina Center for Health Informatics and Statistics (CHIS) reported data for years up to 1998 differently than those on or after 1998. Because of this, infant death probability calculations change to accommodate this difference in reporting. Before 1998, deaths were organized and linked according to the birth cohort; afterwards they were by death cohort. Adjustments were done to approximate a birth cohort estimate from 1998-2000 using a bridge file provided by CHIS with assumptions that death rates were constant over the year 2000.

Simplicial Depth Applications

SD are useful in their own right as a means to assess the centrality of data points in a robust manner, but they also have use as order statistics (Liu 1999). From calculations exclusively involving SD and their properties in two or three dimensions in this specific application, there exist three interesting proxies to dispersion, kurtosis and consistency. Each of the 66 county groups provides information to conduct further analyses that assist in the assessment of the distributions from one group of years to the next. For example, contours based on SD structured in the data cloud, which in this paper are in three dimensions, and the rate of change along contours outward from the center, neatly characterize some aspects of the distribution.

There are different uses of depth possible to characterize a distribution, but only three applications are outlined in this paper. These three applications and interpretations (Liu 1999) use simplicial depths:

1) Data-Depth Plots (DD-plot)
2) Dispersion of data by way of Convex Hull Volume plots
3) Shrinkage Plots (Kurtosis)

Analysis 1 -- Data-Depth Plots

Only the empirical version of the DD-plot is possible since the distribution of data is unknown. The approach is simple: take all points from both distributions in question, determine the depth of each of those points according to one distribution, and plot that value versus the depth according to the other distribution. For example, for the plot of 1989-1991 versus 1992-1994 data, has 132 points, each point with a depth according to the 1989-1991 distribution of points (64) and a depth according to the 1992-1994 distribution of points (64). This two dimensional display will show a line very close to the diagonal if the two distributions from different points in time are identical. If the distributions are not similar, by either skewness, scale, kurtosis, location or another source then the line will differ from the straight-line diagonal. One case would be curvature of the data above the diagonal line, indicating a scale or kurtosis difference.

Analysis 2 – Dispersion by way of Volume Plots

Plots of the volume formed by the data cloud as measured through a convex hull volume and SD values assigned to each point in the three dimensions can show how rapidly some distributions expand out compared to others. Both definitions for volume plots and shrinkage plots rely on the definition of convex hulls with the most simplistic version existing in two dimensions with a hull consisting of lines connecting the outermost points of a data cloud in a two-dimensional plane. A fixed convex hull is defined as $C_{n,p}$, where n denotes the number of points in the 3 dimensional space (Figure 2), and p denotes the p^{th} central region. For example, $C_{66,0.5}$ is the convex hull that encloses 50% of the sixty-six points closest to the center. All convex hull volume calculations were done in MATLAB©.

Plotting the volume of convex hulls $C_{n,p}$ versus p, the p^{th} central region, reflects data dispersion. The degree to which the volume expands as the number of points expands from the center of the data cloud, according to their individual simplicial depth values, shows dispersion. If one time period has larger volume at the proportion of points, p, than some other time period then the conclusion is that the former time period is more disperse than the latter.

Analysis 3 – Shrinkage Plots – Kurtosis and use of Lorenz curve

Up to this point, two methods have been shown to detect distributional changes in data, the first being a DD-plot and the second being a simple plot of the p^{th} central fraction of points versus its corresponding convex hull volume. The latter method addresses data dispersion and the former addresses type of distribution change if one exists. Both methods serve as a check for generally defined changes in scale, kurtosis, or skewness, and have less complex interpretation than a shrinkage plot.

Generation of shrinkage plots is a first step to assess kurtosis. To construct one plot to assess kurtosis, start at a fixed convex hull volume, and then shrink the hull toward the center by a certain amount to yield another hull. Measures include the percent of total volume lost through reduction to that new hull by a factor of s and the percentage of data points lost as you shrunk to a smaller space. That step is repeated many times to produce values of loss of volume, V(s), and loss of points, l(s), both losses expressed as fractions of the starting volume and total data points.

This technique shows proportional changes in volume in tandem with proportional changes in data points, and produces actual quantitative results: a Gini coefficient based on the Lorenz curve concept. The Gini coefficient here functions as a measure of kurtosis of multivariate data (Liu 1999), the larger the coefficient the more kurtotic the data and the farther the line deviates from the diagonal. Plotting V(s) vs l(s) generates a Lorenz curve and the area between the curve and the diagonal line is the Gini coefficient. A coefficient close to one would indicate most data would be concentrated in the middle of cloud with very distant outliers in the space – extremely kurtotic and with a shrinkage plot line far from the diagonal. If the coefficient was close to zero then the line would follow the diagonal indicating points uniformly spread out over the space – not kurtotic at all.

RESULTS

Each of the three SD applications provides different information, one being more complex than others.

