Paper SD16
Missing Data Values: Analyzing their Effects on Rainfall Forecasts Using PROC EXPAND and the SAS® Time-Series Forecasting System

Richard March, *South Florida Water Management District*

ABSTRACT

Missing values are a common problem faced in the analysis of time-series data. Many SAS® time-series procedures (PROC ARIMA, PROC VARMAX, etc.) are intolerant of missing values, particularly when these missing values are embedded in the time-series, rather than occurring at the beginning or end of the series. PROC EXPAND is designed to convert time series from one sampling interval or frequency to another and to interpolate missing values in the time series Daily rainfall data from the National Climatic Data Center (http://www.ncdc.noaa.gov/oa/ncdc.html) and the DBHYDRO database of the South Florida Water Management District (http://sonar.sfwmd.gov:7777/pls/dbhydro_pro_plsql/show_dbkey_info.main_page) contain rainfall (and other) data with irregular patterns of missing values.
This rainfall data has a marked seasonal pattern, with less prominent trend and cyclical components. The sensitivity of long-range and short-range climate forecasts, generated using the Time Series Forecasting System, to alternative options within PROC EXPAND to impute missing values will be examined. Forecasts generated using different imputation methods will be compared.

INTRODUCTION

Missing values are a common problem in data analysis particularly when these data are drawn from secondary data sources. Much of the data analyzed in connection with water management is spatio-temporal, that is there variabiity over both space and time. PROC EXPAND in SAS/ETS® converts time series from one sampling interval or frequency to another and interpolates missing values in time series. PROC VARIOGRAM and PROC KRIGE2D in SAS/STAT® allow for dealing with missing values in spatial data using ordinary kriging. Ordinary kriging (OK) is a geostatistical approach to modeling. Instead of weighting nearby data points by some values. Water managers often have to deal with rainfall and other data that varies both spatially and temporally and which has missing values. This paper looks at the use of PROC EXPAND to fill in missing values in the time-series dimension (series by series over time) and compares parameter estimates and forecasts derived when different options for filling in missing values are specified.

DATA

The data used in the analysis were drawn from the South Florida Water Management District's DBHYDRO database. DBHYDRO is the District's corporate environmental database that stores hydrologic, meteorological, hydrogeologic, and water quality data. In particular daily rainfall data from 20 sites within five basins of the Lake Okeechobee Watershed are analyzed. This analysis is being conducted parallel to the analysis being conducted for the Lake Okeechobee Watershed Project of the Comprehensive Everglades Restoration Program.
http://www.evergladesplan.org/pm/projects/proj_01.cfm
The Lake Okeechobee Watershed is located in parts of five counties north of Lake Okeechobee in the southern portion of Florida. In the Lake Okeechobee Watershed Project document (http://www.evergladesplan.org/pm/projects/project_docs/pdp_01_lake_watershed/hyd_wq_part1r3.pdf), missing daily rainfall data were filled in "by transposing data from the nearest of the twenty rain gages with available data." In the present analysis, these data will be analyzed using PROC EXPAND to examine the effects of different methods of filling in missing values on the resulting output data set and the impacts of these differences in output data sets on forecasts of future rainfall estimated using the Time-Series Forecasting System.

LITERATURE REVIEW

Groundbreaking work in the field of missing data analysis was done by Schafer (1997) and Little and Rubin (1987). In general there are several patterns of missingness discussed in the literature on statistical Analysis with missing data:
1. Missing Completely at Random-
2. Missing at Random-
3. Non-Ignorable. (Chantala and Suchindran, no date)
Chantala and Suchindran note that there is no way to check if the conditions for "missing completely at random" or "missing at random"

Three desirable properties of an estimation method identified by Allison (2000) are
(1) yields "approximately unbiased estimators" of all parameters;
(2) yields "good estimates" of standard errors"; and
(3) is usable with any kind of data and any kind of analysis.
For multiple imputation to have these desirable properties, several conditions must obtain:
1. The data must be missing at random (MAR), meaning that the probability of missing data on a particular variable y can depend on other observed variables, but not on y itself, controlling for the other variables;
2. The model for generating the imputed values must be "correct" in some sense;
3. The model used for the analysis must "match up," in some sense, with the model used in the imputation.

