A Nonparametric Nonlinear Time Series Forecasting Model for Hydrologic Variables

(수문변량의 예측을 위한 비매개변수적 비선형 시계열 기법)

Young-Il Moon[1], U. Lall[2], B. Rajagoplan[3], and M. Mann[4]

1. Introduction

Most hydrologists are familiar with linear regression and its use for developing a relationship between two or more variables. The general linear model (where the variables are presumed to be linearly related after applying some predetermined transform) is used as a building block in many activities ranging from spatial surface estimation, missing value imputation, sediment load estimation to autoregressive time series modeling. Often, such a procedure is not very satisfying. The choice of an appropriate transform to use may not be obvious, and scatterplots of the data may not visually support the assumed model. An alternative to such regression approaches that is capable of representing relatively complex relationships between variables, through local or pointwise approximation of the underlying function, is presented here.

There has been a surge in the development and application of nonparametric regression (Müller, 1987, Eubank, 1988, Scott, 1992, Moon and Lall, 1994(a) and (b)) and density estimation methods(Silverman, 1986, Moon et al., 1993 and Lall et al., 1993) in the last decade as computational resources have become more accessible. Some monographs that make this literature accessible are by Silverman (1986), Eubank (1988), Härdle (1989). Examples of nonparametric estimators include orthogonal series expansions, moving averages, splines, kernel, and nearest neighbor estimators. Some attributes of these methods are :

(1) The estimator can often be expressed as a weighted moving average of the observations.

(2) The estimates are defined locally or using data from a small neighborhood of each point of estimate.

(3) Consequently, they can approximate a wide class of target, underlying functions.

(4) The nonparametric estimator has parameters that control the local weights and the size of the neighborhood used for estimation. However, unlike linear or parametric regression, where the parameters (e.g., intercept and slope) are sufficient to provide an estimate at any point, the nonparametric estimator needs the observations in the neighborhood to provide an estimate at any point. For example the parameter of a moving average is the number of points (e.g., 3) to average over. One still needs the three points surrounding the point of estimate to report the answer. By contrast, once a linear regression has been evaluated, the parameters are all one

[1]Department of Civil Engineering, Seoul City University

[2]Utah Water Research Laboratory, Utah State University

[3]Lamont-Doherty Earth Observatory, Columbia University

[4]Department of Geology and Geophysics, Yale University

needs to provide an estimate at any point. In the parametric case, the overall behavior is of an assumed form (e.g., linear or log-linear), and the parameters of this global form are all one needs to estimate. The parameters of a nonparametric model thus have a different role, since no global form is assumed, and the parameters merely control how local averages are to be formed.

From an one dimensional time series we can reconstruct a multivariate space. Nonparametric estimates of mutual information (Moon et al., 1995(a) and 1996) are used to select appropriate lags (at which the successive values are somewhat independent) for this state space reconstruction.

Background information on locally weighted polynomial regression is provided in the section 2. An application of local regression to time series forecasting is presented in the section 3.

2 Locally Weighted Polynomial Regression

The locally weighted polynomials consider the approximation of the target function through a Taylor series expansion of the unknown regression function in the neighborhood of the point of estimate (Moon et al., 1995(b)). Cross-validatory procedures for the selection of the size of the neighborhood over which this approximation should take place, and for the order of the local polynomial to use are used (Moon et al., 1995(b) and Moon and Lall 1995). The detail of this nonparametric regression approach was shown by Lall (1994) and Moon et al. (1995(b)). Procedures for selecting the smoothing parameters and estimating prediction limits for the local regression estimates was presented in Moon and Lall (1995).

In this section we present only the main idea of a locally weighted polynomials to save the space. The reader is referred to Moon et al. (1995(b)) for the more detail. It is helpful to begin with a simple univariate example. Consider the estimation of the function $f(x) = \text{Sin}(x)e^{-0.2x}$, from the data (x_i, y_i), generated such that the x_i are equally spaced values from 0 to 10, and the y_i are then generated as $f(x_i) + e_i$, $i=1..100$, and $e_i \sim N(0,0.1)$. This data set, the true, underlying function, and three local regressions are shown in Figure 1. The data (small circles) is 100 points generated from $y = \text{Sin}(x)e^{-0.2x} + N(0,0.1)$. The true function is shown as the solid line. The thin dotted line is a linear regression through the full data. It shows the bias incurred if a neighborhood of size 100 is used at any point. The thin dashed line is a quadratic fit through the full data. It shows that the bias may be reduced by going to a higher order polynomial. Estimates are considered at 3 points, x=2.2, 4, and 8. Local linear fits with 10 neighbors are used at the first two points, and a locally quadratic fit with 25 neighbors is used at x=8. The data in each neighborhood are shown with large circles. The local fits are shown as thick solid lines. The approximation at the first two points is quite good. The estimate at x=8 is poorer. Since we know the true function, we can use those values with the local quadratic fit. The resulting fit at x=8 is shown as a thick dashed line. It is seen to coincide with the target function. Consequently, the approximation error at x=8 is a consequence of the local noise realization.

