A NOTE ON THE INFLUENCE OF GEOGRAPHICAL VARIABLES ON BIBI

Steve Obrebski, September 14, 2009

While helping Mike McClean study the influence of physical variables on variation in BIBI estimates I had the occasion to look at the influence of geographical variables on this index. The results of the analysis that follows suggest that although the correlations between some geographical variables and BIBI may be much higher than for the physical variables the data that was available for the analysis may not represent a trend in BIBI that is influenced by a homogeneous set of factors throughout the range of sampling sites represented in the data. The implications of this result are briefly discussed. To reduce the size of this commentary some of the results of the analysis are not included in the text.

In following the text the reader might want to separate the four figures on pages 4 to 7 at the end of this document so they can be examined while reading the material on pages 1 to 3.

Examination of the correlations between BIBI and the geographic variables showed that many of them were higher than the highest correlations between BIBI and the physical variables. All the correlations are not shown here. The three geographic variables that were most highly correlated with BIBI were % Area Forest, % Area Urban and %HiPermGeoUr. The correlations are summarized in the Table 1 below.

Table 1. Correlations: %AreaForest, %AreaUrban, %HiPermGeoUrb BIBI

	%AreaForest	%AreaUrban	%HiPermGeoUr
%AreaUrban	-0.775 = r 0.000 = p < 0.0001		
%HiPermGeoUr	-0.861 0.000	0.954 0.000	
BIBI	0.821 0.000	-0.835 0.000	-0.818 0.000

Note that the three variables are also highly correlated with each other. The correlation of BIBI with the three highly correlated variables was examined in combination using principal component analysis. Principal component analysis is a procedure for reducing the dimensionality of a set of highly correlated independent variables that may influence a dependent variable. This is achieved by a geometric rotation of the data in a fashion as to generate a set of new uncorrelated variables (Principal Components) which are composites of the original correlated variables. The original variables can be examined jointly in a three dimensional scatter plot. Likewise the principal component variables ensuing from the analysis could be examined in a three dimensional scatter plot except that now the correlation between the new rotated variables would be zero. The amount of variation in the system is proportional to the size of the axes (eigenvalues) of the new variable system and the importance of the original variables in defining the new principal components is proportional to the values of the so-called eigenvectors associated with each eigenvalue. The results of the principal component analysis are summarized in Table 2 on the next page.

Table 2. Principal Component Analysis: %AreaForest, %AreaUrban, %HiPermGeoUrb

Eigenvalues

Eigenvalue	2.7281	0.2389	0.0330
Proportion	0.909	0.080	0.011
Cumulative	0.909	0.989	1.000

Eigenvectors

Variable	PC1	PC2	PC3
%AreaForest	0.556	0.802	-0.216
%AreaUrban	-0.579	0.561	0.592
%HiPermGeoUrb	-0.596	0.204	-0.776

The results of principal component analysis (Table 2) of the variables %AreaForest,, %AreaUrban, and %HiPermGeoUrb show that the first principal component (PC1), having an eigenvalue 2.7281, represents 90.9 % of the variance of this new 3 variable system, (Proportion = 0.909 – top of Table 2.). The eigenvector for this first principal component (PC1) is roughly equally dependent on all three variables (0.556,0.579, and 0.596 eigenvector column values for PC1 in Table 2). These eigenvector values are related to the angles of rotation required to produce the uncorrelated principal components. The interpretation of the first principal component is explained in Fig. 1 (page 4) where it is shown that PC1 is highly correlated with the original variables.

As shown in Fig 2 (page 5), the first principal component is associated with an eigenvalue explaining 90.9 % of the variance. The correlation between BIBI and the first eigenvalue is 0.865 ($p < 0.0001$) within an adjusted R-Square of 74.1% . The results suggest that the three variables mentioned, indicators of levels of forestation and the impact of urbanization, may have a major influence on BIBI values. It is possible that adding other geographical variables might improve our understanding of their influence on BIBI. However, a closer look at the distribution of the data may lead to some other implications, as discussed below.