Data-Depth Plots

There is little consistency in the original data depth plots perhaps indicating a change in scale, kurtosis and or skewness in the empirical distribution. After adjusting for the center and variance of the data (Liu 1999), the points draw much closer to the diagonal and distributional changes are more suspect. Yet there are still suggestions of asymmetry about the diagonal indicating possible distributional changes. This pattern is most evident for the 1989-1991 to 1992-1994 dd-plot, and the asymmetry hints of possible kurtosis and scale changes which can be verified with other types of plots. It is also important to note that many points cluster near the origin due to more than one quarter of the data having ties at a zero depth value.

After adjusting the data points, the most central point is not the same for the last two plots, suggesting locations shifts from one time point to another. Accompanying this outcome, the center of the 3-d data cloud goes from (7.7=IMR, 15.9=percent black infants, 20.8=percent not completing high school) in 1989-1991 to (7.9, 19.1, 15.5) for 1992-1994, to (7.3, 18.2, 14.3) for 1995-1997 and finally (7.0, 19.5, 12.9) for 1998-2000. These shifts could be due to levels of any of the three variables but notable are the consistent declines in IMR and education rates.

Dispersion Plots

In this particular case with the North Carolina data having measures of IMR, race of child and education characteristics by county, the volumes of each data cloud per time group are similar and do not show any vast differences in volume change over with the exception of the 1989-1991 time period (Figure 4). After covering 60% of the innermost points to the center of the distribution, the 60th central fraction of points, the consistent trend is one with the earliest time grouping having a denser, less disperse collection of points than the later years. However, the earliest time grouping jumps ahead at the point indicating 100% of the data is covered. What this plot indicates is that the earliest time covers the most total space and has points in its distribution with the largest distance between each other than the other time groups.

One way to test any change in depth values from one group of years to the next, while assuming no location shift, is a Wilcoxon Rank-Sum test on the simplicial depth ranks for pairs of consecutive year groups (1989-1991 and 1992-1994, 1992-1994 and 1995-1997, and 1995-1997 and 1998-2000). Each of the three test statistics are not significant at an alpha level of 0.05 suggesting that the differences in Figure 4 are due to chance. In other words, the depths do not show a significant change from one time to the next indicating that instead of the points in each data distribution expanding or shrinking over time they have a similar dispersion level.

This rank-sum test is not as powerful as parametric tests so a difference may exist, but could go undetected in this case. However, the rank-sum test is the recommended method for testing changes in scale for empirical data (Liu 1993). Besides lower power, this non-parametric test assumes the pairs of years are independent and few ties. In reality, the sequence of years most likely is not independent, and over 30% of the data for any year group have tied SD leaving doubt about the validity of the statistic.

Shrinkage Plots -- Kurtosis

The shrinkage plots in Figure 5 shows this technique for the 60th central hull, enclosing the 60 percent innermost points in the distribution, displaying information on any differences in kurtosis of distributions for each set of years. Typically, Gini coefficients produced from multiple p^{th} central hull values versus p, compose one plot, and that plot provides a complete picture of kurtosis of the data. However, so few observations outside of the 60th central hull without ties at zero, 6 SD values, made it difficult to get a sufficient number of volume calculations. In addition, ties made it impossible to order the ranks as needed for meaningful arrangements of points for successive volume calculations outside the 60th central hull. Despite this situation, single plots based on 60th central hull information still provides useful clues regarding the change in distribution of county infant mortality rates relative to their education and race status.

For the plot based on three-dimensions in figure 5, 1989-1991, the earliest group of years, stands out from other years which remain quite similar to one another. This group of years is also the furthest from the diagonal indicating a more kurtotic distribution than the rest. For example, after approximately covering 40% of the outermost points in the data cloud, the years 1989-1991 lost the most volume relative to its total volume covered in the cloud. Progressing from the outermost hull to the center, 50% of the most extreme points from the center occupy over 90% of the total hull volume, with a sharp uptake in points for the remaining 10% volume before attaining the outer border of the hull for 1989-1991. In contrast, for the later time groupings, 50% of the points occupy less than 84% of the total space – more data points occupying less volume. This trend continues with progression towards the center of the data cloud, 100% of volume lost, thus, the 1989-1991 period, with fewer points on the outer edges of the distribution and more towards the center of the distribution, is more kurtotic.

The largest Gini coefficient of 0.4739 for 1989-1991, in the shrinkage plot based on three-dimensions, further substantiates the status of 1989-1991 as extreme. In more explicit terms, as the p^{th} central hull contracts towards the center of the distribution from its outermost border at the 60th central hull, the increase in proportion of points does not keep up with the proportional change in volume. This means that the proportional change in number of points cannot maintain the same, larger, proportional change in volume and the points are more spread out.

Adding shrinkage plots based on two dimensions, excluding either the education or race variable from the original 3-d data, provides better context for the shrinkage plot based on three dimensions. There is a bigger difference between the 3-d and 2-d plot when removing race than when removing education, subsequently providing evidence for racial and IMR distributional differences along the education gradient changing more than educational and IMR distributional differences along a racial gradient. Additionally, maintenance of differences from one group of years to the next, when switching shrinkage plots, from 3-d to 2-d plot with IMR and race, shows there still exist changes in distribution on the racial and IMR plane. This maintenance rules out the possibility that only educational and IMR distributional changes are responsible for year-to-year differences. Ultimately, the 3-d plot is the primary analysis tool, and the education variable appears to be a good addition to the examination since its distributions also change over time and can better define the racial and IMR distributions.