The rainfall and other climatic data from the National Climatic Data Center and the South Florida Water Management District climatological database clearly are not "missing at random." The National Climatic Data Center has developed an elaborate set of codes for describing common data anomalies. (National Climatic Data Center, 1999). Details on the potential problems with NOAA precipitation data are presented in Kuligowski (1997). In many instances rainfall data are reported as accumulated values.over a period of days when data is not available for each individual day. Kuligowski and Barros have suggested the use of artificial neural networks to estimate missing rainfall data. In a paper presented at Western User's of SAS® Software Conference, Powers compared several SAS® approaches, including PROC EXPAND, to mean imputation. Powers noted that PROC EXPAND " is the simplest approach to missing data imputation in terms of programming. (Powers, 2002)" Similarly, Karp has noted, "Appropriate use of PROC EXPAND can provide users with rapid and statistically valid estimation of values of missing values as well as interpolation of higher-frequency observations from observations collected at lower frequencies. PROC EXPAND's data manipulation options also make its use preferable to the Data Step, especially when several operations (interpolate values of missing observations **and** aggregate from one time period to another) on the same series. (Karp, 2000).

DATA ISSUES

There are several important features of rainfall data that are important considerations in imputing missing values. First, rainfall is by definition non-negative. Thus, any technique used to fill in missing values should not yield negative values. Second, a large percentage of the daily rainfall values are 0. These data considerations need to be taken into account when imputing missing values. Figure 1 below shows a histogram of Peavine daily rainfall amounts. Peavine is selected as a representative climatic data station in Okeechobee County, Florida, north of Lake Okeechobee.

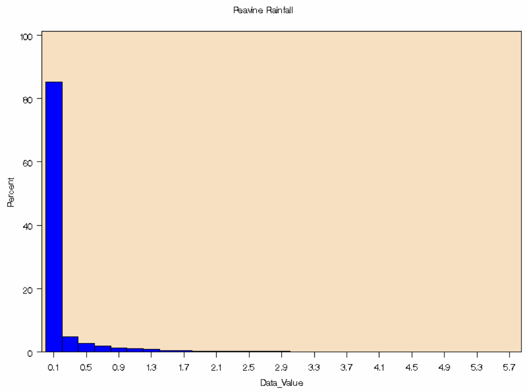

Figure 1: Histogram of daily Peavine rainfall amounts

Because of the high percentage of 0 values and the highly seasonal pattern of rainfall at the Peavine site, as shown in Figure I above, there are particular problems in using PROC EXPAND to fill in missing values, particularly in situations where there are long periods of consecutive missing values

METHODOLOGY

The primary concern in this paper is filling in missing values in the time-domain for multiple series using PROC EXPAND. Within PROC EXPAND there are several options for filling in missing values. In general, the syntax of PROC EXPAND is:

```
PROC EXPAND options ;
BY variables ;
CONVERT variables / options ;
ID variable ;
```

The input and output frequency are controlled by the FROM= and TO= options under the PROC EXPAND statement. To interpolate missing values in time series without converting the observation frequency, the TO= option is omitted. Options available under the PROC EXPAND Statement for converting the data series are given in the METHOD= option statement. The default is METHOD=SPLINE, which fits a cubic spline curve to the impute values. A cubic spline is defined as “A finite sequence of cubic polynomials defined on non-overlapping domains and connected at knots. “ A knot is defined as “A point in the domain space of a function where pieces of a fitted surface join.” The process is to construct a function that balances the twin needs of (1) proximity to the actual sample points, (2) smoothness. So a 'roughness penalty' is defined.” Splines are also referred to as piecewise polynomials.

Constraints, where specified, having one of the following values:

NOTANOT specifies the not-a-knot constraint and is the default.;

NATURAL specifies the natural spline constraint. The second derivative of the spline curve is constrained to be zero at the endpoint.

SLOPE= value
specifies the first derivative of the spline curve at the endpoint.

CURVATURE= value
specifies the second derivative of the spline curve at the endpoint. Specifying CURVATURE=0 is equivalent to specifying the NATURAL option.

The first constraint specification applies to the lower endpoint, and the second constraint specification applies to the upper endpoint. If only one constraint is specified, it applies to both the lower and upper endpoints.