Regressions of scatter plots of BIBI on Area Forest, % Area Urban and %HiPermGeoUrb are shown in Fig. 3 (page 6). Examination of the scatter of points for % Area Urban and %HiPermGeoUrb shows that there is a tight aggregate of points in the upper right corner of the plots of BIBI points that vary between values slightly over 30 to values close to 50. In the plot for % Forest Area there is a somewhat less clumping ensemble of points, but those that are greater then BIBI values of 30 and % Forest Area values above 70 show no particular trend. The five points on the graphs having BIBI values below 30 labeled Ca and Bl were identified in the regression analysis to be so-called high leverage points having a disproportionate effect on the structure of the regression (analysis not shown here). Moreover, if they are not included in the regression analysis, the linear trend observed disappears or is much less significant (analysis not shown here).

The foregoing observation suggests that we may be dealing with two rather different "populations" of BIBI values. Those values higher than a BIBI value of 30 belong to a generally un-impacted set of sampling sites in which the BIBI values vary about a mean of about 40. The

others belong to an impacted population, having BIBI scores below 30, and tending to occur at sites that are influenced by urban development.

The possibility that the values of BIBI above 30 belong to a distinct set of sites varying around a mean BIBI value of about 40 is examined in the analysis that follows as shown in Fig. 4 (page 7). The data were subdivided into two "populations", and their fit to a normal distribution examined. All the 43 BIBI values were plotted in a histogram and tested for normality. The distribution of BIBI values differs significantly from a normal distribution having the same mean and variance (Fig. 4A and 4B. In Fig. 4C and 4D the data with BIBI values greater than 30 is found not to be significantly different from a normal distribution having a mean of 39.95 and standard deviation of 4.558. In Fig. 4E and 4F an example of a simulated random normal distribution having a mean of 39.95 and standard deviation of 4.558 is shown.

The implications of the results summarized in Fig. 4 follow. Perhaps the data for the sites having BIBI values above 30 belong to a relatively un-impacted set of sites having a normal distribution with a mean of about 40 and standard deviation of about 4.6. That is, in areas not seriously affected by human impacts, the average BIBI value is somewhat below the maximum of 50 (in this case ~ 40) and normal ecological or environmental variation about this mean occurs. Within this forest biome, sites differ in various factors affecting the BIBI index, with minimum values slightly above 30 and maximum values close to 50. In contrast, sites with BIBI values below 30, belong to a relatively urbanized area in which variation among sites is influenced by different variables than in the forested area. If such is the case, attempts that use the entire data set to relate BIBI to particular physical or geographical variables, especially for the purpose of discovering remedial measures for human impacts, may confound somewhat different ecological processes accounting for the variation in BIBI values in different ecological situations. This could affect our conception of what constitutes unacceptable BIBI values that indicate degradation in a particular site and may require the application of different "rules of engagement" in planning ecological enhancement programs for improving and preserving salmon habitat.

