PENNSTATE

1 8 5 5

Center for **S**tatistical **E**cology
and **E**nvironmental **S**tatistics

AN ECOLOGICAL ASSESSMENT OF THE UNITED STATES MID-ATLANTIC REGION USING RANK FREQUENCY DISTRIBUTIONS BASED ON WATERSHED QUINTILES

By G. P. Patil, C. Taillie, and Ryan Vraney

Center for Statistical Ecology and Environmental Statistics
Department of Statistics
The Pennsylvania State University
University Park, PA 16802

[Invited paper to appear in *Community Ecology*, 2002]

Technical Report Number 2001-1102
TECHNICAL REPORTS AND REPRINTS SERIES
December 2001

Department of Statistics
The Pennsylvania State University
University Park, PA 16802

G. P. Patil
Distinguished Professor and Director
Tel: (814)865-9442 Fax: (814)865-1278
Email: gpp@stat.psu.edu
http://www.stat.psu.edu/~gpp

An Ecological Assessment of the United States Mid-Atlantic Region Using Rank Frequency Distributions Based on Watershed Quintiles

G. P. Patil, C. Taillie, and Ryan Vraney

Center for Statistical Ecology and Environmental Statistics

Pennsylvania State University

University Park, PA, USA

Abstract

When working with raw data for multiple environmental indicators, it can be difficult to assess quality or 'health' because of the large number of indicators and inconsistencies among the indicators. By grouping the raw data into rankings, the data become more manageable and more comprehensible. We do not, however, want to lose information as a result of the groupings. It is possible to assess the quality of grouping options graphically by seeing if the resulting assessments of 'health' are concordant with the raw data. This can be done through the use of CDF-index values, cumulative distribution function plots, parallel coordinates plots, and scatterplots. A major purpose of this paper is to present approaches and the graphics for comparison and prioritization based on quintiles used, in this case, for ecological assessment of a large region.

Keywords: Comparison and prioritization; Ecological assessment; Multiple watershed indicators; Parallel coordinates plots; Quintiles, Septiles, Nine-tiles; Rank distribution function plots; Rank frequency distribution; Triangle composition plots

Introduction

The EPA has compiled a large body of ecological data regarding a five state subsection of the United States in its An Ecological Assessment of the United States Mid-Atlantic Region (Jones, et al., 1998). The five states are Delaware, Maryland, Pennsylvania, Virginia and West Virginia, as well as the District of Columbia. The region is also partitioned into 114 major watersheds, which are used as spatial units for the data. This paper examines the watershed indicator data that was compiled to assess the region. This is a list of 33 indicators, each a measure of the health of the watershed. Values of these 33 indicators are available for each of the 114 watersheds (Jones, et al, Table A1, pp. 96-101; this table can also be found at http://www.epa.gov/maia/html/la-tablea1.html). The indicator names are abbreviated in Table A1. Figure 1 in the appendix lists the full names of the indicators and gives a brief description of each (see the *Ecological Assessment* Atlas for full details).

Each indicator was then divided into quintile ranks, which represent 20% divisions in the data. The top 20% of the watersheds for each indicator were given a rank of 1, the next 20% a rank of 2, and so on down to the bottom 20% being given a rank of 5. Based on the indicator ranks, we would like to determine which watersheds are the healthiest, and which are in the most need of assistance. In order to do this, we are not limited to looking at the quintiles, but we can also divide the raw data into septiles (divisions of seven) and nine-tiles (divisions of nine). We will attempt to assess the watersheds using these divisions as well.

The indicators have not been weighted in any manner. Therefore, each indicator has the same value as every other indicator. We are not trying to assess the data based on what we think is the most important indicator of the health of a watershed, but rather by collecting multiple indicators for each watershed and assessing the watersheds based on the equal importance of these indicators. A major purpose of this paper is to present approaches and the graphics for comparison and prioritization based on watershed percentiles.

