Using PROC STANDARD and PROC SCORE to impute missing multivariate values

Paul A. Montagna, Harte Research Institute, Texas A&M University-CC, Corpus Christi, TX

ABSTRACT

One of the most frustrating aspects of multivariate analysis is that one missing value in one column will cause the entire sample row to be deleted. This means that all the other values of the sample are lost, and one is not able to classify that sample. There are at least two ways to deal with this problem, what I call the "pre-analysis method" and the "post analysis method." There are many approaches to impute a missing value prior to analyses so the samples are retained. One simple method is to impute the missing value "before" the multivariate analysis using PROC STANDARD to replace missing values based on a statistical characteristics or PROC REG to use a predictive approach. Alternatively, one could impute the missing value "after" the multivariate analysis using the multivariate coefficients as input to PROC SCORE, thus being able to classify the sample. Either approach has its strengths and weaknesses.

INTRODUCTION

Multiple variables are a common property in many data sets in all fields of study. Multivariate data often takes the form of a matrix where samples are rows and the response variables are in columns. Multivariate analysis is the tool of choice to explain correlations or covariances among multiple variables, discover underlying structure in datasets, or summarize relationships among variables. However, one frustrating problem is that just one missing value among the response variables causes all the values of that sample to be deleted and lost to the analysis. This is especially frustrating when the purpose is to classify samples because some samples won't be classified.

There are at least two approaches to deal with the problem of missing values, what I call the "pre-analysis method" and the "post analysis method." Imputing missing values "before" multivariate analyses ensures the samples are retained in the multivariate analysis. One simple method is to replace missing values based on descriptive statistical characteristics, or by using predictive approaches such as linear or non-linear dependencies. Alternatively, one could impute the missing value "after" the multivariate analysis using the multivariate coefficients as predictor variables to classify the sample, but this still requires that missing values be imputed in the original data set.

The example presented here shows how to use both approaches to impute missing values in water quality data sets. Environmental flows (i.e., flow from rivers to estuaries) are vulnerable to water resource development (Montagna et al. 2013). 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 the primary factor controlling estuary condition. The estuary condition drives the biological response and distribution within an estuary. Defining changes in water quality is a multivariate problem.

STATISTICAL METHODS

Study Area

The example data was collected from the Lavaca-Colorado Estuary, also known as the Matagorda Bay System. Freshwater inflow studies have been performed in the ecosystem since 1988, and ecological models have shown that inflow drives estuary condition, which in turn drives productivity (Kim and Montagna 2009). Water quality samples have been taken from six stations (A – F) along two inflow gradients (Fig. 1). Stations A and B are located in freshwater-dominated Lavaca Bay, and represent effects of the Lavaca River. Stations C and D are located in marine-dominated Matagorda Bay and represent the lack of inflow effects. Stations E and F are located in East Matagorda Bay, and are influenced by river inflow from the Colorado River. Stations were sampled quarterly (January, April, July, and October) from April 2004 to July 2010. The average monthly salinity in the ecosystem from 1976 – 2007 was 19.98 practical salinity units (psu), the 25% was 15.74 psu, and the 75% was 25.01 psu (Montagna et al. 2011). The climatic conditions for each sample date was classified based on these

percentiles where wet periods were when the salinity was 25% or less, dry periods were when the salinity was 75% or more, and the climatic period was classified as average when the salinities were between the 25th and 75th percentile.

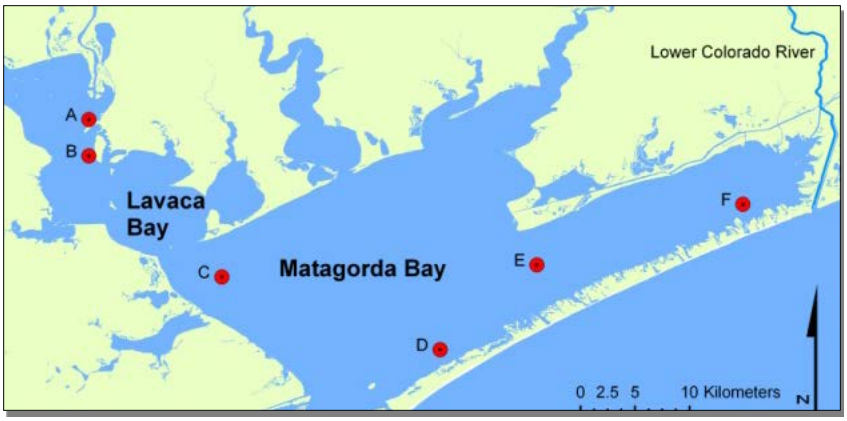


Figure 1. Station locations in the Lavaca-Colorado Estuary (from Pollack et al. 2011).