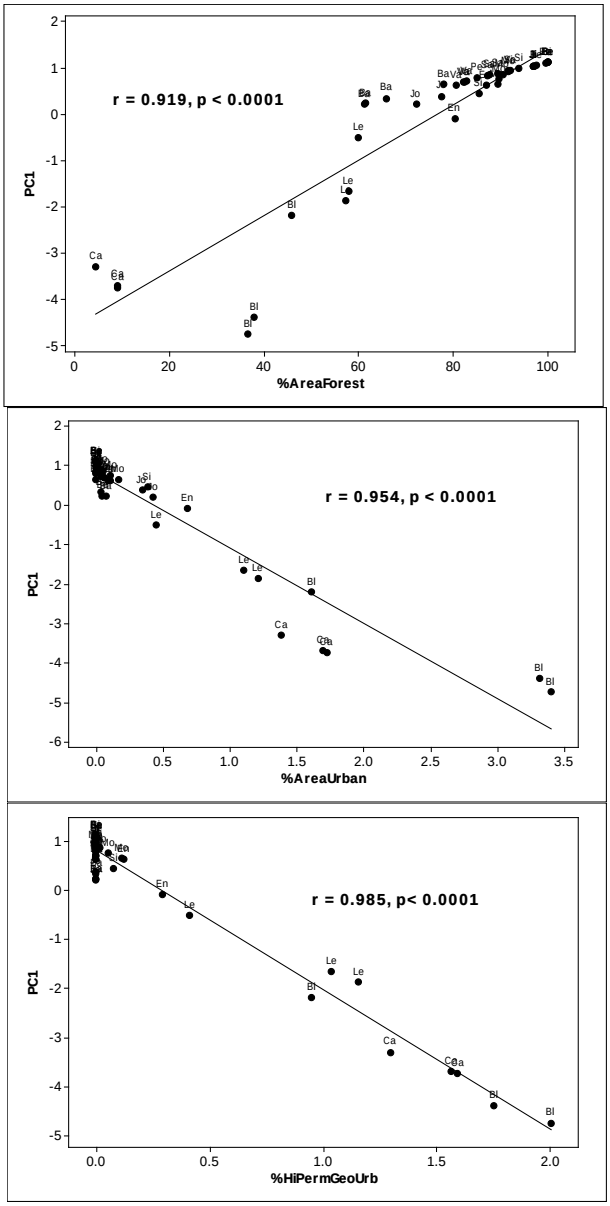

Fig. 1. Plots and correlations between the first principal component (PC1) and the original variables % Area Forest, % Area Urban and %HiPermGeoUr. Correlations (r) and their probabilities are shown in each graph. Linear regression lines fitted to the data are shown. Note that PC1 is a single compound variable that increases with % Area Forest, and decreases with % Area Urban and %HiPermGeoUr. Symbols next to the data points represent different streams. Ba = Bagely, Be = Bear, Bl = Bell, Ca = Cassalery, En = Ennis, Ji = Jimmycomelately, Jo = Johnson, La = lakeTrib, Le = Lees, Mo = Morse, Pe = Peabody, Sa = Salt, Si = Siebret, Va = Valley.

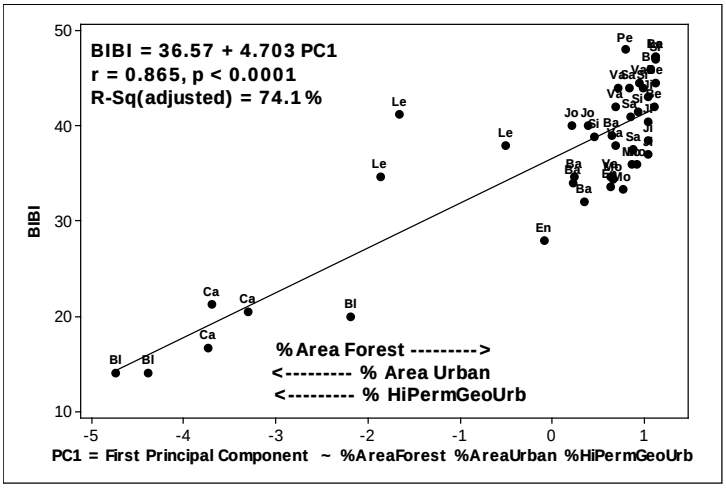

Fig. 2. Linear regression of BIBI on PC1, a principal component positively correlated with %AreaForest, and negatively correlated with %AreaUrban and %HiPermGeoUrb and explaining 90.9% of the variance in the system comprising these three variables. As shown by the arrows in the figure, %AreaForst increases to the right and the other variables increase to the left. The correlation of BIBI with PC1 is quite high and significant ($r = 0.65$, $p < 0.0001$) with 74.1% of the variance explained by the regression. Symbols next to the data points represent different streams. Ba = Bagely, Be = Bear, Bl = Bell, Ca = Cassalery, En = Ennis, Ji = Jimmycomelately, Jo = Johnson, La = lakeTrib, Le = Lees, Mo = Morse, Pe = Peabody, Sa = Salt, Si = Siebret, Va = Valley.