We observe that the quality of the local regressions is quite good. The higher order (quadratic) fit is less biased than the linear, when 100 neighbors are used (global fit). The bias decreases substantially as the size of the neighborhood is reduced from 100 to 10 or 25 points. However, the local regressions can exhibit increased variability of estimate due to the reduced

sample size. This is exacerbated as one moves to a higher order polynomial fit with the same number of data points.

There is a trade-off between bias and variance as one changes the order of the local polynomial and the number of points used to fit it. Parameter selection approaches are usually based on an estimate of the mean square error (MSE) of the estimation scheme. Moon et al. (1995) introduced the use of a Local Generalized Cross validation (LGCV) score that uses data directly from the local regression at the point of estimate. Methods for parameter selection and the estimation of error variances was presented in Moon et al. (1995(b)).

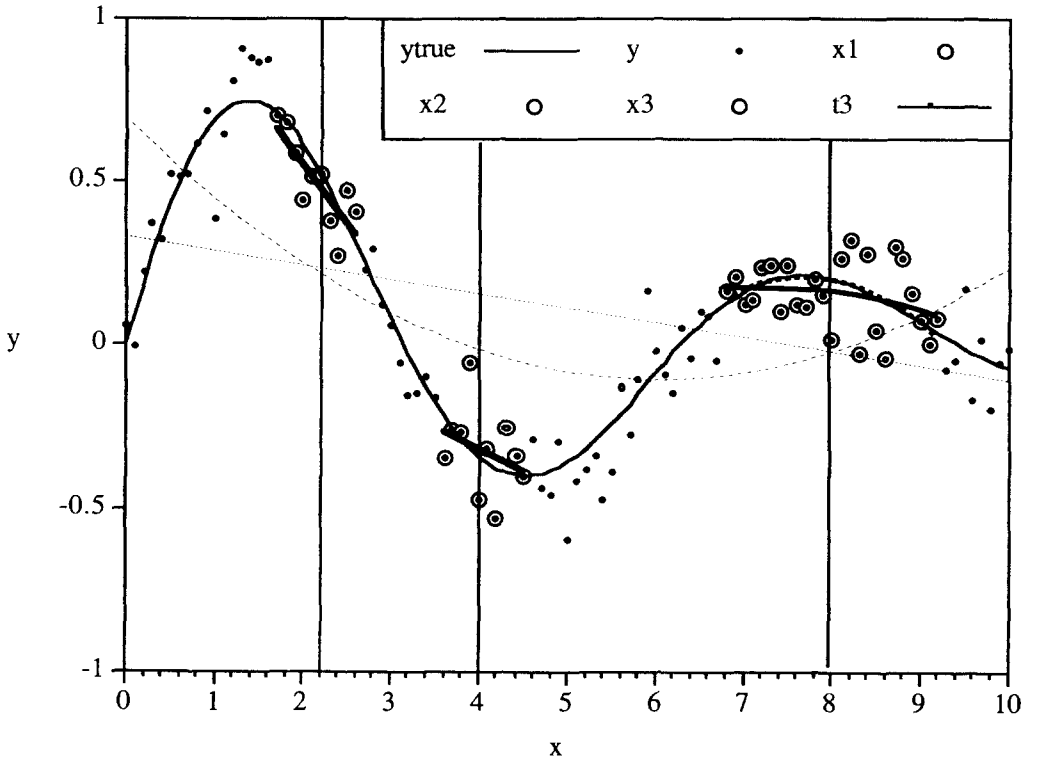

Figure 1. Illustration of local linear and local quadratic regression, with the weights w_{ii} =1/k.

3. Application

The primary application we consider in this paper is the forecast of (1) the volume of the Great Salt Lake (GSL) at key points in time from its 1847-1993, biweekly time series and (2) Southern Oscillation Index. First, we considered blind forecasts of the GSL volume from different states for 1 year into the future from the date of forecast. The forecasted values are then compared with the volumes that were actually recorded subsequently. They are presented in Figure 2. The lag τ was selected as 10 as in the range of the first minimum of the average

mutual information (Moon et al., 1995(a) and 1996) and it was based on experimentation to get the best predictions (min predictive squared error). An embedding of m=5 was selected after experimentation with various values in the range 1 to 9. This value corresponded to the one that most commonly minimized LGCV. We searched over k1=50 to k2=150 nearest neighbors and typically selected 120 to 150. Locally linear and quadratic fits were considered. Typically a linear fit was selected. The results are discussed in the Figure 2. Similar results were obtained in Lall et al. (1996) and Abarbanel et al. (1996).