Other possible values, besides SPLINE for the METHOD= option under PROC EXPAND are:
JOIN, STEP, AGGREGATE, and NONE.
In the JOIN method a continuous curve is fit to the data by connecting successive straight line segments. For point in time output series, the JOIN function is evaluated at the appropriate points. For interval total or average output series, the step function is integrated over the output intervals.
In the STEP method, a discontinuous piecewise constant curve is fit. For point-in-time output series, the STEP function is evaluated at the appropriate points. For interval total or average output series, the step function is integrated over the output intervals.
The AGGREGATE method performs simple aggregation of time series without interpolation of missing values. Since the primary concern of this paper is missing values, the AGGREGATE method is not examined in detail. When METHOD=NONE is specified, no interpolation is performed.

Once missing values are filled in using PROC EXPAND, the time-series forecasting system, in automatic mode, is used to generate forecasts of future rainfall. The time-series forecasting system in the Automatic Mode is applied both to the raw rainfall data and to the expanded data, with missing values filled in. One of the features of missing value imputation made with the various forms of the METHOD=SPLINE interpolation is that this method results in the imputation of negative values, even when there are no negative values in the input data set. The data from the Peavine climate station in Okeechobee County, north of Lake Okeechobee.

RESULTS

The descriptive statistics associated with the fifteen rainfall stations examine are shown in a handout. It can be seen that the mean daily rainfall for the fifteen stations examined ranges from approximately .118 inches per day to approximately .174 inches per day. The seasonality of the rainfall pattern is strong enough that the seasonal effects tend to overwhelm the day-to day fluctuations in rainfall, particularly given the high percentage of 0 daily rainfall events .Of the

An analysis-of variance-was conducted to test for the equality of average daily rainfall by month at the Peavine weather station. The code for this analysis is given below.

```
proc anova;
class month;
model data_value=month;
Means month/Duncan Tukey;
run;
```

This analysis showed that there was a highly significant (F=22.99) difference between months with the highest rainfalls in the summer months (June-September) and the lowest average rainfall in the winter and spring months (November through April). There was too much inter-day variation within any given month for reliable daily rainfall forecasts to be made using PROC EXPAND.. Figure 2 below depicts the mean daily rainfall at the PEAVINE station, with mean daily rainfall by month being shown on the vertical axis and month (with January being month 1) being shown on the horizontal axis. It can be seen that rainfall is highly concentrated in the summer months between May and September.

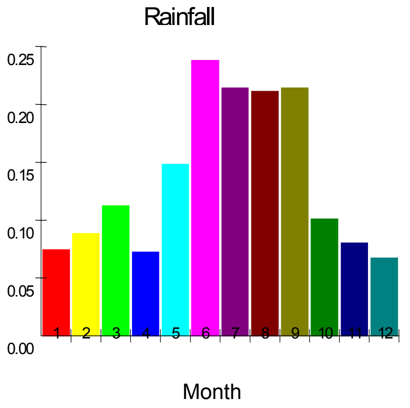

Figure 2: Mean daily rainfall by month at Peavine

Alternatives to PROC EXPAND for filling in missing values rely upon the existence of a moderately dense network of climate stations in South Florida, within SAS using various SAS® PROC's, including PROC VARIOGRAM and PROC KRIGE2D. In addition, a large number of SAS® macros are available for performing spatial analysis. Recently SAS® and ESRI , the producer of the ArcGIS line of spatial analysis software have created a joint product SAS® Bridge for ESRI. SAS® Bridge for ESRI. This product allows organizations to exchange spatial and attribute data as well as metadata between ArcGIS and the SAS® system. There is a large amount of literature on spatial interpolation of rainfall. Among the most common methods are: (1) inverse distance weighting and nearest neighbor, (2) polynomial trend surfaces and splines; (3) Kriging; (4) Likelihood analysis and Bayesian analysis; and (5) Neural networks. (Genton and Furrer, (1998), pp. 12-13). Genton and Furrer utilize "one of the fundamental principles of geostatistics: . . . observations, which are closely located in space, are more likely to be similar than observations which are far away. (Genton and Furrer, p.13) " Hutchinson (1995) and Hutchinson and Corbett (1996) have used thin plate smoothing splines within SAS PROC TPSPLINE for spatial rainfall interpolation.