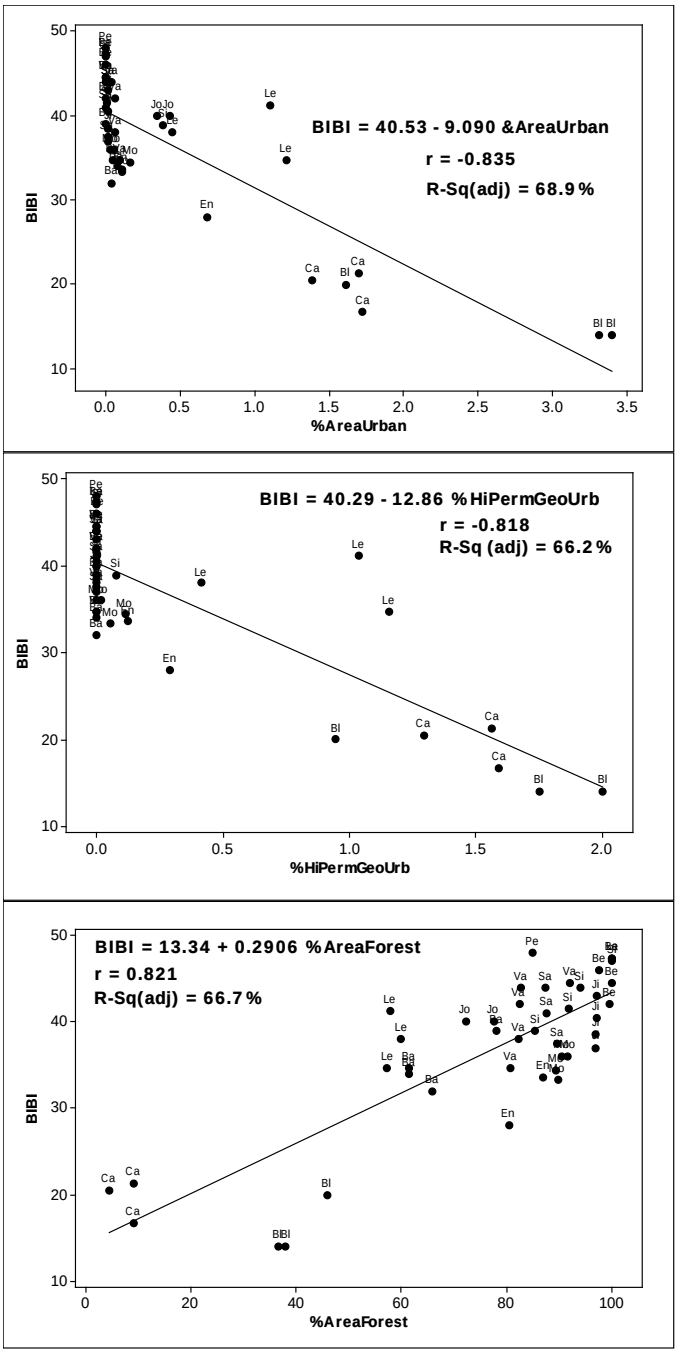

Fig. 3. Regressions of BIBI on Area Forest, % Area Urban and %HiPermGeoUrb. The 5 points labeled Ca and Bl below a BIBI value of 30 are observatiosn whose values give them a large influence on the regression.