Water quality parameters measured include physical characteristics: water temperature, pH, dissolved oxygen (DO), salinity (sal), specific conductivity, turbidity(cond), and secchi disk depth (depth at which the disk is visible); and chemical characteristics: chlorohyll a (Chl), phosphate (PO4), silicate (SiO4), ammonium (NH4), and total nitrite plus nitrate (NOx). The data set name is LC, which is an abbreviation for Lavaca-Colorado estuary.

Exploratory Analysis

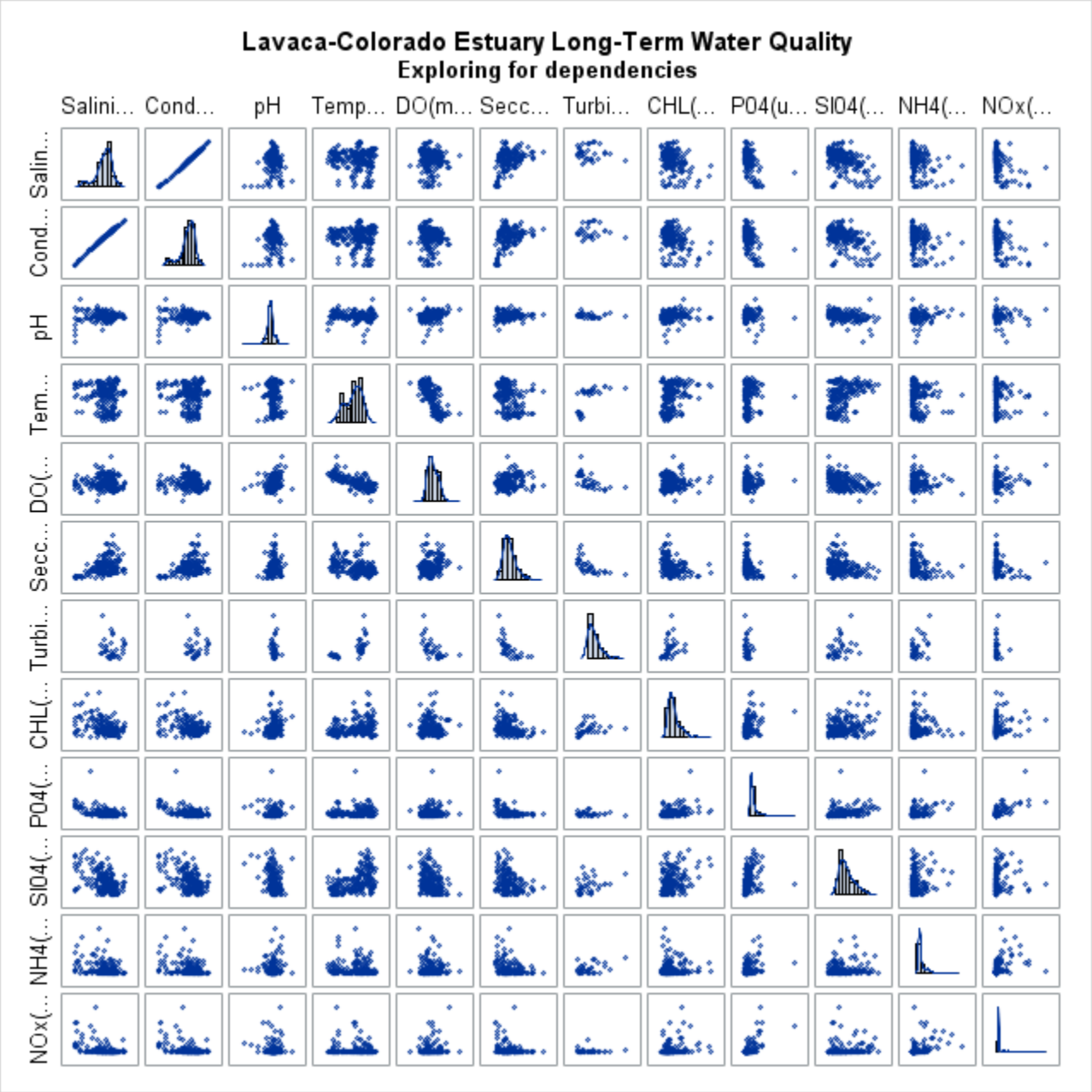
The first step is to perform exploratory analysis to identify the number of missing values. This is easily performed using PROC MEANS.

```
PROC MEANS data=LC;
 var sal cond pH temp DO secchi turbidity Chl PO4 SiO4 NH4 NOx;
 run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
sal	Salinity(psu)	218	24.2743568	8.5425674	0.1300000	40.3150000
cond	Conductivity(uS/cm)	212	37.5390053	12.6661489	0.2800000	60.1400000
ph	pH	218	8.1157217	0.2410604	6.5900000	9.1150000
temp	Temperature(C)	218	22.9688146	5.8209257	10.6600000	32.1249999
do	DO(mg/l)	218	7.5283104	1.2838575	3.6300000	12.9900000
secchi	Secchi(m)	196	0.7478741	0.4111962	0.1000000	2.5000000
turbidity	Turbidity (NTU)	36	13.6513889	12.4736836	1.0000000	57.8000000
chl	CHL(mg/l)	212	7.5038204	5.6937652	0.2275000	35.4450000
po4	P04(umol/L)	206	0.9630780	1.2324326	0	12.8525000
sio4	SI04(umol/L)	206	64.4782626	50.4865378	0	227.0925000
nh4	NH4(umol/L)	206	1.7181769	2.6252088	0	18.2375000
nox	NOx(umol/L)	206	2.9109110	7.6408242	0	61.6375000

There were a total of 218 rows in the data set, and only four of the variables were complete, i.e., no missing values. So the next step is to examine relationships among the variables for possible dependencies using PROC SGSCATTER.