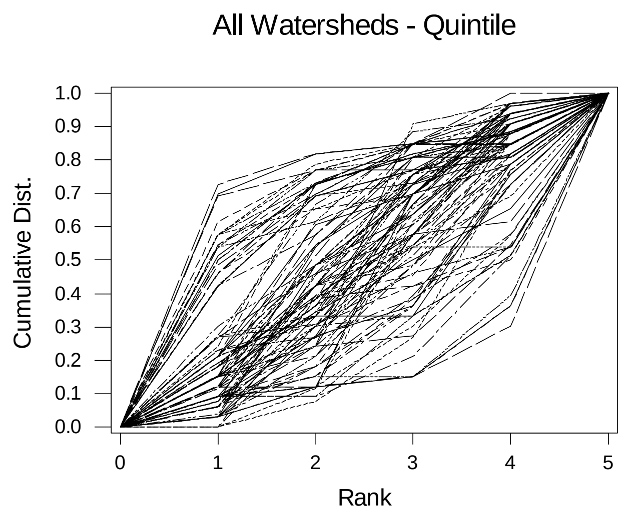
Analysis

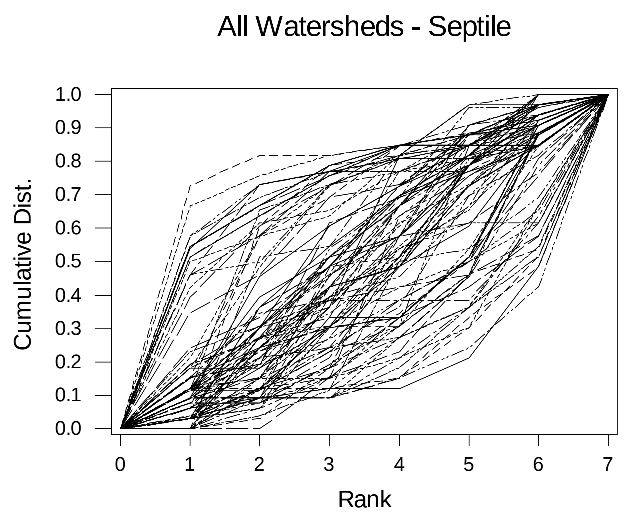
We begin with the raw data that is listed in the EPA's An Ecological Assessment of the United States Mid-Atlantic Region. The data has already been divided into quintile ranks. We have also divided the data into septile and nine-tile ranks. Each septile represents 14.29% of the watersheds for that particular indicator, with the top seventh receiving a rank of one and so on down to the worst seventh receiving a rank of seven. Similarly, the top nine-tile represents the best 11.11% of the watersheds for that indicator.

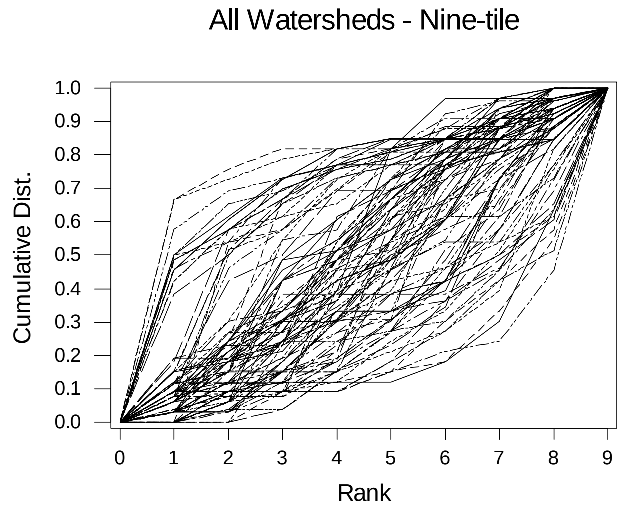
Our next step is to find the cumulative distributions of the different ranks for each watershed. We will use these distribution functions as overall assessments of the quality or 'health' of the corresponding watersheds. We want to find the cumulative distributions for all three ranking options: quintiles, septiles, and nine-tiles. Once we have found the cumulative functions, we can get a brief overview of them by looking at the plots of these functions.

If we were to look at the plots for the cumulative distribution functions of all 114 watersheds (Figures 2 – 4), we would notice that as the number of ranks increases, so does the spread of the graph. If you look at the picture for the quintiles, you notice that it is much more compact than the septile graph, which in turn is more compact than the nine-tile graph. What we would like to know is if there is an effect on comparisons of individual watersheds based on this variability.

Figures 2 – 4







Figures 2-4: Cumulative distribution functions of 114 watersheds with respect to rank occurrence for the quintile, septile, and nine-tile analyses

Now that we have the cumulative distribution functions, we will proceed by taking the sum of the values of the cumulative distribution function at each of the ranks. Thus, for quintiles:

$$\mathbf{\Sigma}[F(x)] = F(1) + F(2) + F(3) + F(4) + F(5)$$

septiles:

$$\mathbf{\Sigma}[F(x)] = F(1) + F(2) + F(3) + F(4) + F(5) + F(6) + F(7)$$

and nine-tiles:

$$\mathbf{\Sigma}[F(x)] = F(1) + F(2) + F(3) + F(4) + F(5) + F(6) + F(7) + F(8) + F(9)$$

We refer to each of these sums as a *CDF-index* value. Larger CDF-index values indicate better watershed health. This is the case because a watershed with a large number of rank 1 scores will have a larger CDF-index value than a watershed with a small number of rank 1 scores. Alternatively, the CDF-index (minus 0.5) equals the area under the graph of the CDF in Figures 2 – 4, and higher graphs indicate better health. Thus, the CDF-index provides an objective way of combining multiple indicators into a single composite index of watershed health. After calculating the CDF-index values for each watershed, we have picked out the top ten, the middle ten, and the bottom ten watersheds for each of the ranking options. These values are listed in Tables 2 – 4:

