Jour. Ind. Soc. Ag. Statistics
57 (Special Volume), 2004 : 191-207

Assessing the Similarity between Neighboring Watersheds and Evaluating the Effect of Watershed Size on Conditional Entropy Profiles[1]

Amy C. Burnicki, G.P. Patil, and Charles Taillie
The Pennsylvania State University, University Park, PA, USA

SUMMARY

The assessment and monitoring of landscape characteristics at different scales has become an increasingly important issue in landscape ecology. And multiscale landscape fragmentation of a watershed is an important landscape characteristic. It is effectively measured by conditional and marginal entropy profiles of the watershed. This paper examines how the conditional and marginal entropy profiles of a large watershed relate to the profiles for the medium watersheds contained within its boundaries. It also examines whether large watersheds are comprised of medium watersheds that are similar to each other in terms of their landscape fragmentation.

Key words : Multiscale landscape fragmentation, Conditional entropy profiles, Marginal entropy profiles, Medium watersheds, Large watersheds, Similarity index.

1. Introduction

The assessment and monitoring of landscape characteristics at different scales has become increasingly important over the last ten years. Some of the new tools and techniques quantifying, classifying and mapping landscape fragmentation using entropy profiles appear in Johnson [3], Johnson *et al.* ([4], [5]), Patil [13] and Patil and Taillie ([16], [17]). These profiles are computed from landcover raster maps. In the case of Pennsylvania, the landcover maps were derived from satellite data in the form of 6-band LANDSAT thematic images. These images measure the reflectance of 30 meter by 30 meter pixel land areas, which helps to indicate the type of landcover existing in that area. Eight landcover categories were used: water, conifer forest, broadleaf forest,

[1] Prepared with partial support from the National Science Foundation Grant for Water and Watersheds Program. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

mixed forest, annual herbaceous (cropland), perennial herbaceous (pastureland, grassland), heterogeneous vegetation, and barren (pavements, buildings, etc.). Pennsylvania was subdivided into watersheds and a landcover map was obtained for each watershed (see Appendix 1a and 1b). An entropy profile characterizes landscape fragmentation in a watershed through the spatial variation of landcover categories across the watershed and is dependent upon the resolution, or scale, of the pixels (Johnson *et al.* [5]). The entropy profile depicts the relationship between resolution and entropy and, based on this relationship, watersheds can be classified into three groups: mostly forested, transitional, and subject to high degrees of agriculture or development. Thus, the profiles can help in determining the relative environmental and ecological health of the watershed (Johnson *et al.* [6]).

The calculated entropy is not that of the marginal landcover distribution because this would measure non-dominance of the landcover types, but would convey no information about the spatial distribution of landcover across the region (Johnson [3]). To capture the spatial pattern, we consider landcover in 2×2 arrangements of contiguous pixels (called 4-tuples). With eight landcover types, there are $8^4 = 4096$ possible configurations of landcover across a 4-tuple. By slicing the map into 4-tuples and counting the occurrence of each possible configuration, we obtain a frequency table with 4096 rows (cells) called the 4-tuple landcover distribution. If landcover were spatially aggregated across the landscape (i.e., large patches), then landcover would be homogeneous across most 4-tuples and the eight homogeneous rows would dominate in the 4-tuple landcover distribution. Accordingly, entropy of the 4-tuple distribution would be small. Conversely, a highly fragmented landscape would lead to a more uniform 4-tuple landcover distribution and larger entropy. Thus, entropy of the 4-tuple landcover distribution is a direct measure of fragmentation at the finest resolution.

In order to assess fragmentation at broader scales, we degrade the original landcover map by aggregating pixels. Specifically, each 4-tuple of pixels in the original map becomes a single pixel four times as large in the degraded map. Landcover is assigned to the aggregate pixel by random selection of one of its four subpixels. Slicing the map into 4-tuples of aggregate pixels then leads to a 4-tuple landcover distribution at the aggregate scale. The frequencies are random variables due to the random selection of subpixels. However, we are able to calculate the expected values of these random frequencies to give a numerical 4-tuple landcover distribution, and hence an entropy for the aggregate scale. The aggregation process can be continued, giving a sequence of successively coarser resolutions labeled as 1, 2, 3, .., with "1" indicating the original finest resolution. Attached to each resolution is a 4-tuple frequency table and corresponding entropy value. The entropy profile is the plot of these entropy values against resolution.

In fact, two different entropy profiles are available. The absolute (or marginal) entropy profile uses entropy of the 4-tuple landcover distributions. The conditional entropy profile uses the entropy of the 4-tuples conditional upon the landcover assigned to their aggregate pixels. More precisely, at any resolution, consider the configuration of landcover types assigned jointly to the members of a 4-tuple and to their aggregate pixel. This gives a two-way frequency table with 4096 rows (labeled as '4-tuples') and 8 columns (labeled as 'aggregates'). As above, these frequencies are random but we are able to calculate the expected frequencies. In the resulting two-way table, the conditional entropy, H('4-tuples'|'aggregates'), is used for the conditional entropy profile. The absolute entropy profile is conceptually and computationally simpler. However, conditional entropy is indicative of predictability of fine scale spatial pattern given coarser scale features and becomes of interest if one were contemplating actual multi-resolution measurements (Johnson *et al.* [7]).

Currently, we are examining two different subdivisions of Pennsylvania into watersheds. The first subdivision contains 102 watersheds and is employed by the Pennsylvania State Water Management Plan. The second subdivision contains 46 watersheds, as delineated by the United States Environmental Protection Agency in its watershed assessment of the Mid-Atlantic Region of the U.S. (Jones *et al.* [11]). Most of these 46 watersheds are unions of the watersheds found in the State water plan. Watersheds in the State's Management Plan will be referred to as 'medium' watersheds. The EPA delineated watersheds will be referred to as 'large' watersheds. Entropy calculations were made through eight resolutions for the medium watersheds and ten resolutions for the large watersheds.