Gallo and Sagnuolo (1998) proposed the use of "Fuzzy B-Splines" for spatial estimation of missing vales. For more general discussions of B-splines for interpolation of missing values see de Boor (1978) and Lee, Wolberg, and Shin (1997).

CONCLUSIONS

Day- to-day variations in .rainfall are large enough and irregular enough that PROC EXPAND yields relatively poor temporal interpolations of missing rainfall data. The first-order autocorrelation in daily rainfall at Peavine (.13545), while significantly different from 0, is too small for accurate imputation of missing values, based on past values alone. The spatial patterns of rainfall variability appear to provide a stronger basis for imputing missing values, given a sufficiently dense network of rainfall stations, than temporal patterns of rainfall variability.

REFERENCES

Allison, Paul D., (2000), "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research* 28, 301-309. <http://www.ssc.upenn.edu/~allison/MultInt99.pdf>

Chantala, Kim and C. Suchindran, (no date) "Multiple Imputation for Missing Data," Center for Population Studies, University of North Carolina, Chapel Hill. <http://www.cpc.unc.edu/services/computer/presentations/mi_presentation2.pdf>

Cressie, Noel A. (1993), *Statistics for Spatial Data*, Wiley, New York.

De Boor, Carl, 1978. *A Practical Guide to Splines*, Springer-Verlag, New York.

Gallo, Giovanni and Michela Spagnuolo, 1998. "Rainfall Estimation from Sparse Data With Fuzzy B-Splines," *Journal of Geographic information and Decision Analysis*, vol. 2, no. 2, pp. 194-203.

Genton, Marc G. and Reinhard Furrer (1998), "Analysis of Rainfall Data by Simple Good Sense: Is Spatial Statistics Worth the Trouble?," *Journal of Geographic information and Decision Analysis*, vol. 2, no. 2, pp. 12-17.

Hutchinson, M. F., (1995), "Interpolating Mean Rainfall Using Thin Plate Smoothing Splines. *International Journal of GIS,* pp. 306-403.

Hutchinson, M. F. and J. D. Corbett, (1996), Spatial Interpolation of Climate Data Using Thin Plate Smoothing Splines, Paper Prepared for the FAO Expert Consultation on the Coordination and Harmonization of Databases and Software for Agroclimatic Applications, Rome, 29 November-3 December, 1993,.

Karp, Andrew H. (2000) , "Working With SAS ® Date And Time Functions," Paper Presented at 2000 Northeast SAS® User's Group Meeting, http://www.ats.ucla.edu/stat/sas/library/nesug00/bt3007.pdf

Kuligowski, Robert J., (1997), "An Overview of National Weather Service Quantitative Precipitation Estimates," U. S. Department of Commerce, National Oceani and

Atmospheric Administration, Techniques Sevelopment Laboratory, TDL Office Note 97-4. http://205.156.54.206/pub/im/tdl97-4.pdf

Kuligowski, Robert J. and Ana P. Barros (1998), "Using Artificial Neural Networks to Estimate Missing Rainfall Data, *Journal of the American Water Resources Association,* v.34, No. 6, pp. 1437-1447.

Lee, Seungyong, George Wolberg, and Sung Yong Shin, 1997. Scattered Data Interpolation With Multilevel B-Splines, *IEEE Transactions on Visualization and Computer Graphics* ,v. 3. No. 3., pp. 228-244.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

National Climatic Data Center, (1999), *Climatological Data, Florida, January, 1999,* "Reference Notes," Asheville, North Carolina.

Powers, Keiko, (2002), "Comparisons of Several SAS Approaches to Mean Imputation for Time-Series Data with Missing Data Points," paper presented at Western Users' of SAS® Software Conference. http://www.wuss.org/Conference/papers/CC11.pdf

Schafer, J. L., (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London

South Florida Water Management District and U. S. Army Corps of Engineers, 2003, *Comprehensive Everglades Restoration Plan, Lake Okeechobee Watershed Project, Project Document, Section 6.0, Hydrologic /Water Quality Characterization of the Watershed.* <http://www.evergladesplan.org/pm/projects/project_docs/pdp_01_lake_watershed/hyd_wq_part1r3.pdf>