Tables 2 – 4

Quintile Analysis					
Top 10		**Middle 10**		**Bottom 10**	
Watershed	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**
2050203	4.2425	2080106	3.0303	2060003	1.6969
2050202	4.2425	2050302	3.0303	2050306	1.7272
5050005	4.1213	6010206	3.0001	2070008	1.7272
5070201	4.0770	3010205	3.0000	2070009	1.7575
5070101	4.0769	3010202	3.0000	2040203	1.7878
2050205	4.0607	2050206	2.9697	2040205	1.8181
5070202	4.0385	5020003	2.9394	2040202	1.8484
5050007	3.9698	5010007	2.9394	2040105	1.9091
2080201	3.9697	2080110	2.9230	2060006	1.9696
5050009	3.9616	2050107	2.9091	2070010	2.0303

Septile Analysis					
Top 10		**Middle 10**		**Bottom 10**	
watershed	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**
2050202	5.9092	2050304	4.0606	2060003	2.0303
2050203	5.8486	2080107	4.0303	2050306	2.0908
2050205	5.6971	2050206	4.0303	2070008	2.0908
5050005	5.6365	2050302	4.0001	2040205	2.2423
5070202	5.5771	5010007	3.9394	2040201	2.3333
5050007	5.5456	2080104	3.9393	2070009	2.3333
5020004	5.5456	3010205	3.9231	2040203	2.3636
5070101	5.5385	6010206	3.9230	2040105	2.3939

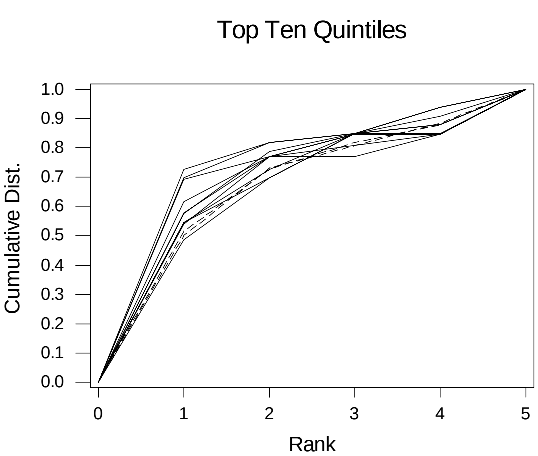
5070201	5.5000	5020003	3.8788	2060006	2.5151
5050009	5.5000	2080206	3.8484	5030102	2.5384

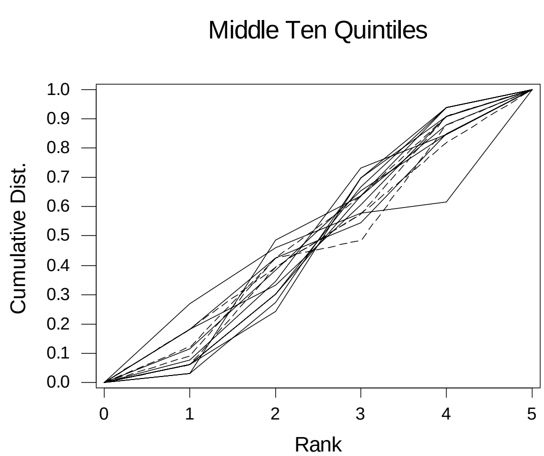
Nine-tile Analysis					
Top 10		**Middle 10**		**Bottom 10**	
watershed	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**	**watershed**	**Σ [F(x)]**
2050203	7.4850	2080107	5.0909	2050306	2.4241
2050202	7.4850	2080106	5.0909	2060003	2.4544
5050005	7.1213	2050304	5.0606	2070008	2.6060
5070101	7.0770	2050206	4.9394	2070009	2.6969
2050205	7.0607	5020003	4.9091	2040201	2.7575
5070202	7.0002	6010206	4.8846	2040203	2.8181
5050007	6.9395	3010202	4.8482	2040205	2.8181
5070201	6.9231	5010007	4.8182	2040105	2.9999
5050009	6.8847	2050303	4.8182	5030102	2.9999
5050006	6.8462	2080104	4.8181	2060006	3.1514