This paper examines how the conditional and marginal entropy profiles of a large watershed relate to the profiles for the medium watersheds contained within its boundaries. It also examines whether large watersheds are comprised of medium watersheds that are similar to each other in terms of their landscape fragmentation.

2. Conditional Profile Comparisons

For each large watershed comprised of more than one medium watershed, the conditional entropy profile for the large and the medium watersheds contained within it were plotted. Profiles for twenty-seven large watersheds were plotted in this manner. Furthermore, an area-weighted average conditional entropy profile was also graphed. This weighted average profile was calculated by multiplying the conditional entropy value for an individual medium watershed by its area, summing this value for each medium watershed, and dividing by the total area of the large watershed. Of the 27 large watersheds examined, only two displayed weighted averages that were greater than the conditional entropy profile for the large watershed. The two exceptions were

watersheds 23 and 25. Furthermore, an additional seven large watersheds displayed a crossing of the weighted and large conditional entropy profiles at one or two resolutions. These crossings mainly occurred at the first two resolutions (1, 2) and the highest two resolutions (7, 8). However, the majority of the plots showed that the large conditional entropy profile was above that of the weighted average for the medium watersheds contained within. The following figures illustrate this result for four selected large watersheds. The complete set of conditional entropy profile comparisons can be found in Appendix 2 of Burnicki *et al.* [1].