```
PROC SGSCATTER data=LC;
 matrix sal cond pH temp DO secchi turbidity Chl PO4 SiO4 NH4 NOx
        / diagonal=(histogram kernel);
 run;
```



Two trends are obvious: that salinity and conductivity linearly related, and that turbidity and secchi depth are inversely related. Only 6 of the conductivity values are missing, which is about 3% of the dataset, so

we could estimate these values using a linear regression model as a function of salinity. However, the relationship between salinity and conductivity is so strong (R^2 = 0.997) that the conductivity value is totally redundant, and can be dropped for the purposes of a multivariate analysis. We could use an exponential decay nonlinear regression model to estimate the missing values for turbidity based on the secchi values, but there are only 36 turbidity values meaning 83% of the values are missing. A convention is that we should not estimate more than 5% of missing values, so the turbidity values should be dropped altogether.

Another approach to estimate missing values is to use PROC STDIZE, which calculates various descriptive characteristics. There are two important options for this procedure: METHOD and REPONLY.

```
PROC STDIZE data=LC out=LCm method=mean missing=mean REPONLY ;
 var sal cond pH temp DO secchi turbidity Chl PO4 SiO4 NH4 NOx;
 run;
```

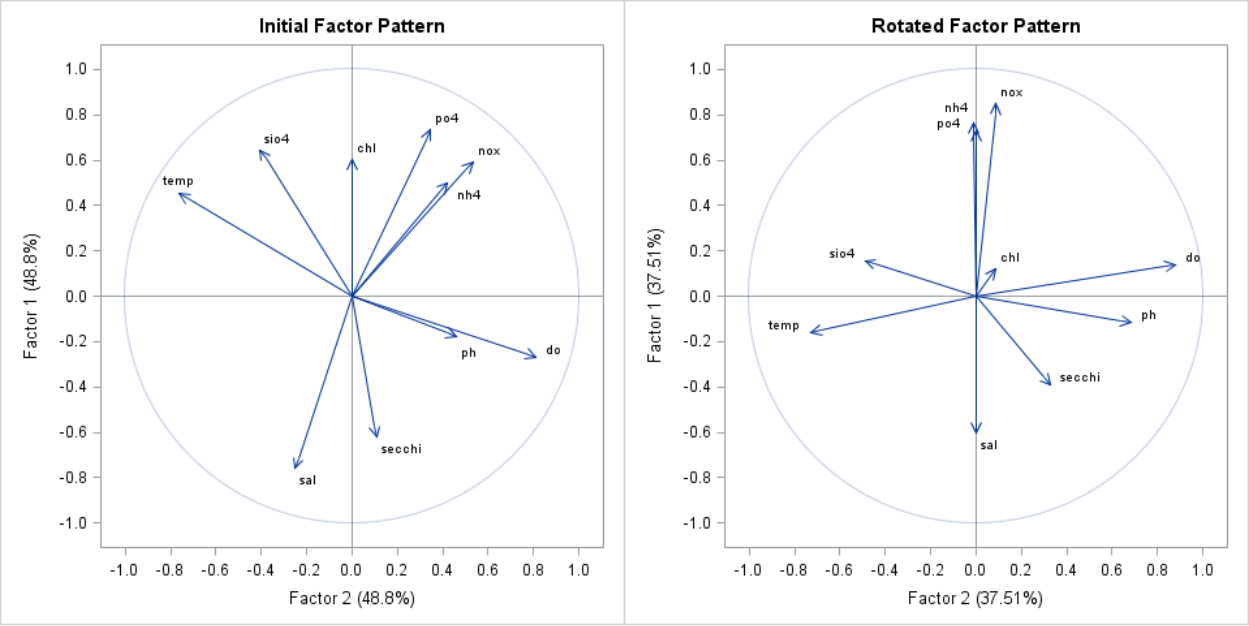