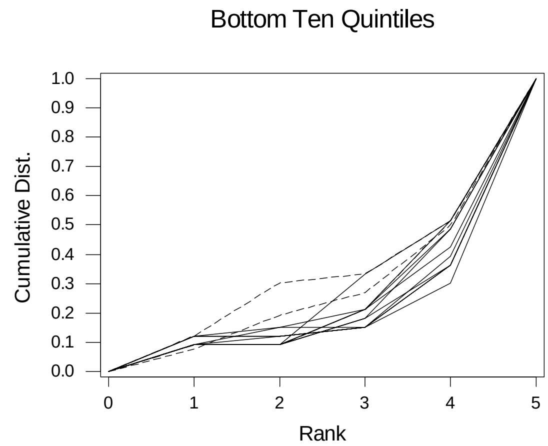
Table 2-4: List of top ten, middle ten, and bottom ten watersheds, and their corresponding CDF-index values for the quintile, septile, and nine-tile analyses

We can also look at the cumulative distribution functions of the top, middle, and bottom ten watersheds in the quintile analysis (Figures 5 – 7). The solid lines represent the top, middle and bottom ten quintiles, and the dotted lines represent watersheds that appeared in the top, middle or bottom ten under the septiles or nine-tiles analysis but not appear under the quintile analysis.

Figures 5 – 7





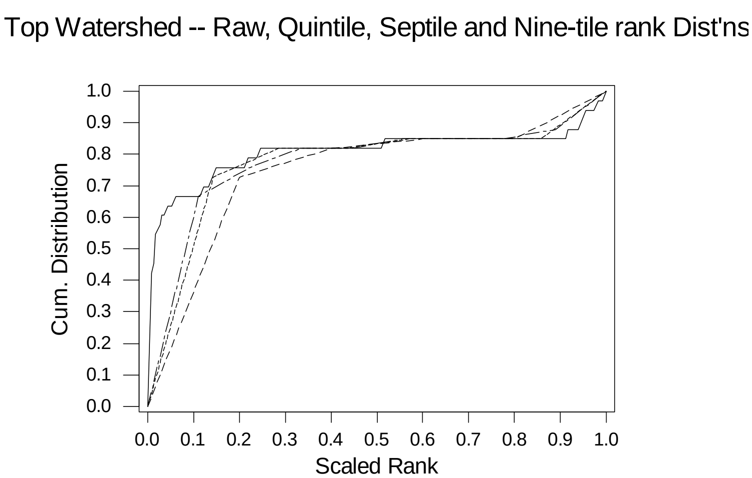


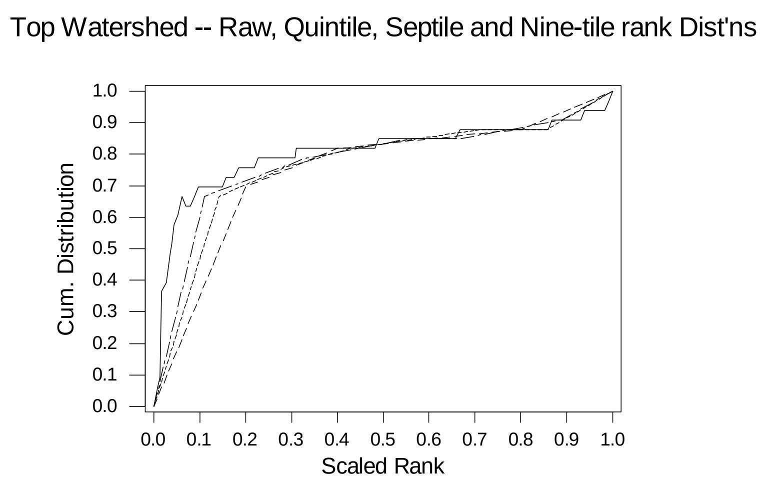
<u>Figures 5-7</u>: Cumulative distribution functions for the top, middle, and bottom ten watersheds based on rank occurrence in the quintile analysis