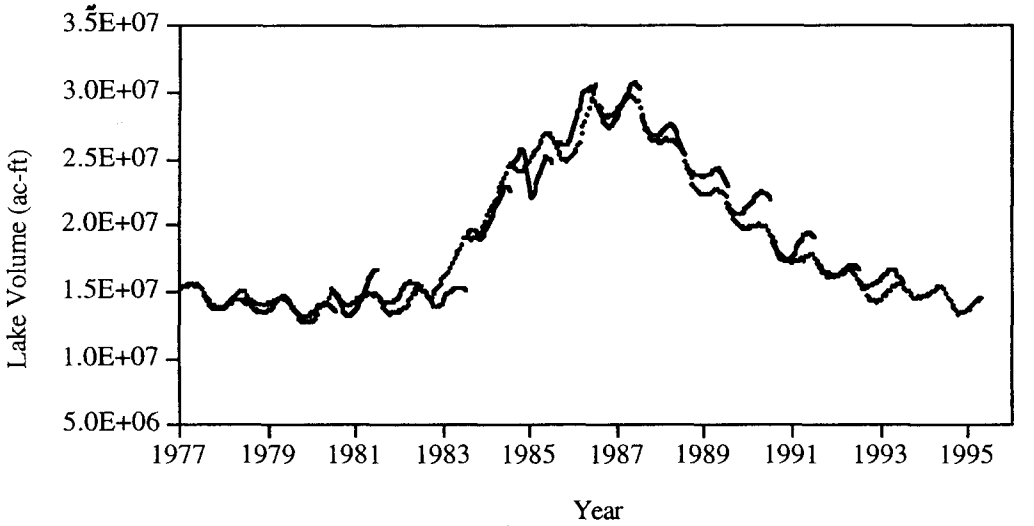

Figure 2. A sequence of 1 year blind forecasts of the GSL from July 1977 to Jan 1996. The dots represent the observed GSL time series. The solid lines represent 12 forecasts, one for each month of the next year. Only data available up to the beginning of the forecast period is used for fitting and forecasting.

The SOI is a measure of El Nino Southern Oscillation (ENSO) pattern. The SOI data consist of time series of difference in the monthly mean Sea Level Pressure(SLP) between Tahiti (150W, 13S) and Darwin (130E, 18S) from Jan. 1899 to Dec.1993 (Moon, 1994 and Moon and Lall, 1996). The 2 yr filtered low frequency time series of the normalized SOI time series explain well the trend of the El Nino events and La Nina events. The normalization consists in by subtracting the mean SOI at each month and dividing the monthly anomalies obtained by the corresponding standard deviation. Negative episodes correspond to El Nino events and positive episodes identify La Nina events. In Figure 3, Jan.1984-Dec.1989 (yp1) and Jan.1989-Dec.1994 (yp2) blind forecasts of SOI (τ=6 and M=5) are presented, using only data from Jan 1899 to Dec. 1993 and from Jan 1899 to Dec. 1994. The behaviors in the forecast and 2 yr. filtered data are coincidental. Figure 4 shows Jan 1994-99 blind forecasts of SOI (τ=6 and M=5), using only data from Jan 1899 to Dec. 1993.

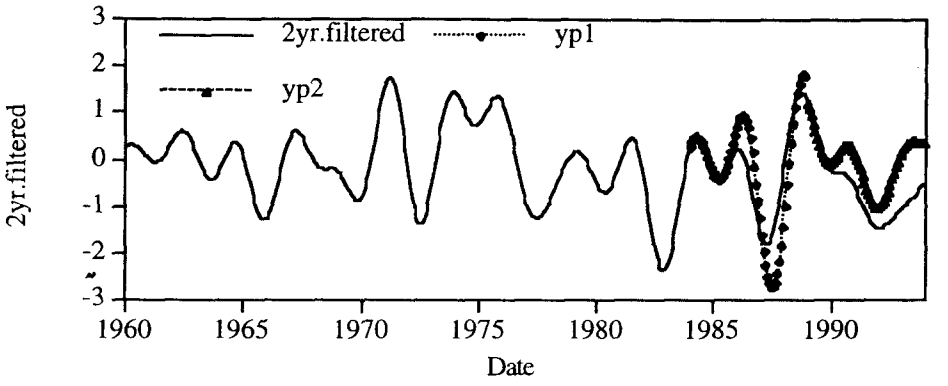

Figure 3. Jan.1984-Dec.1989 (yp1) and Jan.1989-Dec.1994 (yp2) blind forecasts of SOI (τ=6 and M=5), using only data from Jan 1899 to Dec. 1983 and from Jan 1899 to Dec. 1988 respectively.