The REPONLY option is used to replace missing values without standardization. There are many method options that can be used to calculate the missing values, in this example the overall sample mean is used.

Multivariate Analysis

There are several multivariate method procedures, but I prefer PROC FACTOR because of the options available. The most important option is to create a dataset with the scoring coefficients.

```
PROC FACTOR data=LC
            method=principal
            rotate=varimax
            nfactors=3
            plots(flip)=(scree initloadings(vector) loadings(vector))
            out=Scores
            outstat=pcastat;
  var sal pH temp DO secchi Chl PO4 SiO4 NH4 NOx;
  run;
```

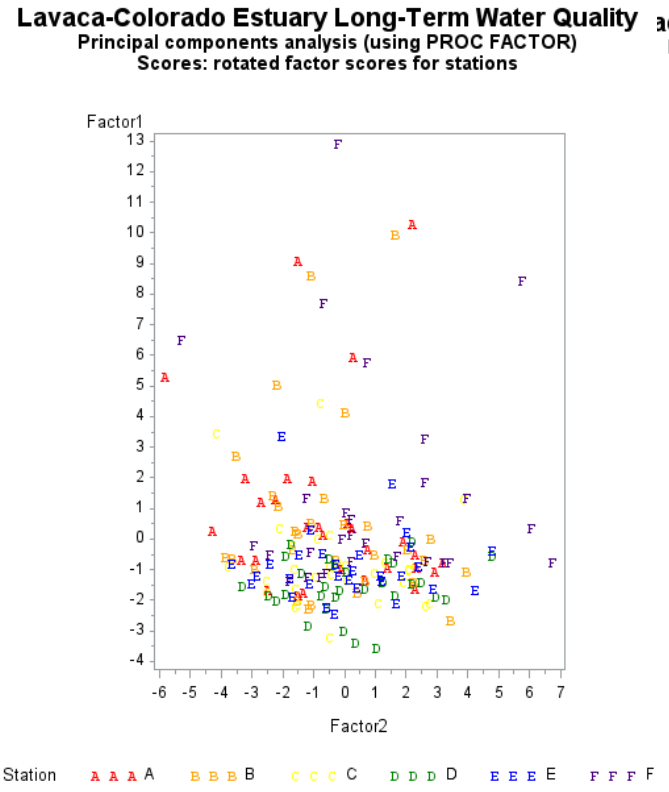
In this case only 180 out of 218 data lines were used in the analysis because of the number of missing values. Thus, 17% of the samples were not used in the analysis, and thus not classified.

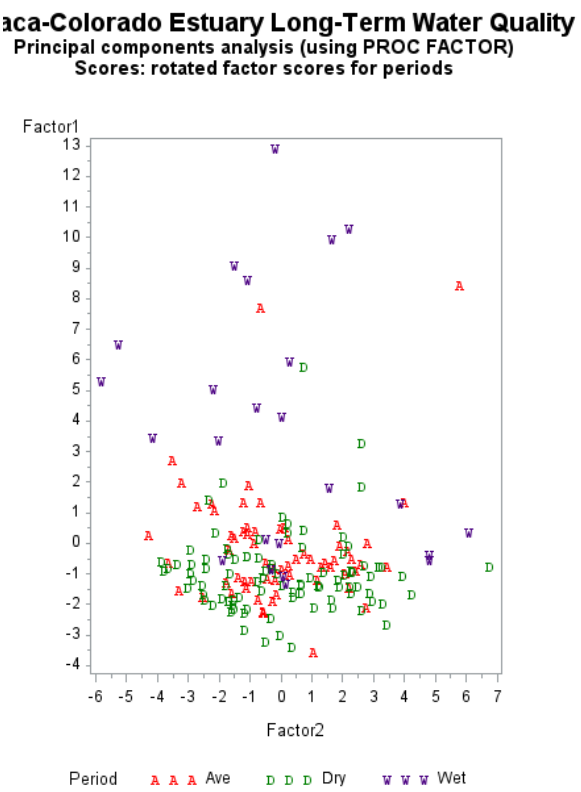


PROC FACTOR in principal components mode can compute variable loads using rotation procedures. Rotating is important because it allows the loads to be aligned along axes in a perpendicular method (above right). The scores for each station are contained in the output data set “Scores”. However, the default output is scores without rotation. To obtain rotated scores, use PROC SCORE with the type='PATTERN' option.