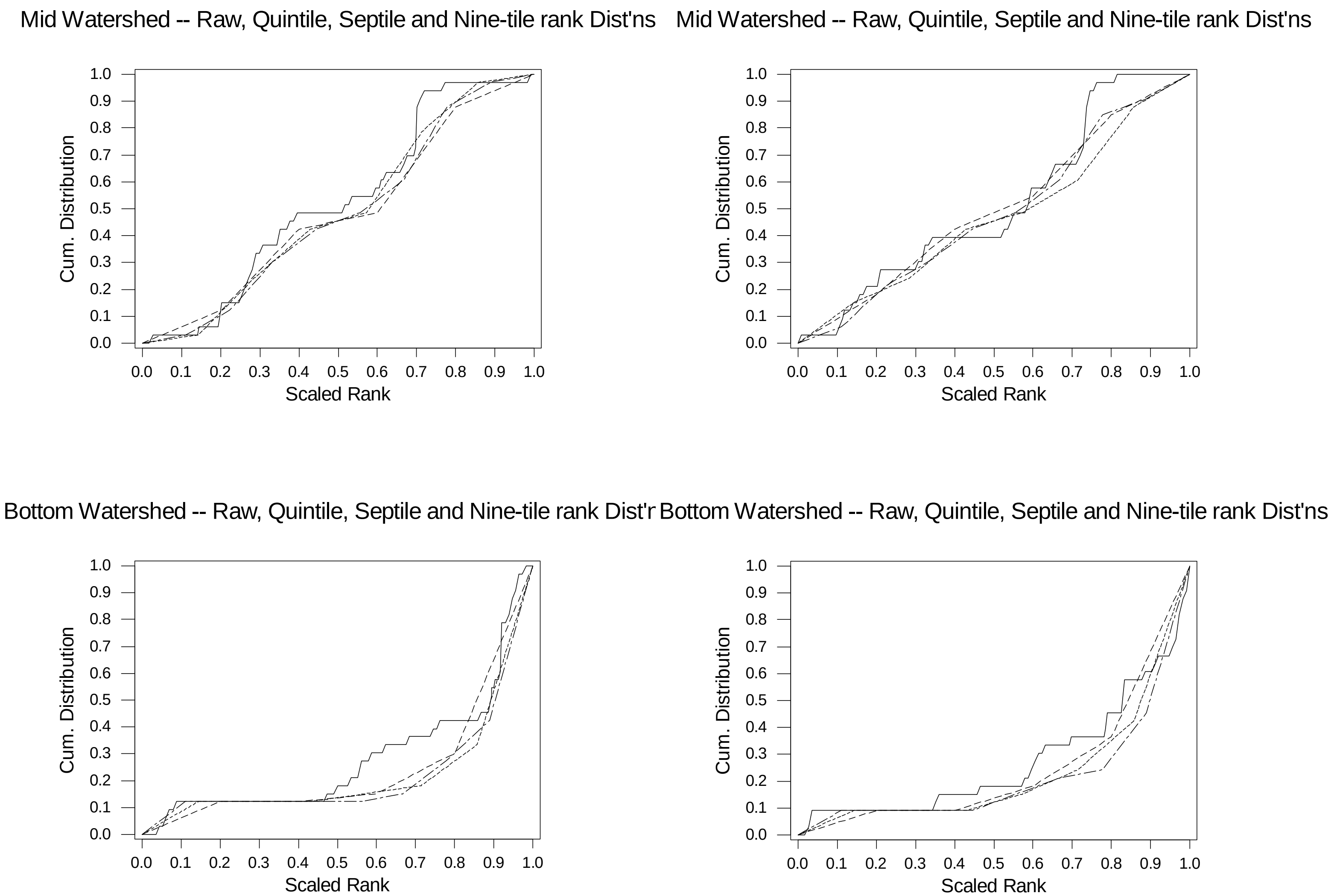
There is a great deal of agreement across the three grouping options in identifying the watersheds the top ten, middle ten, and bottom ten watersheds. Nine of the top ten watersheds appeared in all three lists, six appeared in all three lists for the middle ten, and eight appeared in all three lists for the bottom ten.

We can also look at the graphs of the cumulative distributions for two of the top ten, two of the middle ten, and two of the bottom ten watersheds plotted against their scaled ranks and superimposed on top of each other. These are presented in figures 8 – 13.