DISCUSSION

Methods producing characteristics of an empirical unknown distribution stand on their own for exploratory analysis, but further conclusions exist in that these characteristics reflect behavior of the variables themselves and the points in 3-d space they represent. Changes in distances between the points in three-dimensional space directly link to changes in dispersion or kurtosis. Furthermore, the distance between two points on the infant mortality and racial plane signifies an infant mortality differential by race. Figure 1 demonstrates this concept in two dimensions for the North Carolina IMR data while figure 2 does so for three dimensions. Adding the third dimension allows the consideration of education levels simultaneously, a version of a control, in the racial differential in infant mortality.

If a distribution becomes less kurtotic over time then a direct conclusion would be that the distribution is less dense around the center and the points in the data cloud are migrating away from the center and vice versa for a distribution becoming more kurtotic over time. For these data in particular, the dispersion and shrinkage plots based on 3-d data indicate that after the 1989-1991 period, the distribution becomes less concentrated around the center. The earliest time has lower volume of its data cloud up to a certain point outward from the center, then that period of time jumps ahead in total volume, a finding confirmed in the shrinkage plot from Figure 5, which shows a more kurtotic distribution for 1989-1991. One caveat is the limited nature of this shrinkage plot, based only on the 60th central hull.

It is apparent from the shrinkage plot based on three-dimensions (Figure 5, Plot A) that 1998-2000 remains less kurtotic than the 1989-1991 kurtosis measure (a Gini coefficient of 0.39 versus 0.47) with an implication that the absolute distance between a large group of points along at least one axis, in particular the race and IMR axis, is growing. The difference between 1998-2000 and 1989-1991 values also imply less concentrated values about the center over the course of a decade, not a positive development if movement of points away from the center towards a higher IMR for certain racial levels. The dispersion plot also shows that points may be growing apart on the race and IMR plane over time, but 1989-1991 having the largest scale (Figure 4) shows it having the most distant outliers from the center of the distribution implying the largest IMR differentials by race either with or without (not shown) the education factor.

The other two groups of time, 1992-1994 and 1995-1997 grow farther apart from 1989-1991 towards the diagonal, and this differential in IMR could be due more to education since this trend was not apparent in the 2-d shrinkage plot excluding education. The change contributes to an intriguing observation that differentials in IMR may be growing over time by education levels for the earlier groups of time, when relying on the 'the less kurtotic, the bigger the differentials' interpretation.

Testing these results would be desirable to substantiate these claims based on entire population values, but the only option available is to test for a shift in SD between the time distributions, four in total, and one for each time. This test, if significant, renders a change in scale as being statistically significant, yet this does not occur as noted in the results. So there is no formal test here showing the trend in decreasing scale is a significant one. However, while there is no change in location, an assumption for the test and shown by the DD-plots (Figure 3), this test is not ideal for this particular situation since the data have many ties and lack independence from one time group to the next.

CONCLUSION

Characterizations of the data via SD have led to certain conjectures about the status of infant mortality differentials. Primary among them is the conclusion that county-level infant mortality differences on the racial plane, when considering education levels simultaneously, increased somewhat after 1989-1991. This result comes from indications that the data distribution is getting to be less kurtotic and more disperse from the earliest period, 1989-1991. At this point, using SD to target changes in distribution with implications for IMR differentials remain exploratory at best given limitations of the data and method, but do pose interesting questions. Among them are: 1) Given the data depth plots, is the data becoming more reliable or are there IMR changes occurring? 2) Are these changes subtle enough to be missed by a parametric analysis, given the data at hand? and 3) If the distances between certain IMR are becoming larger across race values, what reasons exist for this development?

Of course, confirmation of these findings via a parametric approach is possible, either on a county or individual level, with parametric coefficients indicating an average difference in IMR between racial groups. However, the reason behind this analysis was to apply a new non-parametric approach robustly exploring the data with minimal assumptions and perhaps lending some evidence of the direction of infant mortality differentials by race. Another unique outcome from this SD application is the examination of extreme points in the distribution, something a coefficient from a parametric analysis omits.

These characterizations of the data provide a different perception of the data, before a parametric modeling process occurs, yielding quantitative estimates of association and their related standard errors. Considering a simpler alternative in terms of non-parametric statistics, how would this application provide more information than a simple median based on adjusted data? By obtaining the center of multivariate data from SD, the equivalent to a median in one dimension, the dispersion in all three dimensions contributes to the calculations. Estimates are also available of dispersion and kurtosis of the distribution, again incorporating all three variables and their multivariate dispersion.

