Using SAS for multivariate analysis of water quality parameters to determine freshwater inflow needs of estuarine organisms in Texas bays

Paul A. Montagna, Harte Research Institute, TAMU-CC, Corpus Christi, TX

ABSTRACT

Freshwater inflow is important to maintain estuary health. Effects of water flow are dynamic, and it is impossible to sample all conditions as they vary over space and time. Benthos, however, are fixed in place, continuously sample the overlying water conditions, and benthic indices of biotic integrity (BIBIs) are useful for assessing aquatic health. Many biotic characteristics can be measured. Likewise, the water column condition can be described using multiple water quality parameters. So, how do we relate multivariate water column condition with multivariate benthos response? The metrics were reduced to two variables using principal component analysis (PCA). Then the BIBI metrics and water column PC variables were significantly correlated, indicating that benthic communities respond to changes in salinity and do so in a relatively predictable manner. If inflow is reduced (i.e., salinity increased), it will cause upstream communities to take on characteristics of downstream communities. This presentation will demonstrate how SAS was used to solve three problems related to data management, reducing dependent and independent data sets, and testing the relationship between water column condition and biotic response.

INTRODUCTION

Environmental flows (i.e., flow from rivers to estuaries) are vulnerable to water resource development. Freshwater inflows serve a variety of important functions in estuaries including creation and preservation of low-salinity nurseries, driving movement and reproductive timing of estuarine species, and transport of sediments, nutrients, and organic matter. Dilution of sea water and subsequent changes in salinity are is a primary factor controlling estuary condition. The estuary condition drives the biological response and distribution within an estuary.

Bottom-dwelling invertebrates (called benthos or macrofauna) are especially sensitive to changes in salinity over time. The benthos live long, are sessile, continuously sample the overlying water conditions, integrate ephemeral hydrologic conditions over time, and provide a long-term record of short-term changes. Thus, benthic monitoring is an important tool to assess the ecological health and integrity of aquatic ecosystems.

The present study examined water column and benthic data collected from the Lavaca-Colorado estuary on the central Texas coast in the western Gulf of Mexico. The goal was to determine identify salinity zones and how changes in salinity regimes (as a proxy for freshwater inflow) affect benthic populations (as a proxy for ecosystem health).

STATISTICAL METHODS

The statistical and data base problem is to link environmental variables (i.e,, measures of water column conditions) with biological variables (i.e., measures that indicate ecosystem health). Both data sets are multivariate.

A benthos sample contains different species of different abundances (or number values), so the design matrix should contain samples as rows and species as columns (Table 1). However, we never know which species will be found, and new species are found all the time. So, most biological data is stored as a vector, not a matrix, as shown in the example table to the right. Note that species 1 occurs in both samples, but in sample 2, we found two new species (3 and 4). This storage approach creates a problem because simply computing sample means with PROC MEANS would return the wrong values. For example the average for Species 2 would be 5, because the zero

Table 1. Biological data.

Sample	Species	Value
1	1	6
1	2	5
2	1	9
2	3	1
2	4	2

value for Species 2 was not found in the second sample. So, the first task is to put the zeros in the data set. This is easily accomplished by transposing the dataset twice. The first transpose will create the missing cells by using the PREFIX option and the ID statement so that each species is placed in its own column. The second transpose returns the data set to the original format, and then it is easy to replace the missing values with zeros as in the example below.

```
PROC SORT DATA=sp;
    BY sample species;
    RUN;
PROC TRANSPOSE DATA=sp OUT=tsp PREFIX=s;
    BY sample;
    ID species;
    VAR value;
    RUN;
PROC TRANSPOSE DATA=tsp OUT=sp1;
    BY sample;
    RUN;
DATA SP2;
    SET sp1;
    IF value=. THEN value=0;
    RUN;
```

The next task is to calculate various measures of biodiversity, such as the Shannon diversity index (H´) or the number of species (richness). This is best accomplished by transposing the data and treating each sample as an array. We can also calculate total abundance of organisms (i.e., to sum values). In this case, there is no need for the missing species cells, so the array for each sample contains all the information necessary for the calculations. So, we do not use the ID statement as in the following statements.

```
PROC TRANSPOSE DATA=sp OUT=tsp2 PREFIX=s;
    BY sample;
    VAR value;
    RUN;
DATA diversity;
    SET tsp2;
    ARRAY s s1-s100;                    /* Expect less than 100 species, =MAX(ts) */
    DROP _NAME_ s1-s100;
    Ts=N(OF s1-s100);                   /* Ts = total number of species in a sample */
    Tn=SUM(OF s1-s100);                 /* Tn = total number of individuals in a sample */
    hprime=0;
    jprime=0;
    IF tn=0 THEN GOTO nosp;             /* Skip samples without species */
    DO OVER s;                          /* Loop to calculate indices */
        IF s=. THEN GOTO msp;           /* Skip species with missing values */
        hprime = hprime+(-(s/Tn)*LOG(s/Tn));    /* Shannon's diversity index */
        msp:
        END;
    n1=EXP(hprime);                     /* Hill's Number of dominant species index */
    jprime=LOG(n1)/LOG(Ts);             /* Pielou's evenness index */
    nosp: ;
    OUTPUT;                             /* Add new calculated values to dataset */
    RETURN;                             /* Start the next sample */
    RUN;
```