<u>Figures 8 – 13</u>





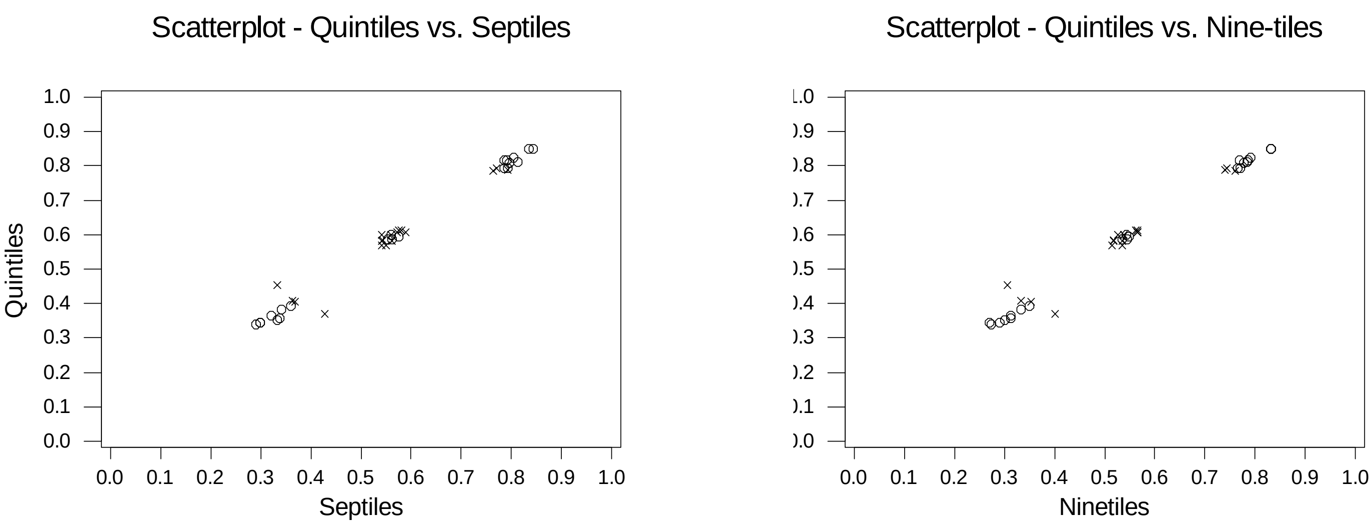


<u>Figures 8-13</u>: These graphs show the watershed with the highest 2 CDF-index values, the middle 2 CDF-index values, and the lowest 2 CDF-index values and plot their cumulative distribution functions for the quintile, septile, and nine-tile analyses against their CDF for the raw rank analysis.

In these figures, the solid lines represent the cumulative distributions for the raw ranks, with each indicator being ranked on a scale of 1-114, with the best value for that indicator receiving a one, and down to the worst receiving a 114. The scaled ranks are the rank values scaled using the transformation 1/N, with N equal to the number of ranks in that grouping option. From these distributions, we can see that the ranked distribution functions are very similar to the raw rank distribution functions, which would indicate that little information is lost by choosing one of the grouping options.

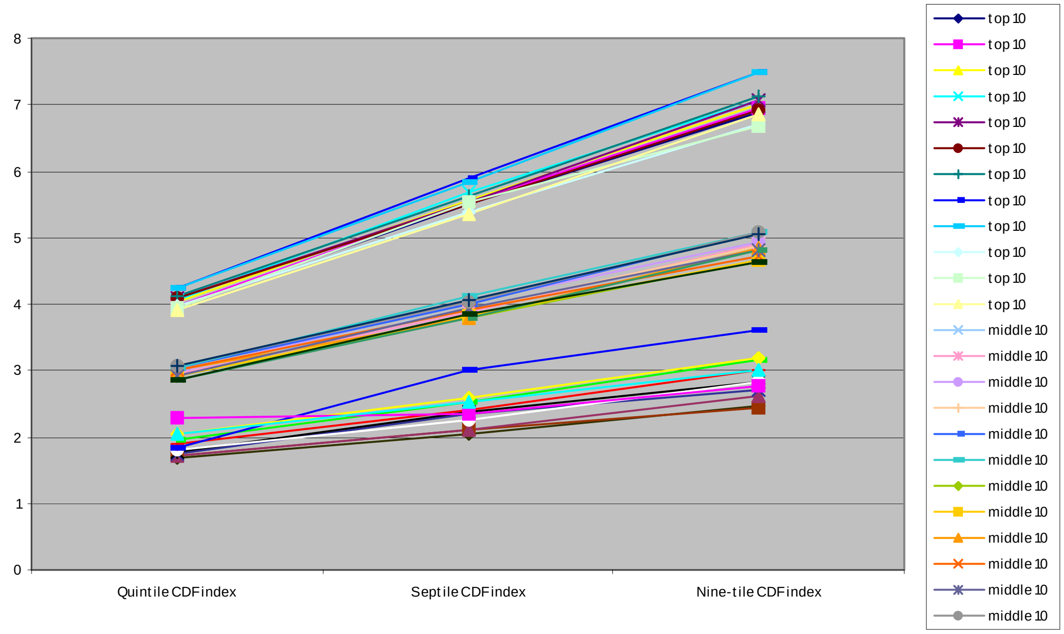
Finally, we can examine scatterplots of the CDF-index values. In these plots, we have transformed the sums listed in the tables above, again using the transformation 1/N, with N again representing the number of ranks within that ranking option. We plotted the scaled quintile sums against both the scaled septile sums and the scaled nine-tile sums. Included in the plots were the watersheds that appeared in any of the top, middle or bottom ten lists in Tables 2 – 4. The plots are shown in Figure 14 – 15:

Figures 14 – 15



Figures 14-15: Scatterplots of quintile values vs. septile values (Figure 14) and quintile values vs. nine-tile values (Figure 15) for those watersheds that were in the top ten watersheds. The circles represent watersheds that appeared in the top ten of both analyses, while the x's represent those watersheds that only appeared in the top ten in one of the plotted analyses.

We also produced a parallel coordinates plot (Wegman, 1990) showing CDF-index values versus grouping method (**Figure 16**):



<u>Figure 16</u>: Parallel coordinates plots of the watersheds that fell into the top ten, middle ten, and bottom ten watersheds based on CDF-index values.