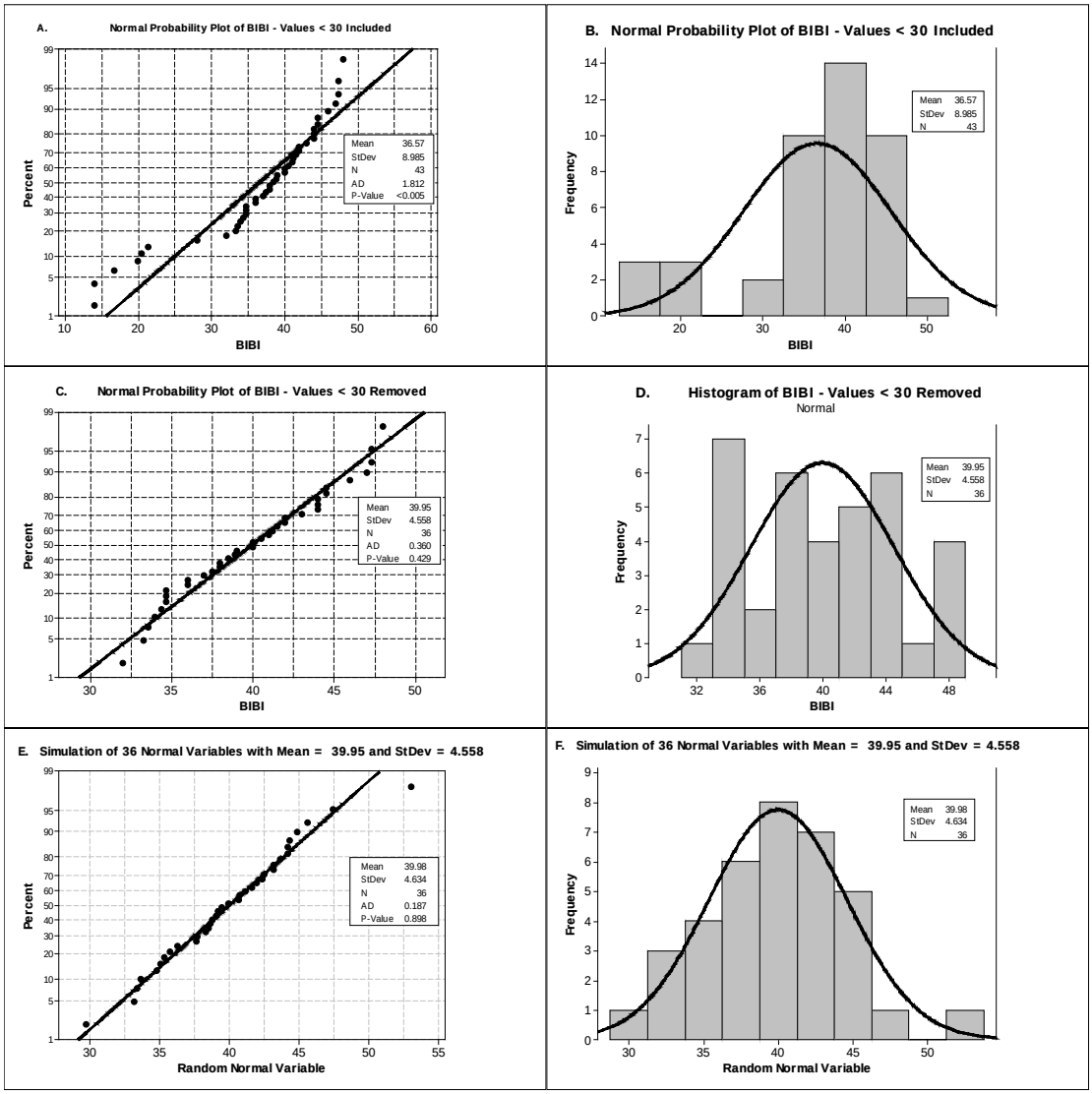

Fig. 4. A, (Top row, left). Probability plot of BIBI values in the physical variables data set, The distribution is significantly different from normal with $p < 0.005$. B,(Top row, right). Histogram of BIBI values in A. with normal distribution having the same mean and variance superimposed. C. (Middle row, left). Probability plot of 36 BIBI values with BIBI scores less than 30 removed. The data is not significantly different from a normal distribution with a mean of 39.95 and standard deviation of 4.558. D. (Middle row, right). Histogram of BIBI values in C with normal distribution having the same mean and variance superimposed. E. (Lower row, left). Probability plot of an example of a simulation of 36 random normal variables having the same mean and variance as the 36 BIBI values in plots C and D. F. (Lower row, right). Histogram of BIBI values in E with normal distribution having the same mean and variance superimposed.