The environmental data is simpler to analyze because it is a matrix where each sample is a row and all the measured responses are columns. A multivariate analysis is performed with a goal to reduce the number of variables and extract the largest contribution of variance in sequential order. While there are

many approaches to perform this analysis, principal components analysis is the best approach. The goal is to compute scores for each sample that will be correlated to the biological responses. Either PROC FACTOR or PRINCOMP can be used, but FACTOR has some options that will be useful so it is used here (Long et al. 2003). The example below uses an environmental data set (env), which is first log transformed, and outputs the statistics to create unique data sets of variable loads and sample scores. Then files are created to make plotting the data convenient.

```
TITLE 'Principal Components Factor Analysis on water column parameters [(Log x)+1]';
PROC FACTOR DATA=env METHOD=PRINCIPAL ROTATE=VARIMAX NFACTORS=3 COV
        PREPLOT NPLOT=5 PLOT OUT=scores OUTSTAT=pcsstat;
    VAR do sal temp din po4 sio4 chl pH;
     RUN;
TITLE2 'Scores data: factor scores for samples';
PROC SORT DATA=scores;
    BY factor1 factor2 factor3;
    RUN;
DATA _NULL_;                        /* This step creates a text file for plotting scores (Figure 2) */
    SET scores;
    FILE 'c:\temp.txt';
    WHERE factor1 NE .;
    IF _N_=1 THEN PUT  @1 'Date' @11'Sta' @22 'Score1' @32 'Score2' @42 'Score3';*Header;
    PUT @1 sdate @10 sta @21 factor1 8.4 @31 factor2 8.4 @41 factor3 8.4 ;
    RUN;
TITLE2 'Loads data: loading scores for water column variables';
DATA ostat;                                                 /* This step selects the loads */
    SET pcsstat;
    WHERE _type_='PATTERN';
    RUN;
PROC TRANSPOSE DATA=ostat OUT=loads;
    ID _name_;
    RUN;
DATA _NULL_;                    /* This step creates a text file for plotting vector labels (Figure 1) */
    SET loads;
    FILE 'c:\temp1.txt';
    IF _N_=1 THEN PUT  @1 'Loads' @16 'Factor1' @26 'Factor2' @36 'Factor3' ;        *Header;
    PUT @1 _NAME_ @15 factor1 8.4 @25 factor2 8.4 @35 factor3 8.4;
    RUN;
DATA _NULL_;                        /* This step creates a text file for plotting vectors (Figure 1) */
    SET loads;
    FILE 'c:\temp2.txt';
    IF _N_=1 THEN PUT  @2 'Vector1' @12 'Vector2' @22 'Vector3' ;
    PUT @3 '0.0'    @13 '0.0' @23 '0.0' /
        @1 factor1 8.4 @11 factor2 8.4 @21 factor3 8.4/;
RUN;
```

Principal Component Analysis (PCA) of water-column environmental variables revealed that salinity was the primary factor driving the system and it is inversely related to nutrient concentrations (Fig. 1). Thus, PC 1 represents a linear scale of freshwater inflow effects on hydrology. Dissolved oxygen and temperature also exhibited secondary roles, thus PC 2 represents seasonal effects because temperature increases in summer and dissolved oxygen decreases in summer. PC 1 & 2 vectors explained 82% of the variation.

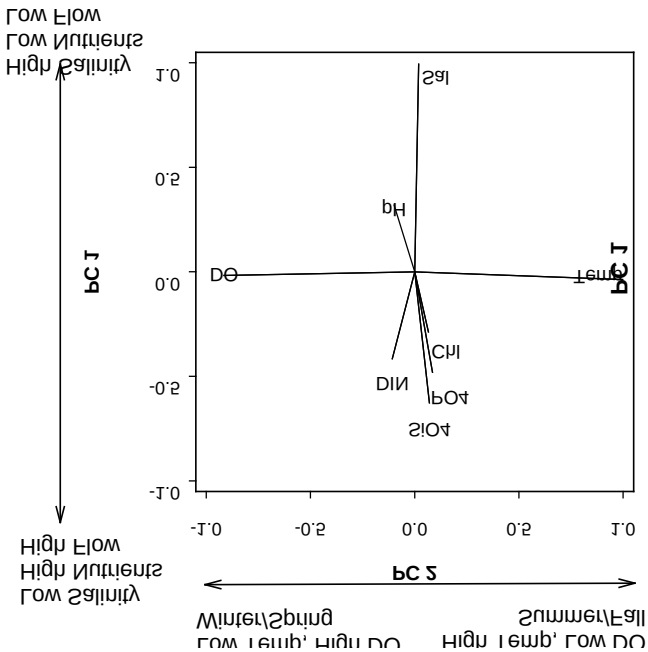

Figure 1. Vector loads computed from the principal components analysis of environmental variables.