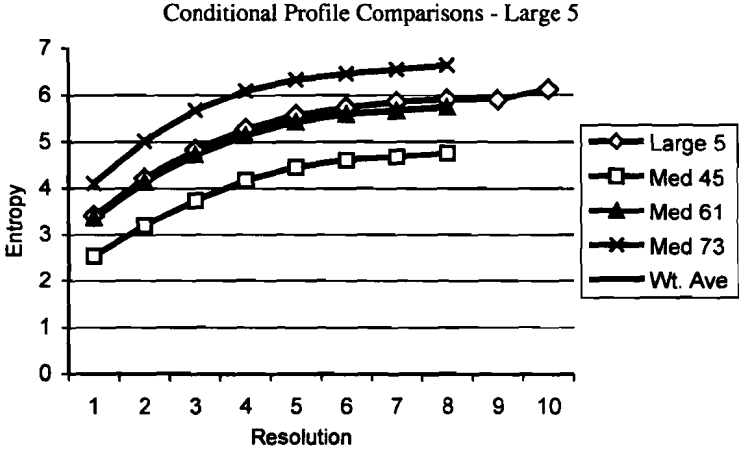

Fig. 1. Conditional Entropy Profile Comparisons for Large Watershed 5

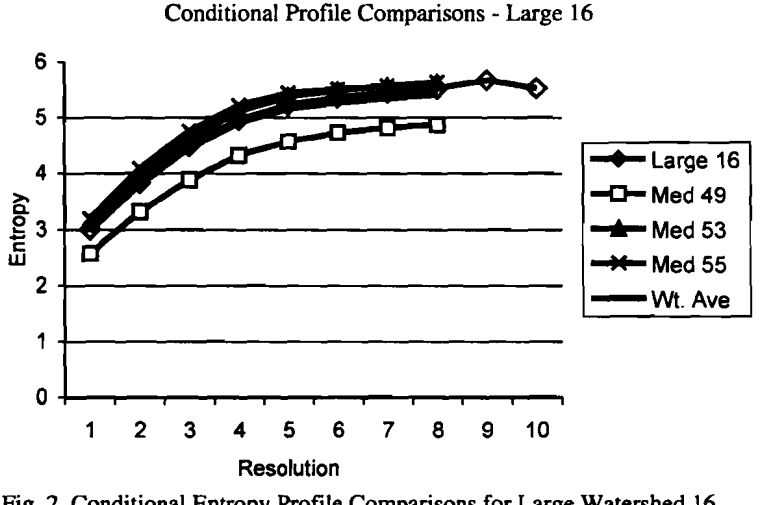

Fig. 2. Conditional Entropy Profile Comparisons for Large Watershed 16

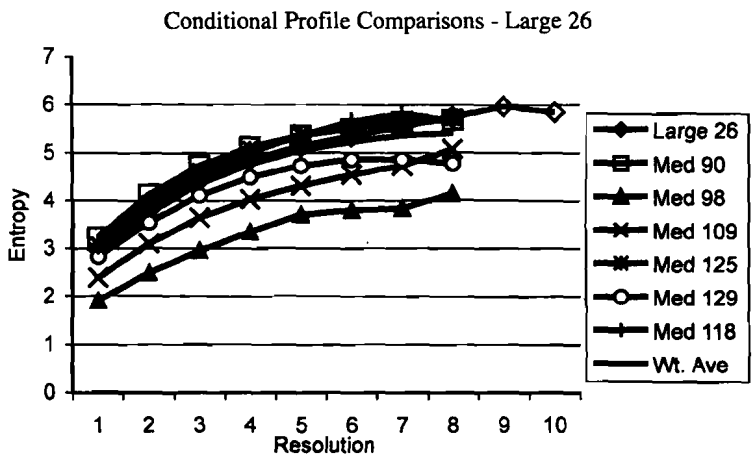

Fig. 3. Conditional Entropy Profile Comparisons for Large Watershed 26

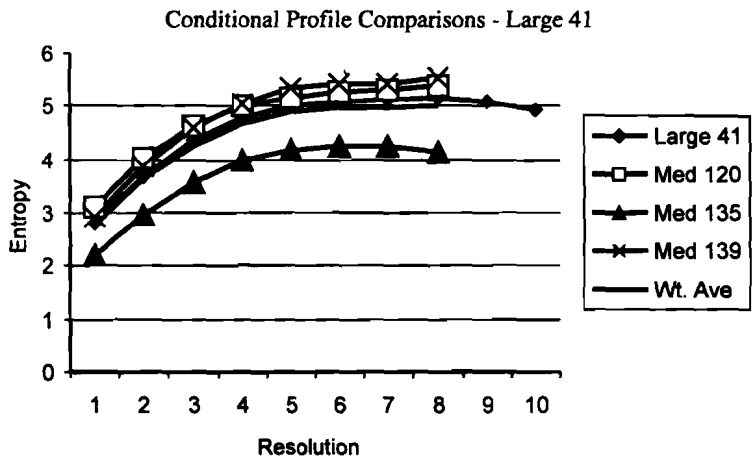

Fig. 4. Conditional Entropy Profile Comparisons for Large Watershed 41

3. Marginal Profile Comparisons

Similarly, the same graphs were generated for the 27 large watersheds using the marginal entropy values. An examination of these plots revealed that only large watershed 25 had the weighted average profile above that of the large marginal entropy profile. Furthermore, large watersheds 9, 23, 33, 35, and 38 had crossings of the weighted average and the large entropy profiles at only one or two resolution levels. Similar to the conditional entropy profile comparisons, these crossings occurred mostly at resolutions 1 and 2 and resolutions 7 and 8. Thus, in the case of the marginal entropy profiles, fewer crossings were present. Again, the majority of the graphs show that the weighted average entropy profile

falls below that of the large marginal entropy profile. The following figures examine the same four watersheds illustrated in the previous section, but display their marginal profile comparisons. The complete set of marginal entropy profile comparisons can be found in Appendix 3 of Burnicki *et al.* [1]).