```
title3 'Scores: rotated factor scores for stations';
PROC SCORE data=LC score=pcastat type='PATTERN' out=rotate_score;
  var sal pH temp DO secchi Chl PO4 SiO4 NH4 NOx;
  run;
PROC SORT data=rotate_score; by sta mdate; run;
symbol1 interpol=none value='A' color=red;
symbol2 interpol=none value='B' color=orange;
symbol3 interpol=none value='C' color=yellow;
symbol4 interpol=none value='D' color=green;
symbol5 interpol=none value='E' color=blue;
symbol6 interpol=none value='F' color=indigo;
axis1 length=4.25 in;
axis2 length=3.25 in;
PROC GPLOT data=rotate_score;
 plot factor1*factor2=sta / vaxis=axis1 haxis=axis2;
 run; quit;
```

When rotated sample scores are plotted, only 180 of the 218 samples are plotted (below).



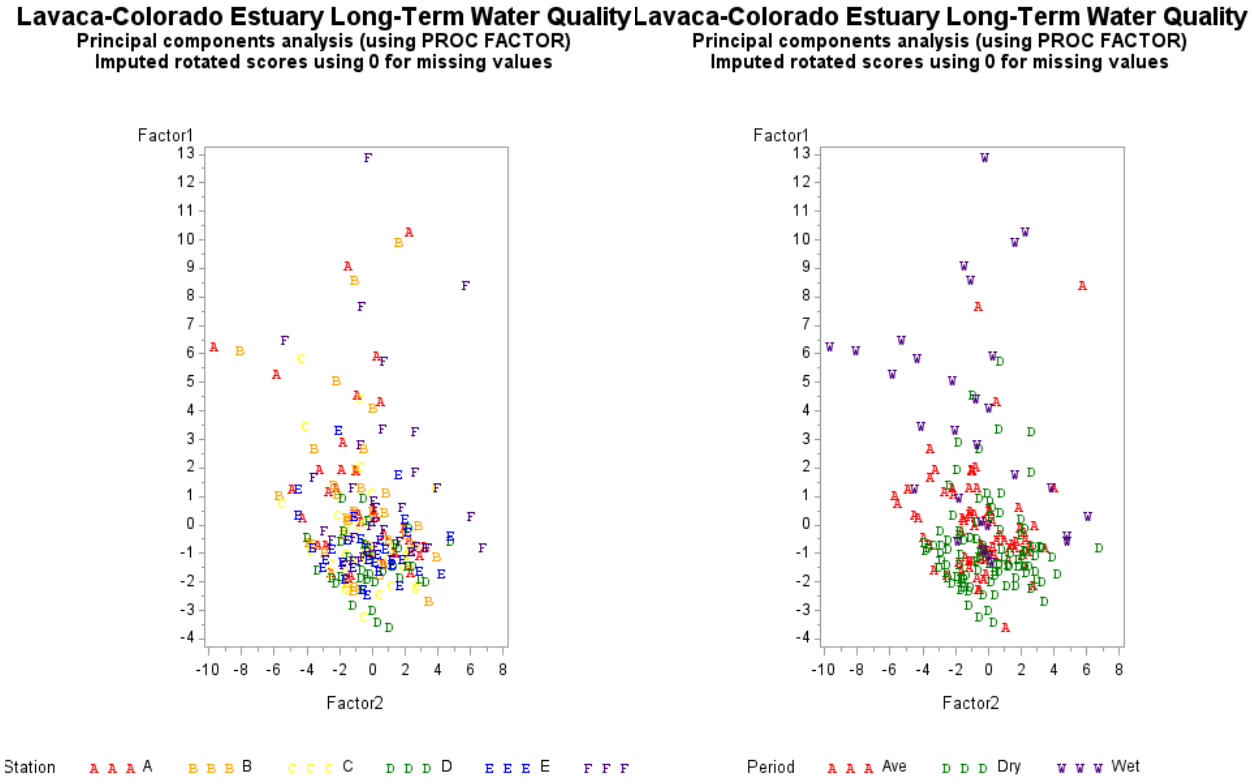


Imputing Multivariate Values

The PROC SCORE procedure multiplies values from two SAS data sets, one containing coefficients (for example, factor-scoring coefficients or regression coefficients), and the other containing raw data to be scored using the coefficients from the first data set. The result is a SAS data set containing linear combinations of the coefficients and the raw data values. Because PROC SCORE will multiply the original data set with the coefficients, the product of a missing value is still a missing value. One approach is to simply remove the missing values by replacing them with a zero, and thus they have no effect on the computation of the