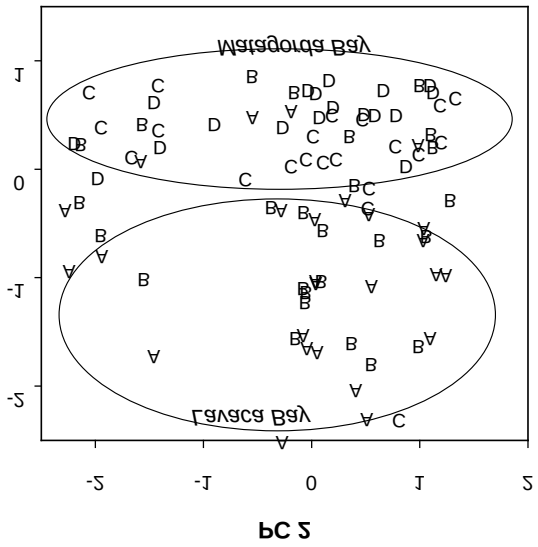

Figure 2. Principal component scores for samples from two bays.

Stations A, B, C, and D were aligned from the Lavaca River to the Gulf of Mexico respectively. Station loading scores exhibited a fairly distinct spatial pattern along the salinity gradient, that is PC 1, with station A generally exhibiting the most negative relationship with salinity and station D the most positive (Figure 2).

The next step is to merge the biological and environmental data by sample. After the data sets are merged, then the biological response can be correlated with the environmental drivers to link environmental condition to ecological response (Green and Montagna 1996).

```
PROC CORR DATA=pc;
    VAR factor1 factor2;
    WITH gm2 nm2 Hprime;
    RUN;
```

PC 1 scores for the environmental variables were significantly correlated with all biological variables, but were not correlated to PC 2 scores (Table 2). Although seasonal differences exist, the biological community is mainly affected by freshwater inflow. No significant correlations were found between the water-column and sediment PCAs, indicating that freshwater inflow and salinity is more important at controlling the benthic community than the sediment structure.

Table 2. Correlation between environmental condition and biological response.

Biological Variables	Hydrographic PC Factor 1	Hydrographic PC Factor 2
Biomass (g m^{-2})	r = 0.34754 p = 0.0002	r = -0.09274 p = 0.3397
Abundance (n m^{-2})	r = 0.29419 p = 0.0020	r = -0.08058 p = 0.4071
Diversity (H′)	r = 0.59740 p <.0001	r = -0.03215 p = 0.7411

CONCLUSION

There is a direct relationship between freshwater inflow and changes in the water column characteristics of estuaries. High inflow is directly related to low salinity and high nutrient levels in the system as well. Significant relationships were found between multivariate measures of the biological community and between salinity zones allowing communities to be defined by two salinity zones in two different bay systems. These salinity regimes act as a proxy for measuring the effects of freshwater inflow, and the results indicate that benthic communities respond to changes in inflow and do so in a relatively predictable manner. If inflow is reduced, then it will cause the upstream communities to take on downstream characteristics (Kim and Montagna 2009).

The SAS system provides powerful tools for data management, statistical analysis, and data visualization. If fact, it would very difficult and time consuming to perform the analyses presented here without the ability to pass derivative data sets from one procedure to another. While many software packages contain multivariate statistical analyses, few, if any, have the flexibility and options that allow for the specific analyses described above.

REFERENCES

Green, R.H. and P. Montagna. 1996. Implications for monitoring: Study designs and interpretation of results. *Canadian Journal of Fisheries and Aquatic Sciences* 53:2629-2636.

Kim, H.-C. and P.A. Montagna. 2009. Implications of Colorado River freshwater inflow to benthic ecosystem dynamics: a modeling study. *Estuarine, Coastal and Shelf Science* 83:491-504.

Long, E.R., R.S. Carr, and P.A. Montagna. 2003. Porewater toxicity tests: value as a component of sediment quality triad assessments. In: R.S. Carr and M. Nipper (eds.) *Porewater Toxicity Testing: Biological, Chemical, and Ecological Considerations.* Society of Environmental Toxicology and Chemistry (SETAC) Press, Pensacola, FL. Chapter 8, pp. 163-200.

ACKNOWLEDGMENTS

The field work for this study was supported by several grants from the Texas Water Development Board, 523 Research and Planning Fund, Research Grants, authorized under the Texas Water Code, Chapter 524 15, and as provided in §16.058 and §11.1491; and a contract from the Lower Colorado River Authority PO No. 49032. Much of the analytical work was supported by grant number 527 NA09NMF4720179 from the National Oceanic and Atmospheric Administration under the 528 Comparative Assessment of Marine Ecosystem (CAMEO) program.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul A. Montagna
Harte Research Institute for Gulf of Mexico Studies
Texas A&M University-Corpus Christi
6300 Ocean Drive, Unit 5869
Corpus Christi, Texas 78712
paul.montagna@tamucc.edu
Office (361) 825-2040
http://harteresearchinstitute.org/

TRADEMARK INFORMATION