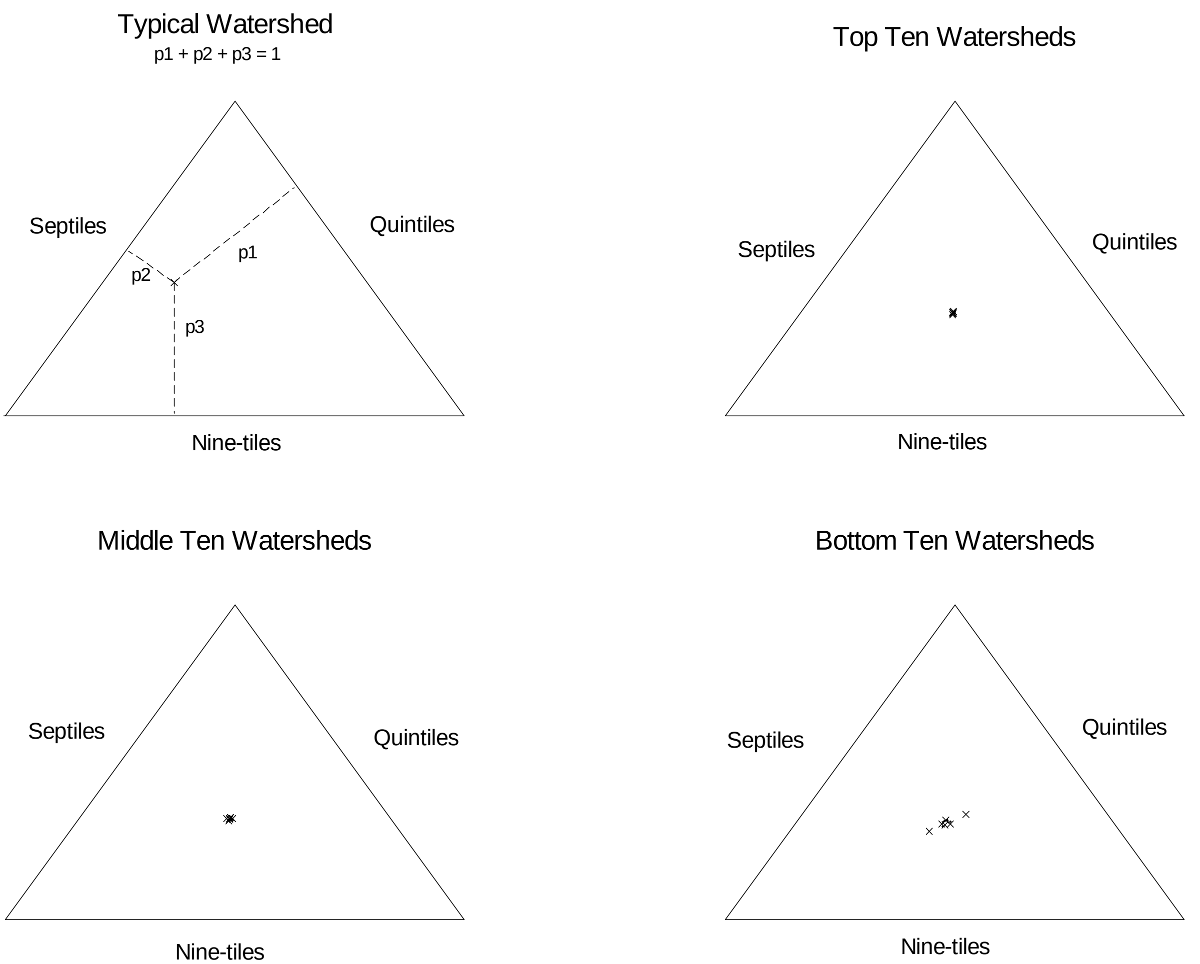
In the parallel coordinates plot, the variables are plotted parallel to one another, rather than the orthogonal plots that consist of the x and y axes. This is another way to visually recognize patterns that appear in the data. Using the quintile, septile and nine-tile CDF-indices, the watersheds that appeared in the top ten, bottom ten, and middle ten lists remain in those groupings, with no intersection of watersheds from one group with another. There are also a minimal number of intersections within the groups.

We have also produced triangular scatter plots showing concordance among the three grouping methods (quintiles, septiles, and nine-tiles). Each watershed has three CDF-index values corresponding to the three grouping methods. These are first scaled to range between 0 and 1, giving x, y, z as the scaled CDF-index values. The triangular coordinates of x, y, z are

$$p_1 = x/(x+y+z)$$
$$p_2 = y/(x+y+z)$$
$$p_3 = z/(x+y+z).$$

The (p_1, p_2, p_3) can be represented as points in an equilateral triangle, giving a *triangular scatter plot*. Dispersion about the centroid of the triangle indicates a lack of concordance among the three grouping methods. These plots, along with an example of a typical watershed, can be seen in Figures 17 – 20.