While these findings are of interest, limitations of the analysis exist, some being data-specific while others are methodological. Specifically, during data analysis, certain issues cropped up. First, counties with small populations receive just as much weight in simplicial depth calculations as the largest ones affecting the determination of central point of data. It is logical to have larger counties with more births affect the center more than counties with fewer births. Weighting in this type of calculation remains a challenge and possible direction for further effort due to the geometric approach. Assigning a weight to a simplicial depth after its calculation is not sufficient. Instead, the point must place multiple times into the same space to calculate depths relative to all other counties, contrary to continuous data assumptions. In addition, making a county value more frequent in the space before depth calculations, will affect other points in the space, not just its own value, meriting a complete reevaluation of all points.

Secondly, the nominal nature of the mortality and race variable does not allow for SD calculations on an individual level. If the data were all continuous for each person, then SD values are possible for individuals instead of counties. As a result, findings apply to county level observations and not individuals. Due to any type of ecologic fallacy these results may not be applicable to individuals, yet the results are still of note since it's valid for counties and allows them to be targeted.

On a somewhat different note, simplicial depth calculations are by far more complicated than other methods like the Mahalanobis depth and require more computational resources. Adding dimensions to the computation is neither straightforward nor is it time efficient, and there exists no software package to determine SD. The effort required to design an algorithm and process SD values from even a minor data set of 100 observations using naïve algorithms is to an order of n^4 steps, n being 66 in this case and 4 being the number of dimensions (Rousseeuw 1996).

After completing the simplicial depth calculations, other issues arose. The large number of ties and properties of time dependent data hampers the formal test of depths, indicating a need for another type of test besides what currently exists in the literature.

Despite these limitations, application of SD provides a different way to assess characteristics of a distribution and identification of outliers with no distributional assumptions required for the data. This aspect alone is important given those types of assumptions for other methods such as the Mahalonobis depth. In addition, the nature of geometric calculations provides a robust way to detect distributional characteristics regardless of shape of the data. Finally, this analysis enables an indirect approach to examine mortality differentials through multivariate exploratory data anlaysis over the broadest perspective possible in as robust a manner as possible.

REFERENCES

Howard W. Odum Institute for Research in Social Science, UNC, Chapel Hill. 2002. Retrieved June 10, 2002, from http://www.irss.unc.edu/ncvital/biddown.html.

Hummer, R. A. 1993. "Racial Differentials in Infant Mortality in the U.S.: An Examination of Social and Health Determinants." *Social Forces* 72(2): 529-554.

--., Biegler, Monique, De Turk, Peter B., Forbes, Douglas, Frisbie, W. Parker, Hong, Ying, Pullum, Starling G. 1999. "Race/Ethnicity, Nativity, and Infant Mortality in the United States." *Social Forces* 77(3): 1083-1117.

Johnson, J. H. 1987. "U.S. Differentials in Infant Mortality: Why Do They Persist?" *Family Planning Perspectives* 19(5): 227-232.

Kahan, W. 2002. What Has the Volume of a Tetrahedron To Do With Computer Programming Languages? (March) Internet. Available from http://www.cs.berkeley.edu/~wkahan/VtetLang.pdf; accessed 10 September 2002.

Liu, R. Y. 1988. "On a Notion of Simplicial Depth." *Proceedings of the National Academy of Sciences of the United States of America* 85(6): 1732-1734.

--. 1990. "On a Notion of Data Depth Based on Random Simplices." *Annals of Statistics* 18(1): 405-414.

--., Singh, K. 1992. "Ordering Directional Data: Concepts of Data Depth on Circles and Spheres." *Annals of Statistics* 20(3): 1468-1484.

--., Singh, K. 1993. "A Quality Index Based on Data Depth and Multivariate Rank Tests." *Journal of the American Statistical Association* 88(421): 252-260.

--. 1995. "Control Charts for Multivariate Processes." Journal of the American Statistical Association 90(432): 1380-1387.

MATLAB, Copyright © 1984-2002 The Mathworks, Inc.

--., Parelius, J.M., Singh, K. 1999. "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference." *Annals of Statistics* 27(3): 783-858.

Rousseeuw, P. J., Ruts, I. 1996. "Algorithm AS 307: Bivariate Location Depth." *Applied Statistics* 45(4): 516-526.

SAS, Copyright © 2001 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademark of SAS Institute Inc., Cary, NC, USA.

Shyrock, H.S., Siegel, J.S. 1976. *The Methods and Materials of Demography*. San Diego: Academic Press, Inc.

Singh, G. K., Yu, S.M. 1995. "Infant Mortality in the United States: Trends, Differentials, and Projections, 1950 through 2010." *American Journal of Public Health* 85(7): 957-964.

Stokes, M.E., Davis, C.S., Koch, G.G. 2000. *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute Inc.

Von Holle, A. 2003. *Data Depth and Infant Mortality Differentials*. Masters Paper. Department of Biostatistics, The University of North Carolina.

Weisstein, E.W. 1999. Convex Hull. (n.d.). Retrieved May 29, 2003, from http://mathworld.wolfram.com/ConvexHull.html

Zuo, Y., Serfling, R. 2000. "General Notions of Statistical Depth Function." *Annals of Statistics* 28(2): 461-482.