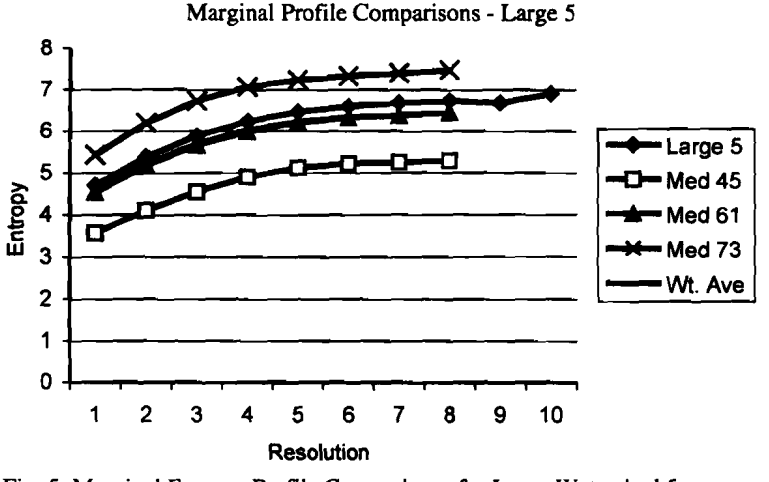

Fig. 5. Marginal Entropy Profile Comparisons for Large Watershed 5

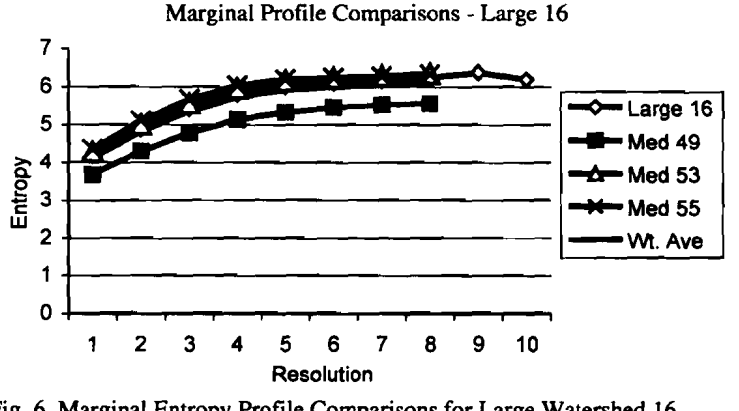

Fig. 6. Marginal Entropy Profile Comparisons for Large Watershed 16

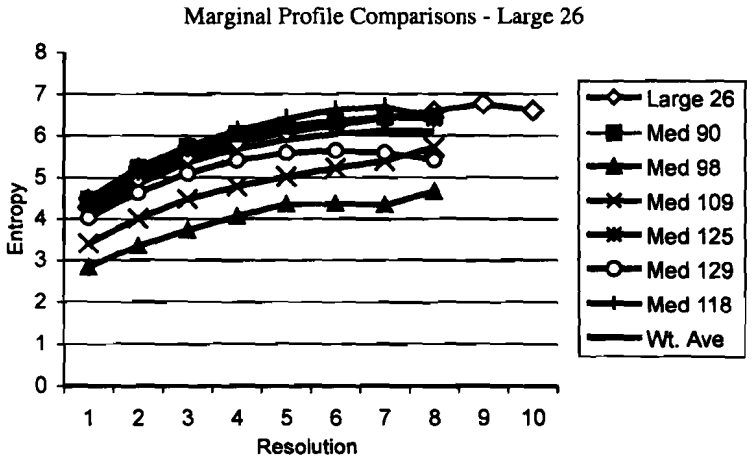

Fig. 7. Marginal Entropy Profile Comparisons for Large Watershed 26

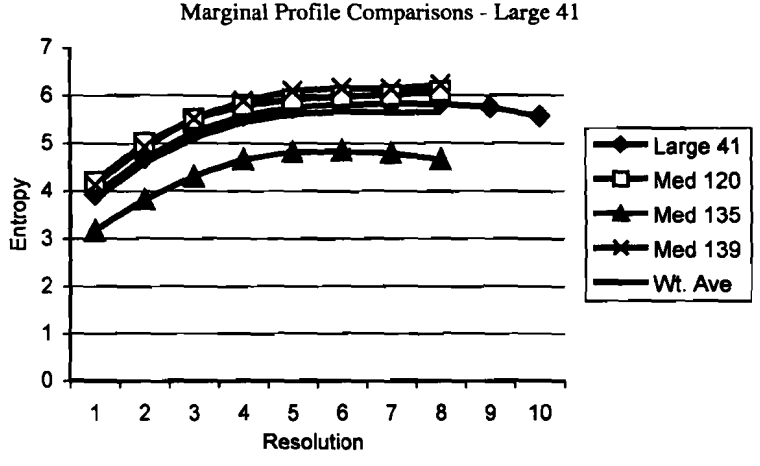

Fig. 8. Marginal Entropy Profile Comparisons for Large Watershed 41

4. Mathematical Explanation

Our empirical finding for most of the large watersheds is that their entropy profile is above the area-weighted profile for their medium sub-watersheds. We use Jensen's inequality to give a mathematical explanation for this phenomenon. First, we recall Jensen's inequality. If $f(x)$ is a concave function of $k-1$ variables and if X is a $(k-1)$-dimensional random vector, then $E[f(X)] \leq f(E[X])$. Further, if $f(x)$ is strictly concave, then the inequality is strict unless X is degenerate at a single point.

The entropy function is strictly concave, so Jensen's inequality can be applied. Consider a table in which the rows are the 4096 different 4-tuple configurations. We have a column T_L of frequencies for the large watershed and columns $T_1, T_2, \ldots T_a$ for each of the medium sub-watersheds. For now, consider absolute instead of relative frequencies. Let $T_s = T_1 + T_2 + \ldots T_a$. Every 4-tuple of pixels in a medium watershed is also a 4-tuple in the large watershed. The converse is not quite true, however, since certain 4-tuples in the large watershed may overlap several medium watersheds. Thus, we can only say that T_s and T_L are approximately the same. Now, let $C_L, C_1, \ldots C_a, C_s$ be the columns of relative frequencies. The i^{th} component of C_s is

$$C_{si} = \frac{T_{si}}{T_{s\cdot}} = \frac{T_{1i} + T_{2i} + \ldots T_{ai}}{T_{s\cdot}}$$

$$= \frac{T_{1\cdot}}{T_{s\cdot}} \frac{T_{1i}}{T_{1\cdot}} + \ldots + \frac{T_{a\cdot}}{T_{s\cdot}} \frac{T_{ai}}{T_{a\cdot}}$$

$$= \frac{T_{1\cdot}}{T_{s\cdot}} C_{1i} + \ldots + \frac{T_{a\cdot}}{T_{s\cdot}} C_{ai}$$

Thus, C_s is the area-weighted average of $C_1, \ldots, C_a$. Since C_L and C_s are approximately equal, Jensen's inequality implies that

$$H(C_L) \approx H(C_s) \equiv H(\overline{C}) \geq \overline{H(C)}$$

where H is the entropy function and 'bar' denotes the area-weighted average. Furthermore, the last inequality is strict unless $C_1, C_2 \ldots, C_a$ are all equal, i.e. unless the medium watersheds within the large watershed have the same 4-tuple landcover distribution. Consequently, the difference $H(C_s) - \overline{H(C)}$, is a measure of landscape pattern heterogeneity within the large watershed and at the scale of the medium watersheds.

5. Examination of Result

Both the marginal and conditional entropy profile comparisons supported this result suggested by Jensen's inequality. As seen previously, the weighted average entropy profile for the medium watersheds was located below the entropy profile for the large watershed. Since the proof was based on the marginal entropy function, the marginal profile comparisons showed fewer deviations from this result. A possible source of error affecting both sets of profile comparisons can be attributed to the construction of the large watershed. The entropy of each large watershed reflects the entropy contained within the averaging of the medium watersheds plus additional entropy due to overlapping that occurs when mapping adjacent medium watersheds. Because the satellite

maps 30 meter by 30 meter square pixels, pixels located on the boundaries of a medium watershed sometimes extend into neighboring watersheds.

6. Motivation

The motivation behind examining the relationship between large and medium watersheds is to determine whether the medium watersheds were grouped so that a large watershed was comprised of medium watersheds that were similar in terms of landscape fragmentation. Since conditional entropy profiles are used to characterize the fragmentation pattern that is present within a watershed, the constituent medium watersheds should have entropy profile values that are similar to one another. Each conditional entropy profile can be summarized with respect to three parameters: (1) a = amount of information lost as the resolution becomes coarser, (2) b = rate of information loss as the resolution becomes coarser, and (3) c = asymptotic maximum conditional entropy attainable (Johnson *et al.* [6]). These three defining parameters are illustrated by Figure 9.

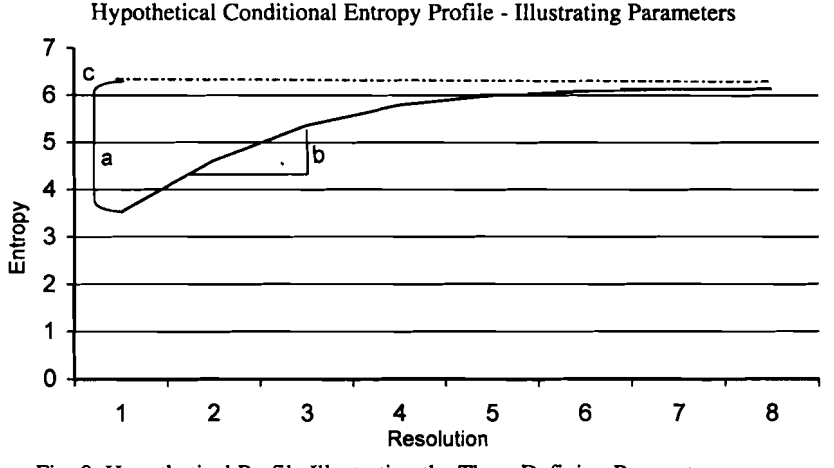

Fig. 9. Hypothetical Profile Illustrating the Three Defining Parameters

Thus, similar medium watersheds should have similar values for these three parameters. In order to determine if the watersheds contained within a large watershed have greater similarity to each other than to watersheds contained in another large watershed, a cluster analysis was performed on the conditional a, b, c values. Only large watersheds that were comprised of more than one medium watershed were included in the analysis. Additionally, medium watersheds that overlapped several large watersheds were excluded. A total of 81 medium watersheds were included in the analysis. The clustering was performed on standardized variables using average linkage and Euclidean distance. Euclidean distance uses the square root of the sum of squared