<u>Figures 17 – 20</u>



<u>Figures 17-20</u>: Triangular scatter plots for a typical watershed, and the top ten, middle ten, and bottom ten watersheds based on CDF-index values

In figures 14 and 15, the best watersheds will have points plotted in the upper right-hand corner, while those with the worst will have points plotted in the lower left hand corner. The scales have been adjusted to better see the data points. There is a strong positive correlation indicating that the better the score in the quintile analysis, the better the score in the septile and nine-tile analysis.

As these figures indicate, there is almost no variation in the watersheds chosen for the top ten, the middle ten, and the bottom ten due to a change in the grouping option. We do notice two watersheds that appear to be outliers in all three graphs, and they are both among the bottom ten watersheds. These were watershed #2040202, which appeared in the bottom ten quintiles and not the other two groupings, and watershed #2040201, which appeared in the bottom ten septile and nine-tile groupings but not the quintiles.

Conclusions

Jones, et. al. (1997) present the raw data in a way that is more comprehensible and manageable for analysis and interpretation. It is done by breaking the data into quintiles. It appears easier, however, to see comparisons, and interpret the data, by using the rank values rather than the raw data, which are often "soft" in any case. By using the quintile ranks instead of the raw indicators, we have values that are on a common scale with a common directionality (rank 1 is always the best).

A major concern in using ranks is potential information loss. We have shown that the watershed comparisons are quite robust to changes in the grouping options.

References

Jones, Bruce K., Ritters, Kurt H., Wickham, James D., Tankersley, Roger D. Jr., O'Neill, Robert V., Chaloud, Deborah J., Smith, Elizabeth R., Neale, Anne C. 1997. *An Ecological Assessment of the United States Mid-Atlantic Region: A Landscape Atlas*. EPA 600-R-97-130, United States Environmental Protection Agency, Office of Research and Development, Washington, D.C.

Wegman, Edward J. 1990. "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association* 85, 664-675.

Environmental Protection Agency, (1999). Data on "Additional Information about Watershed Indicators," downloaded from http://www.epa.gov/maia/html/la-tablea1.html.

Appendix

Figure 1

POPDENS	Population density (number of people per square kilometer)
POPCHG	Population change (percentage change from 1970 to 1990)
UINDEX	Human use index (proportion of watershed area with agriculture or urban land cover)
RDDENS	Road density (average number of kilometers of roads per square kilometer of watershed area
NO3DEP	Average annual wet deposition of nitrate (1987 and 1993)
SO4DEP	Average annual wet deposition of sulfate (1987 and 1993)
OZAVG	Average annual value of the W126 ozone index (1988 and 1989)
RIPFOR	Proportion of total streamlength with adjacent forest land cover
RIPAG	Proportion of total streamlength with adjacent agriculture land cover
STRD	Proportion of total streamlength that has roads within 30 meters
DAMS	Number of impoundments per 1000 kilometers of stream length
CROPSL	Proportion of watershed with crop land cover on slopes that are greater than three percent
AGSL	Proportion of watershed with agriculture land cover on slopes that are greater than three percent
STNL	Potential nitrogen loading to streams
STPL	Potential phosphorus loading to streams
PSOIL	Proportion of watershed with potential soil loss greater than one ton per acre per year
FOR%	Percent of watershed area that has forest land cover
FORFRAG	Forest fragmentation index
EDGE7	Proportion of watershed area with suitable forest edge habitat (7 hectare scale)
EDGE65	Proportion of watershed area with suitable forest edge habitat (65 hectare scale)
EDGE600	Proportion of watershed area with suitable forest edge habitat (600 hectare scale)
INT7	Proportion of watershed area with suitable interior forest habitat (7 hectare scale)
INT65	Proportion of watershed area with suitable interior forest habitat (65 hectare scale)
INT600	Proportion of watershed area with suitable interior forest habitat (600 hectare scale)
INTALL	Proportion of watershed area with suitable interior forest habitat at three scales
FORDIF	Departure of the largest forest patch size from the maximum possible for given amount of anthropogenic land cover
NDVIDEC	Decrease in normalized difference vegetation index from 1975 to 1990
NDVIINC	Increase in normalized difference vegetation index from 1975 to 1990
NDVITOT	Total change in normalized difference vegetation index from 1975 to 1990
1STDEC	Difference between observed and expected decreases in normalized difference vegetation index from 1975 to 1990 in first-order stream regions
1STINC	Difference between observed and expected increases in normalized difference vegetation index from 1975 to 1990 in first-order stream regions
1STTOT	Difference between observed and expected total change in normalized difference vegetation index from 1975 to 1990 in first-order stream regions
NDVI3%	Proportion of watershed with normalized difference vegetation index decreases from 1975 to 1990 on slopes greater than three percent

I