sample score. The PREDICT option can be used to accomplish this because a variable with a coefficient of -1 can tolerate a missing value and still produce a prediction score. Also, a variable with a coefficient of 0 can tolerate a missing value.

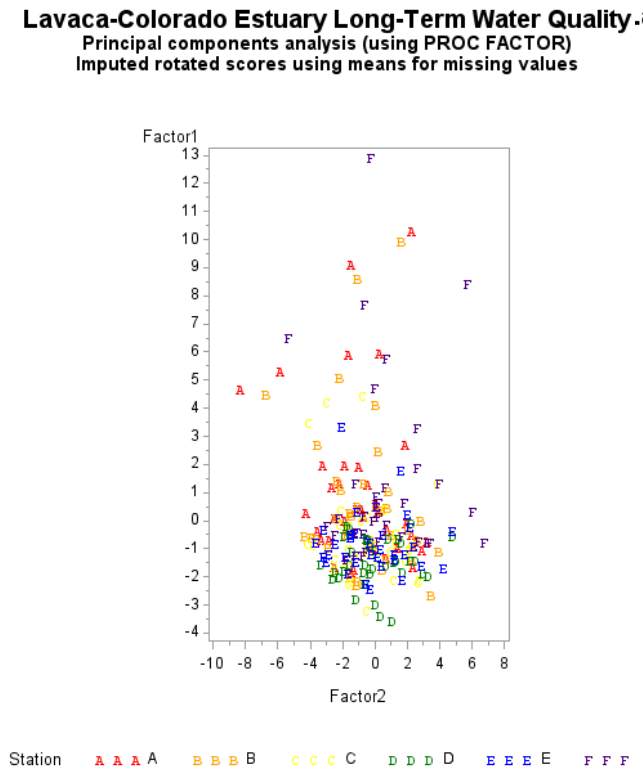
First, a DATA step is used to replace missing values with -1, and then PROC SCORE is used to impute the missing values. In this case the missing values are essentially set to 0, which means the missing values are not contributing to the scores for the samples that had missing values. Below is the SAS code to produce this analysis and the resulting scatter diagram with all 218 samples classified.

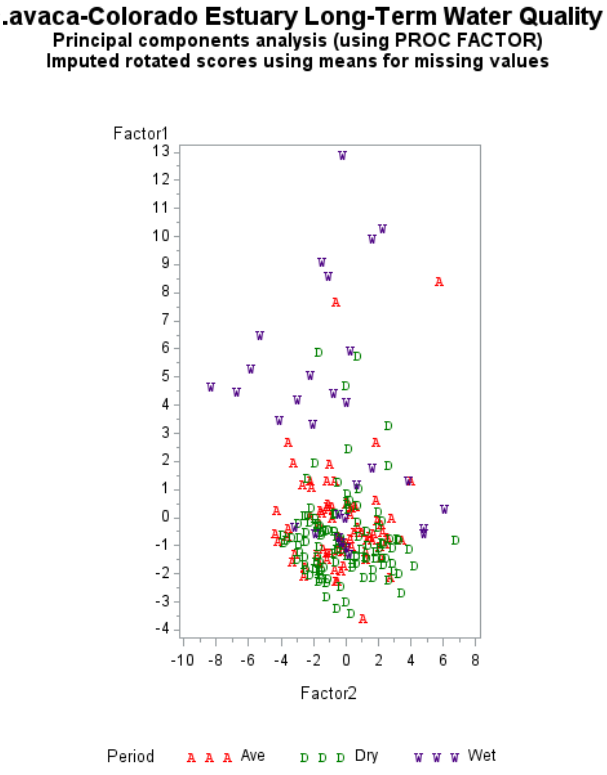
```
title3 'Imputed rotated scores using 0 for missing values';
data iscore;
 set LC;
 if secchi=. then secchi=-1;
 if chl=. then chl=-1;
 if po4=. then po4=-1;
 if sio4=. then sio4=-1;
 if nh4=. then nh4=-1;
 drop turbidity cond;
 run;
PROC SCORE data=iscore score=pcastat type='PATTERN' out=imp_score predict ;
  var sal pH temp DO secchi Chl PO4 SiO4 NH4 NOx;
  run;
```



Another approach is to use PROC STDIZE to replace the missing values with the overall sample mean for each missing value. The net effect of using the mean is that the variable is neutral for scoring purposes when the multivariate score is calculated using PROC SCORE. Below is the SAS code and the resulting scores for this type of analysis.