differences to define the distance between any two observations. The linkage method defines how the distance between two clusters is calculated. Average linkage uses the mean distance between an observation from one cluster and an observation from another cluster. The resulting dendrogram, which illustrates clusters through branches that join together at calculated levels of similarity, is displayed in Figure 10 (Johnson and Wichern [10]).

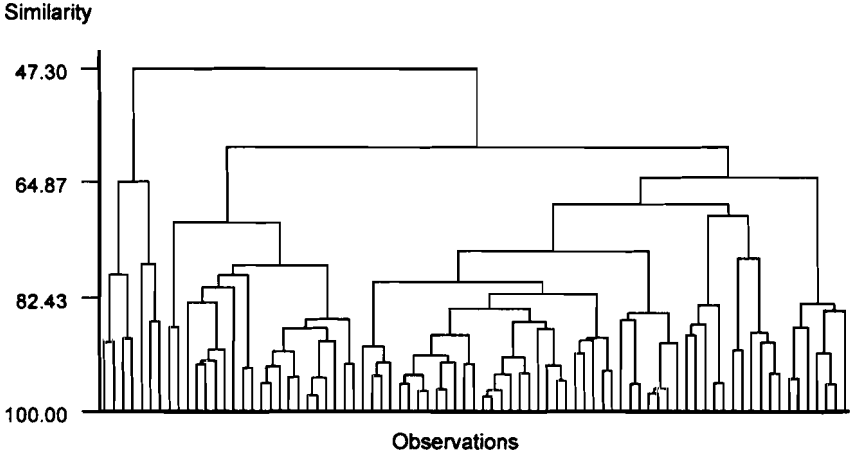

Fig. 10. Dendrogram from Cluster Analysis on Conditional a, b, c values.

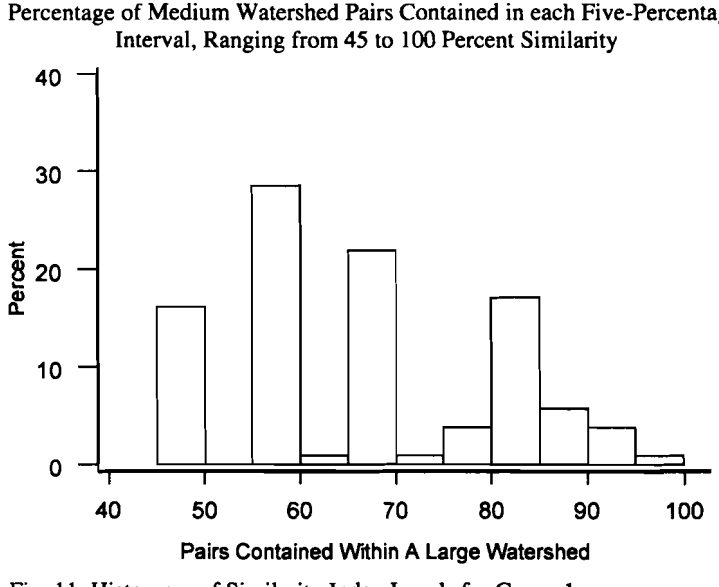

Fig. 11. Histogram of Similarity Index Levels for Group 1.

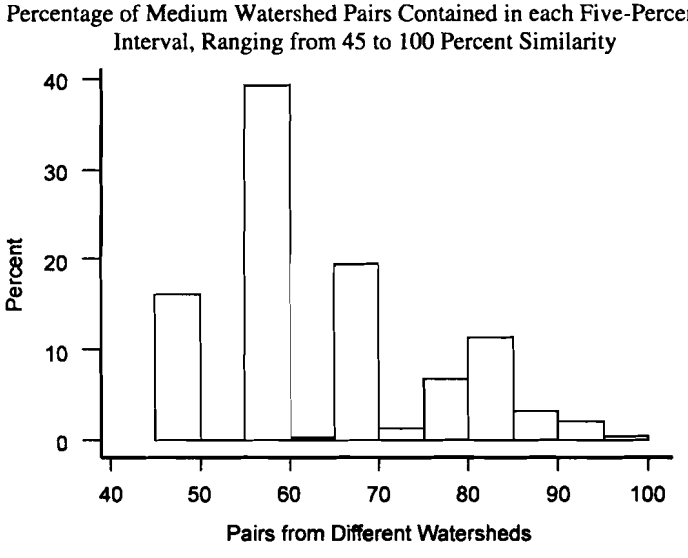

Fig. 12. Histogram of Similarity Index Levels for Group 2

For each pair of medium watersheds in the analysis, a similarity level between the two medium watersheds, as computed by the cluster analysis, was determined. The 3240 medium watershed pairs were then further divided into two groups: (1) pairs with both members contained within the same large watershed, and (2) pairs with members contained within two different large watersheds. Thus, if the large watersheds are comprised of similar medium watersheds, the similarity levels of group 1 should be higher than those of group 2. To assess this hypothesis, the percentage of pairs having similarity index levels in five-percentage point intervals were plotted in Figures 11 and 12 for each of the two groups.