FIGURE 1: SIMPLICIAL DEPTH CONTOUR PLOT FOR IMR VERSUS PERCENT BLACK BIRTHS.

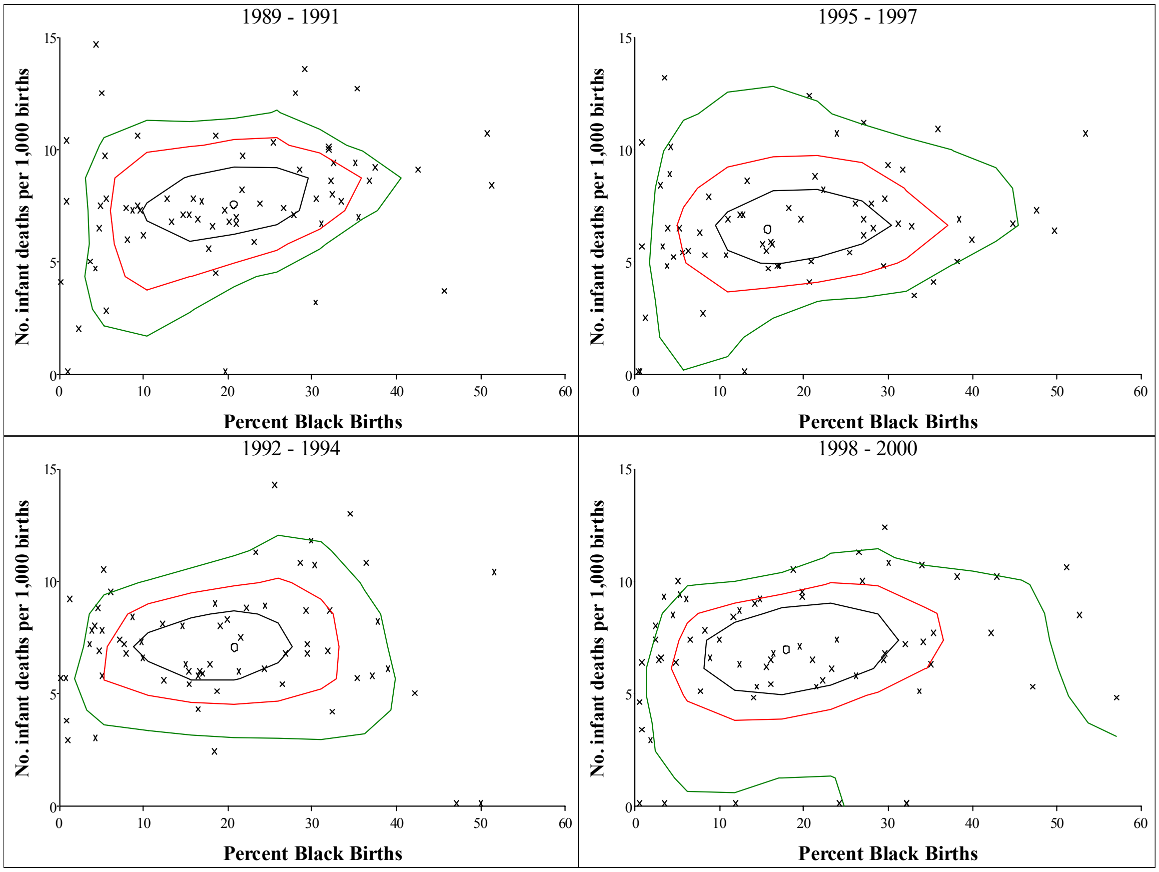


FIGURE 2: THREE-DIMENSIONAL PLOT OF IMR, PERCENT MOTHERS WITHOUT HS EDUCATION AND PERCENT BLACK BIRTHS.