```
title3 'Imputed rotated scores using means for missing values';
PROC STDIZE data=LC out=iscorem method=mean missing=mean REPONLY;
 var sal pH temp DO secchi Chl PO4 SiO4 NH4 NOx;
 run;
PROC SCORE data=iscorem score=pcastat type='PATTERN' out=imp_scorem predict ;
  var sal pH temp DO secchi Chl PO4 SiO4 NH4 NOx;
  run;
```





CONCLUSION

It is possible to classify all samples in a multivariate analysis, even those with missing values. The approach is to impute the missing values. There are two main approaches to imputation: using univariate – pre-analysis, or using multivariate – post-analysis. Post-Analysis can be done in at least two ways: using zero for missing values so they have no effect in the computation of the sample score, or using the univariate sample mean for the missing value so that the effect on that multivariate score is neutral.

REFERENCES

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. http://dx.doi.org/10.1016/j.ecss.2009.04.033

Montagna, P.A., J. Brenner, J. Gibeaut, and S. Morehead. 2011. Coastal Impacts. In: Schmandt, J., G.R. North, and J. Clarkson (eds.), The Impact of Global Warming on Texas, second edition. University of Texas Press, Austin, Texas, pp. 96-123.

Montagna, P.A., T.A. Palmer, and J. Beseres Pollack. 2013. Hydrological Changes and Estuarine Dynamics. SpringerBriefs in Environmental Sciences, New York, New York. 94 pp. http://dx.doi.org/10.1007/978-1-4614-5833-3

Pollack, J.B., T.A. Palmer, and P.A. Montagna. 2011. Long-term trends in the response of benthic macrofauna to climate variability in the Lavaca-Colorado Estuary, Texas. *Marine Ecology Progress Series* 436: 67–80. http://dx.doi.org/10.3354/meps09267

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 MX954526 from the U.S. Environmental Protection Agency, Gulf of Mexico 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