Comparing these two figures, the first notable difference involves the percentage of medium watershed pairs that have similarity levels greater than 80%. One can observe that pairs that are contained within the same large watershed, group 1, have slightly higher percentages for each five-point interval ranging from 80 to 100 per cent similarity. Second, medium watershed pairs form different large watersheds, group 2, have a larger percentage of pairs with similarity levels less than 60%.

An alternate means for comparing these two groups is through the use of inverse cumulative relative frequency distributions. For each group, the relative frequencies in the five-percentage point intervals were determined. The relative frequencies for each interval were added successively to arrive at the cumulative relative frequency distribution. The inverse cumulative relative frequency distribution is simply one minus the cumulative distribution. Figure 13 compares the resulting distributions for each paired watershed group. The relative

frequencies for each five-percentage point interval and the inverse and cumulative relative frequencies distributions for both groups are found in Appendix 7 of Burnicki *et al.* [1].

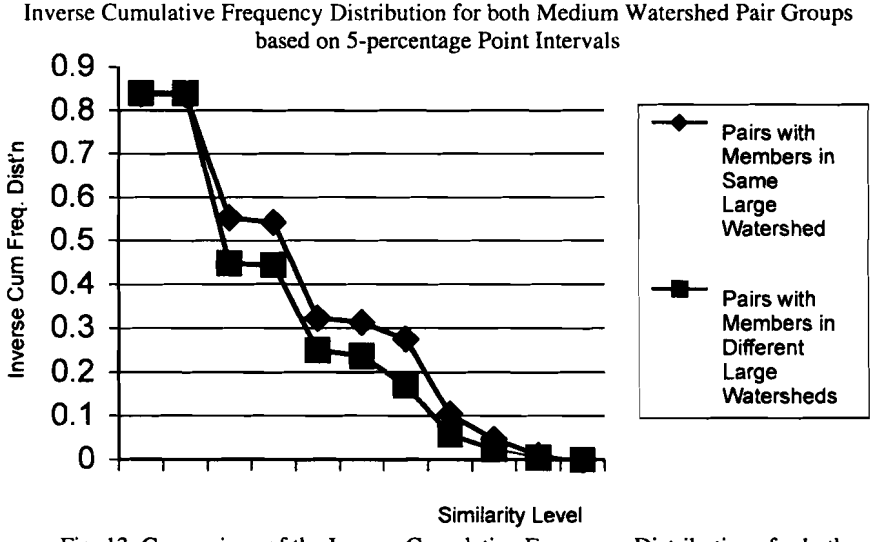

Fig. 13. Comparison of the Inverse Cumulative Frequency Distributions for both Watershed Groups.

Examining this figure, one can observe that the distribution for the pairs with members in two different watersheds lies below the distribution for the pairs with members in the same large watershed. This illustrates that the inverse cumulative relative frequency distribution for the pairs with members in the same large watershed is stochastically greater than the distribution for the pairs with members in two different large watersheds. In other words, the probability of a pair with members in the same large watershed having a similarity level above 55% is greater than the same probability for a pair with members in two different large watersheds. A two-sample Kolmogorov-Smirnov test was conducted to determine if these two sets of similarity levels for groups 1 and 2 could have come from the same distribution. At a significance level of 0.05, one can conclude that the cumulative distributions for the two groups were significantly different (KS test statistic = 0.1372, p-value = 0.04). Thus, the two sets of similarity levels do not come from a similar distribution. This result, as well as the inverse cumulative relative frequency graph, support the results found previously that the similarity levels from the two groups are significantly different and the levels are higher for pairs that have members in the same large watershed.

The above p-value of 0.04 is to be taken as indicative only. This is because the null distribution for the two-sample Kolmogorov-Smirnov test is derived under the assumptions that (1) each sample is a random sample and (2) the samples are independent. Here, both assumptions are violated since two pairs of

watersheds may have a watershed in common, thereby inducing dependencies among the similarity values used in the analysis. Additionally, the p-value was calculated using an S-Plus function for two-sided alternatives. Here, a one-sided alternative would be more natural and appropriate. Although the S-Plus function does not return one-sided p-values, the on-line help guide states that the one-sided p-value is approximately one-half of the two-sided value, i.e. $0.04/2 \cong 0.02$.

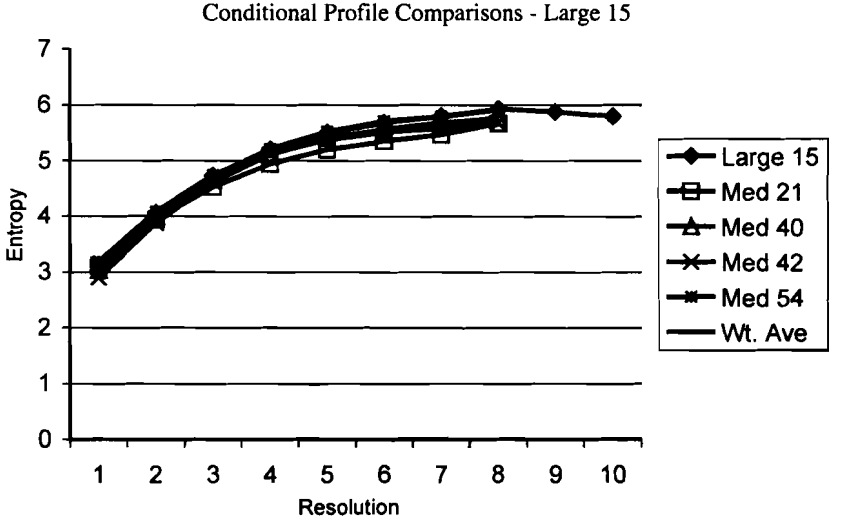

Fig. 14. Conditional Entropy Profile Comparisons for Large Watershed 15.