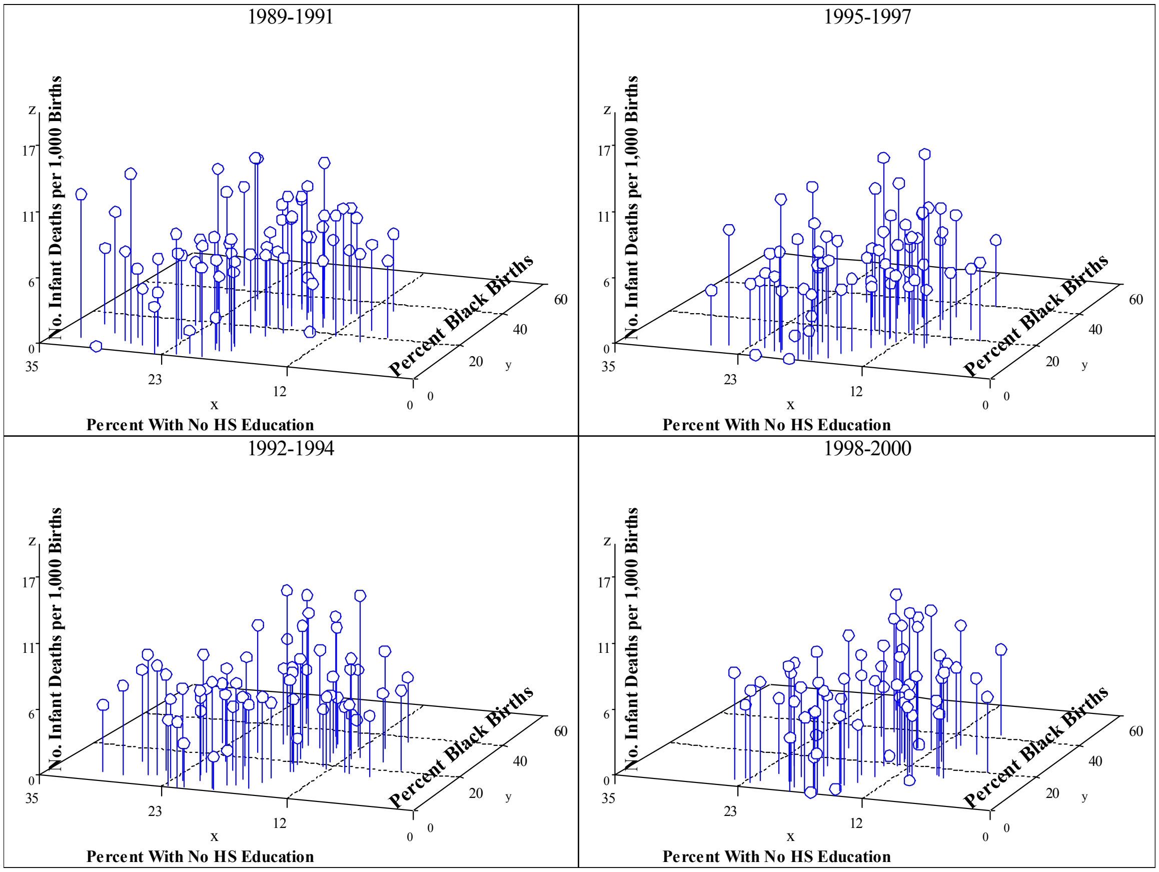


FIGURE 3: DATA-DEPTH PLOTS FOR IMR, RACE, AND EDUCATION DATA POINTS.