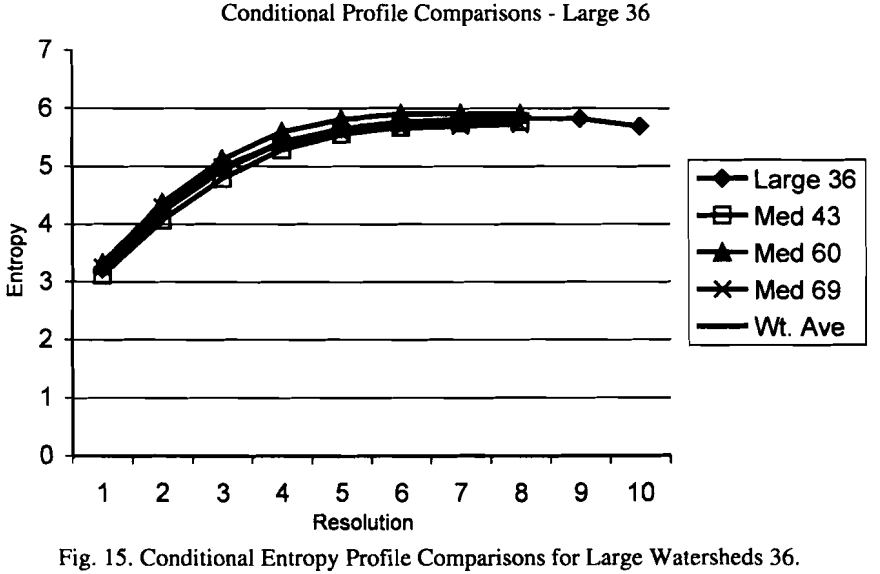

Fig. 15. Conditional Entropy Profile Comparisons for Large Watersheds 36.

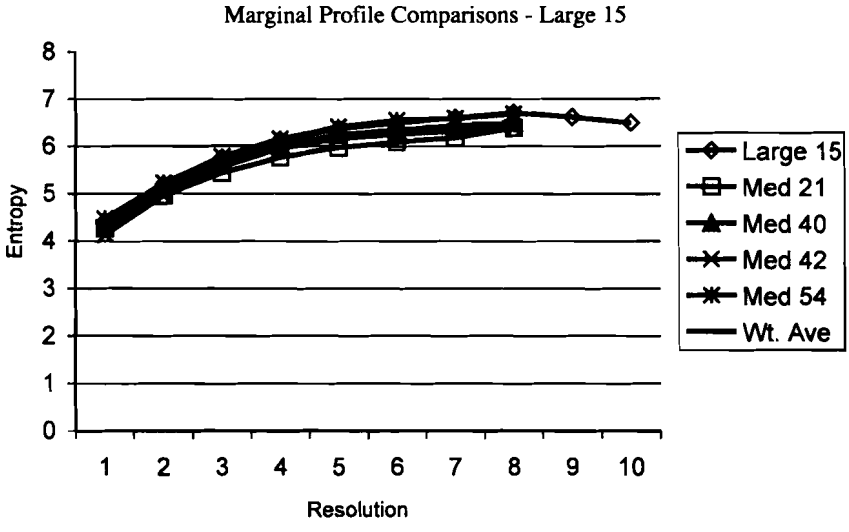

Fig. 16. Marginal Entropy Profile Comparisons for Large Watershed 15.

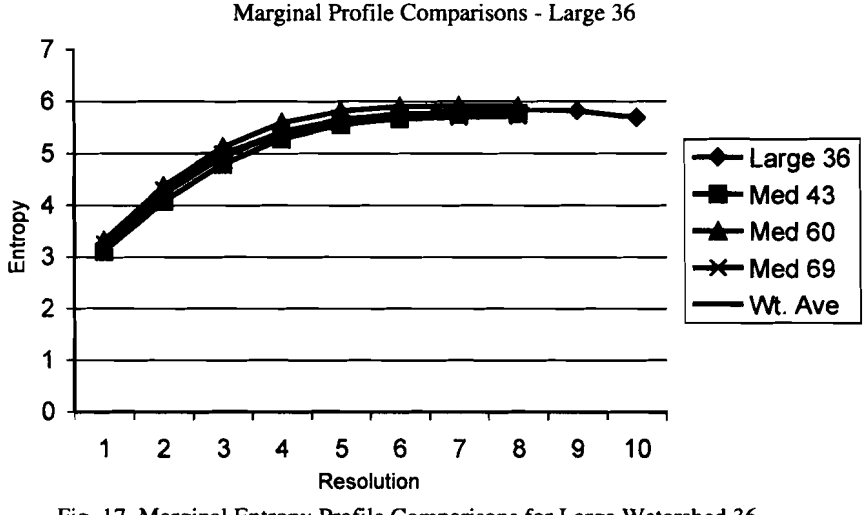

Fig. 17. Marginal Entropy Profile Comparisons for Large Watershed 36.

comparison plots previously investigated, one can observe that the watersheds with the greatest distance between medium watershed entropy profiles correspond to situations where the separation between the large and weighted average profile is easier to determine. For example, comparing the figures for large watersheds 16 and 26, the separation between the large and weighted average entropy profiles for large watershed 26 is easier to observe. It is also observed that watershed 26 displays the greater amount of variation in its constituent medium watersheds. Additionally, if one examines all 27 large

watershed comparisons, 18 out of the 27 plots show small separation between the large and weighted average profiles, thus suggesting that a majority of the large watersheds are comprised of similar medium watersheds. Figures 13-16, illustrate this result for a pair of conditional and marginal profile comparisons.

7. Multivariate Analysis of Variance

A multivariate analysis of variance procedure was used to test whether the mean conditional a, b, c values differed among the large watersheds. The following model was fit to the data: $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1, \ldots, 27$ and $j = 1, \ldots, n_i$, where n_i = total number of medium watersheds contained within large watershed i and α_i = effect due to large watershed. The results of the multivariate analysis of variance, as well as the results of the univariate analysis of variance on each conditional a, b, c value separately and the between and within sum of squares matrices, are found in Appendix 8 of Burnicki *et al.* [1].

Using Wilk's test statistic, which had a corresponding p-value of approximately 0.00, one can conclude that there is a significant effect due to large watershed. In other words, the mean conditional a, b, c values differ among the large watersheds. Additionally, a comparison between the between and within sum of squares matrices revealed that the between sum of squares matrix is larger. This implies that more variation in the conditional a, b, c values occurs between the large watersheds and not within them. Thus, in terms of conditional a, b, c values, the medium watersheds that comprise the large watersheds are more alike then they are different. This supports the results found in the previous section.

Regarding the marginal Analysis of Variance procedures, the p-value for each of the conditional a, b, c, values is highly significant indicating that each is an effective discriminator of the large watersheds. It is interesting to note that the p-value for c – a is smaller than that of either c or a separately. Referring to Figure 9, c – a is the 4-tuple entropy at the finest scale of resolution.

8. Direction for the Future

There is still a great deal of work to be done examining the issue of watershed size/extent. One future area of study involves the determination of whether the averaging process inherent in the formation of the large watershed units results in watershed units that are more distinct from one another. A second area of interest concerns how the grouping/clustering of watersheds across Pennsylvania changes when the medium versus large watershed units are used.

REFERENCES

[1] Burnicki, A.C., Patil, G.P., and Taillie, C. (2001). Assessing the similarity between neighboring watersheds and evaluating the effect of watershed size on conditional entropy profiles in the State of Pennsylvania. Technical Report 2001-0601, Center for Statistical Ecology and Environmental Statistics, Penn. State University, University Park, PA.

[2] Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York.

[3] Johnson, G.D. (1999). Landscape Pattern Analysis for Assessing Ecosystem Condition: Development of a Multi-resolution Method and Application to Watershed Delineated Landscapes in Pennsylvania. Ph.D. Thesis, Penn. State University, University Park, PA.

[4] Johnson, G.D., Myers, W.L., Patil, G.P., O'Connell, T.J., and Brooks, R.P. (2003). Predictability of bird community-based ecological integrity using landscape measurements. In : *Managing for Healthy Ecosystems*, (Eds. D.J. Rapport, W.L. Lasley, D.E. Rolston, N. Ole Nielsen, C.O. Qualset, A.B. Damania) Lewis Publishers, Boca Raton, 617-659.

[5] Johnson, G.D., Myers, W.L., Patil, G.P. and Taillie, C. (1998). Multiresolution fragmentation profiles for assessing hierarchically structured landscape patterns. *Ecological Modelling*, **116**, 293-301.

[6] Johnson, G.D., Myers, W.L., Patil, G.P. and Taillie, C. (2001a). Characterizing watershed-delineated landscapes in Pennsylvania using conditional entropy profiles. *Landscape Ecology*, **16**, 597-610.

[7] Johnson, G.D., Myers, W.L., Patil, G.P. and Taillie, C. (2001b). Fragmentation profiles for real and simulated landscapes. *Environmental and Ecological Statistics*, **8**, 5-20.

[8] Johnson, G.D., Myers, W.L., Patil, G.P., and Walrath, D. (1998). Multiscale analysis of the spatial distribution of breeding bird species richness using the echelon approach. In : *Assessment of Biodiversity for Improved Forest Planning*, (Ed. P. Bachmann, M. Kohl, and R. Paivinen), Kluwer, Boston, 135-150.

[9] Johnson, G.D. and Patil, G.P. (1998). Quantitative multiresolution characterization of landscape patterns for assessing the status of ecosystem health in watershed management areas. *Ecosystem Health*, **4**, 177-187.

[10] Johnson, R.A. and Wichern, D.W. (1982). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

[11] Jones, K.B., Ritters, K.H., Wickham, J.D., Tankersley, R.D. Jr., O'Neill, R.V., Chaloud, D.J., Smith, E.R. and Neale, A.C. (1997). An Ecological Assessment of the United States Mid-Atlantic Region. USEPA, ORD, EPA/600/R-97/130.

[12] Patil, G. P. (2002). Conditional entropy profiles. *Encyclopedia of Environmetrics*, **1**, 413-417.

[13] Patil, G.P. (2003). Overview: Landscape health assessment. In : *Managing for Healthy Ecosystems*, (Eds. D.J. Rapport, W.L. Lasley, D.E. Rolston, N. Ole Nielsen, C.O. Qualset, A.B. Damania), Lewis Publishers, Boca Raton, 559-565.

[14] Patil, G.P., Brooks, R.P., Myers, W.L., and Taillie, C. (2003). Multiscale advanced raster map analysis system for measuring ecosystem health at landscape scale—A novel synergistic consortium initiative. In : *Managing for Healthy Ecosystems*, (Eds. D.J. Rapport, W.L. Lasley, D.E. Rolston, N. Ole Nielsen, C.O. Qualset, A.B. Damania), Lewis Publishers, Boca Raton, 567-576.

[15] Patil, G.P., Johnson, G.D., Myers, W.L., and Taillie, C. (2000). Multiscale statistical approach to critical-area analysis and modeling of watersheds and landscapes. In : *Statistics for the 21st Century: Methodologies for Applications of the Future*, (Eds. C.R. Rao and G. J. Szekely), Marcel Dekker, New York, 293-310.

[16] Patil, G.P., and Taillie, C. (1999). A Markov model for hierarchically scaled landscape patterns. In : *Bulletin of the International Statistical Institute*, Voorburg, 89-92.

[17] Patil, G.P., and Taillie, C. (2000). A multiscale hierarchical Markov transition matrix model for generating and analyzing thematic raster maps. *Environmental and Ecological Statistics*, **8**, 71-84.