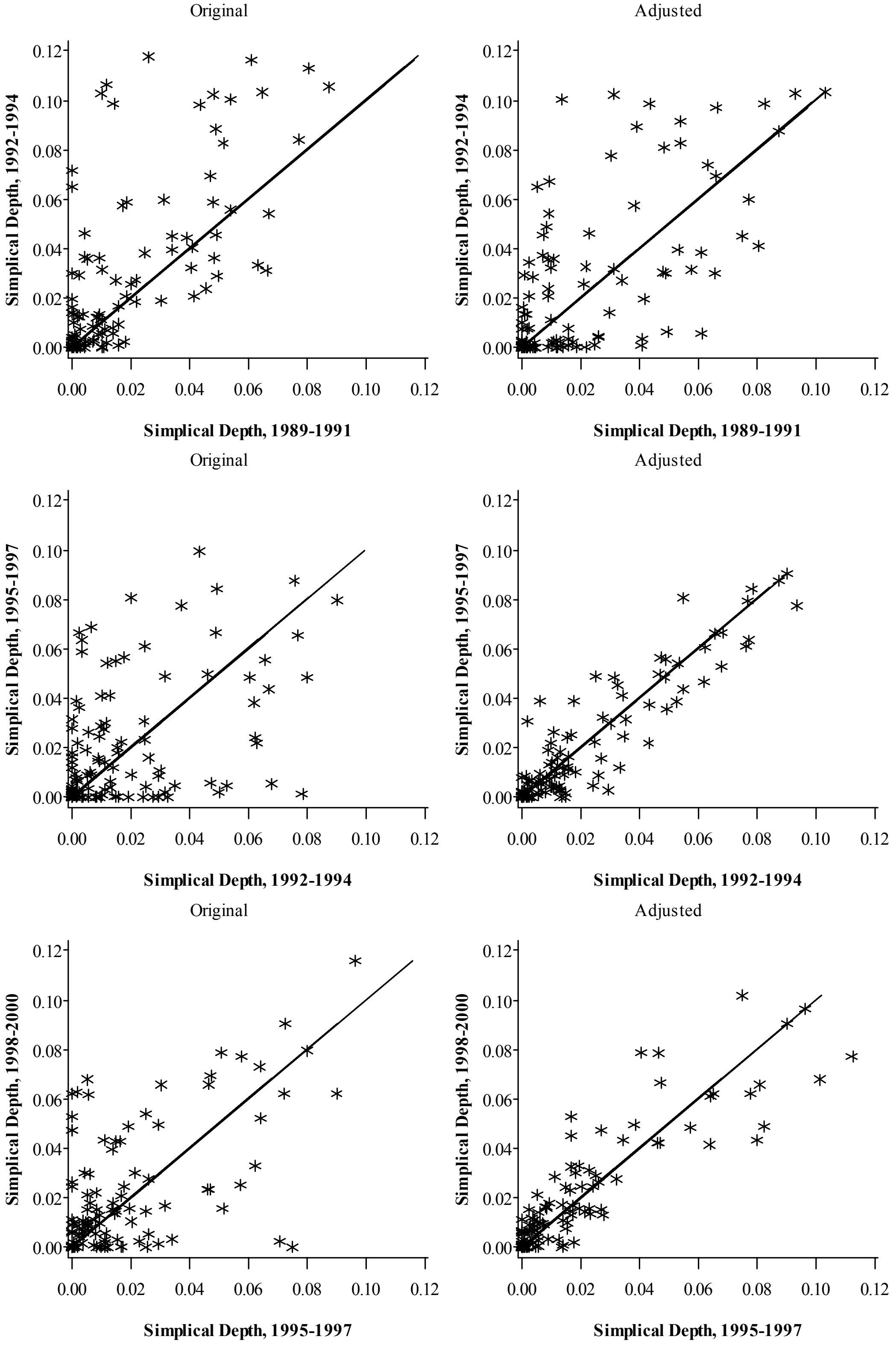


FIGURE 4: DISPERSION PLOT FOR IMR, RACE, AND EDUCATION DATA POINTS

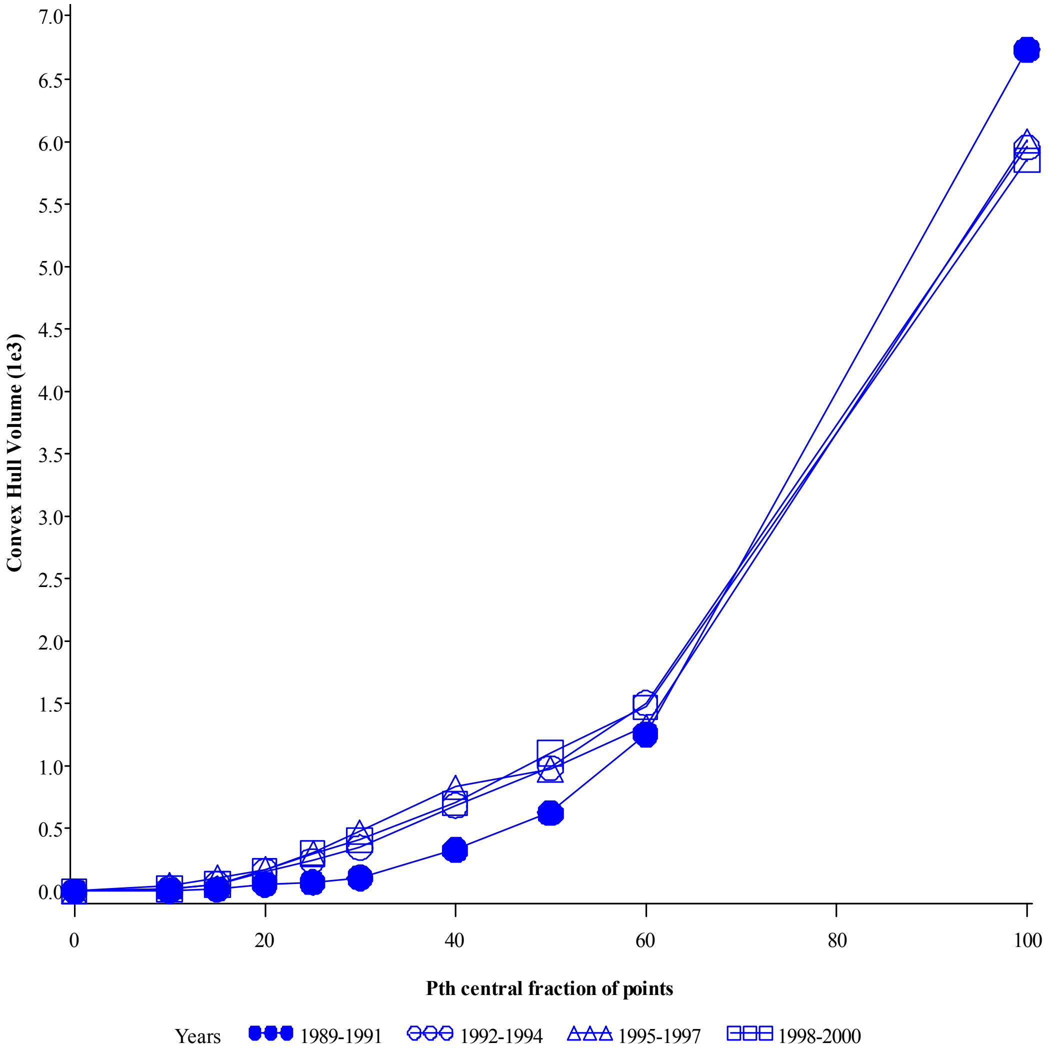
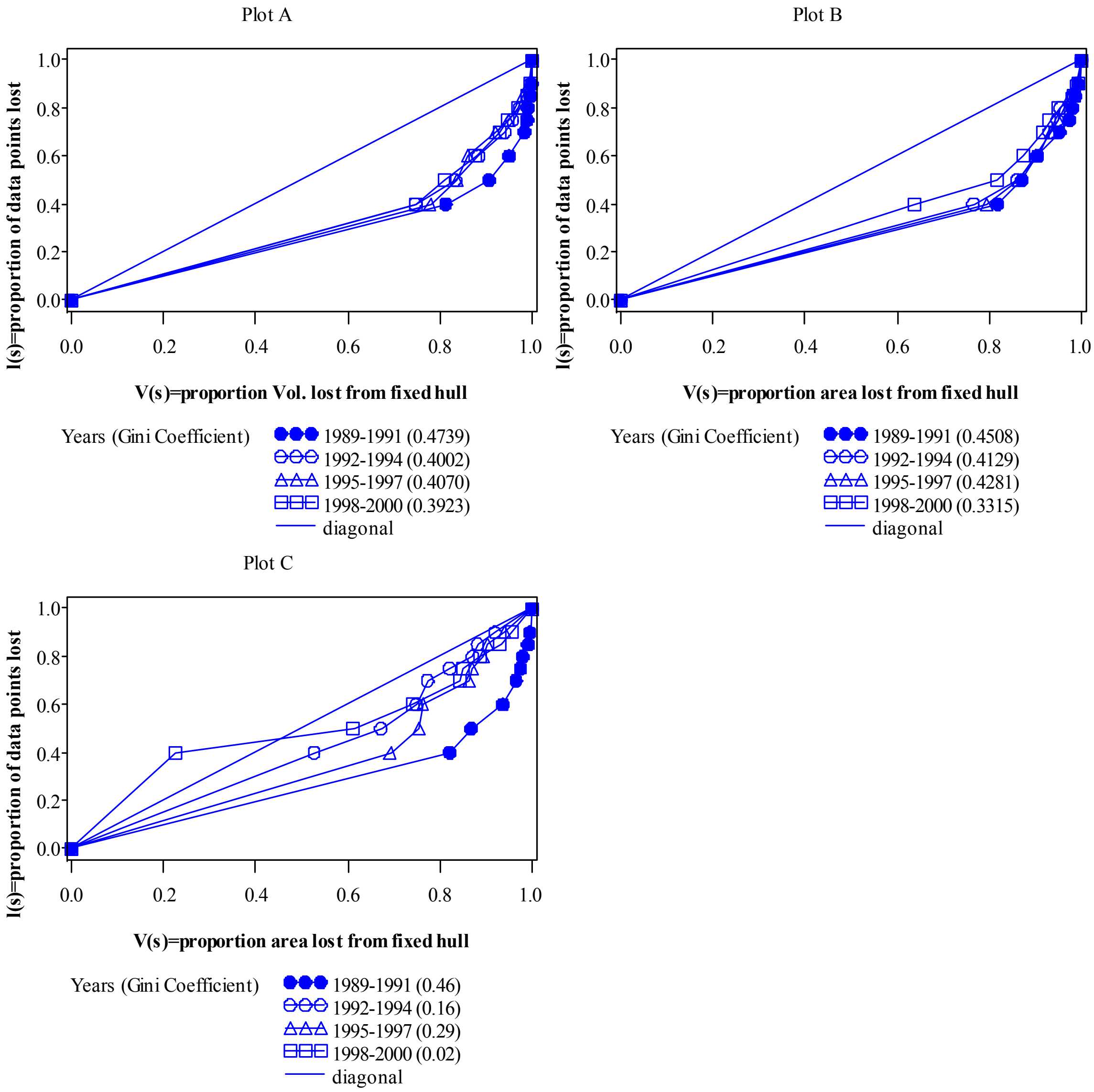


FIGURE 5: SHRINKAGE PLOTS FOR: PLOT A) IMR, RACE, AND EDUCATION DATA POINTS, PLOT B) IMR AND RACE DATA POINTS, AND PLOT C) IMR AND EDUCATION DATA POINTS


Plot A
l(s)=proportion of data points lost
V(s)=proportion Vol. lost from fixed hull
Years (Gini Coefficient)
●●● 1989-1991 (0.4739)
○○○ 1992-1994 (0.4002)
△△△ 1995-1997 (0.4070)
□□□ 1998-2000 (0.3923)
diagonal
Plot B
l(s)=proportion of data points lost
V(s)=proportion area lost from fixed hull
Years (Gini Coefficient)
●●● 1989-1991 (0.4508)
○○○ 1992-1994 (0.4129)
△△△ 1995-1997 (0.4281)
□□□ 1998-2000 (0.3315)
diagonal
Plot C
l(s)=proportion of data points lost
V(s)=proportion area lost from fixed hull
Years (Gini Coefficient)
●●● 1989-1991 (0.46)
○○○ 1992-1994 (0.16)
△△△ 1995-1997 (0.29)
□□□ 1998-2000 (0.02)
diagonal