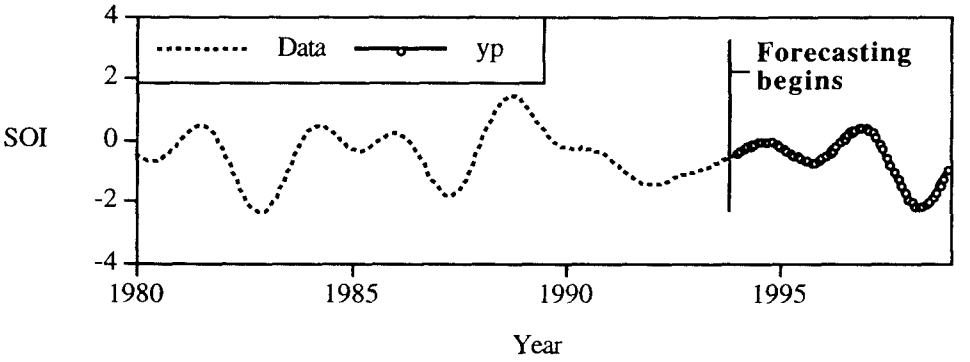

Figure 4. Jan 1994-99 blind forecasts of SOI (τ=6 and M=5), using only data from Jan 1899 to Dec. 1993.

4. Conclusions

A locally weighted polynomial regression methodology for approximating nonlinear regressions was introduced in this paper.

The methodology presented was then applied to the forecast of selected time series with encouraging results. These methods are still evolving. We can expect improvements in procedures for estimating prediction intervals and for selecting parameters of the method. Algorithmic improvements for more efficiently exploiting multivariate data structures are also to be expected.

From a hydrological point of view, these methods provide new directions for the exploration of data as well as the possibility of dramatic improvements in time series forecasting and spatial surface reconstruction.

Acknowledgements

Partial support of this work by the USGS grant # 1434-92-G-226 and NSF grant # EAR-9205727 is acknowledged.

References

Abarbanel, H.D.I., U. Lall, M. Mann, Young-Il Moon, and T. Sangoyomi, Nonlinear Dynamics and the the Great Salt Lake: A Predictable Indicator of Regional Climate, Energy, in press, 1996.

Eubank, R., Spline smoothing and nonparametric regression, Marcel Dekker, New York, 1988.

Härdle, W., Applied Nonparametric Regression, in Econometric Society Monographs, pp. 333, Cambridge University Press, Cambridge, 1989.

Lall, U., Young-Il Moon, and K. Bosworth, Kernel Flood Frequency Estimators: Bandwidth Selection and Kernel Choice, Water Resources Research 29 (4), 1003-1015, 1993.

Lall, U., Nonparametric function estimation: Recent hydrologic contributions, Contributions in hydrology, U.S. National report to the IUGG 1991-1994, 1994.

Lall, U., T. Sangoyomi, H.D.I. Abarbanel, Nonlinear dynamics of the Great Salt Lake: Nonparametric short term forecasting, Water Resources Research, in press, 1996.

Moon, Young-Il, U. Lall, and K. Bosworth, A Comparison of Tail Probability Estimators, Journal of Hydrology 151, 343-363, 1993.

Moon, Young-Il and U. Lall, A Kernel Quantile Function Estimator for Flood Frequency Analysis, Water Resources Research, Vol. 30(11), 3095-3103, 1994(a).

Moon, Young-Il and U. Lall, A Kernel Quantile Function Estimator for Flood Frequency Analysis, Extreme Values: Floods and Droughts, Kluwer Academic Publisher, Ontario, Canada, 1994(b).

Moon, Young-Il, Low Frequency Relationships Between Great Salt Lake and Atmospheric Variability, American Geophysical Union's Fourteenth Annual Hydrology Days, 305-316, 1994.

Moon, Young-Il, Rajagopalan, B., and U. Lall, Estimation of Mutual Information Using Kernel Density Estimators, Physica Review E, V52(3), 2318-2321, 1995(a).

Moon, Young-Il, Lall, U., and K. Bosworth, Locally weighted polynomial regression:Parameter choice and application to forecasts of the Great Salt Lake, WP-95-HWR-UL/014, Utah State University, Logan, 1995(b).

Moon, Young-Il and U. Lall, Nonparametric short term forecasts of the Great Salt Lake using atmospheric indices, WP-95-HWR-UL/015, Utah State University, Logan, 1995.

Moon, Young-Il and U. Lall, Atmospheric Flow Indices and Interannual Great Salt Lake variability, Journal of Hydrologic Engineering in American Society of Civil Engineers, Vol.1 (2), 1996.

Moon, Young-Il, Rajagopalan, B., and U. Lall, Nonlinear Dependence in Hydrologic Time Series through Mutual Information, Water Resource Research, submitted, 1996.

Müller, H.-G., Weighted local regression and kernel methods for nonparametric curve fitting, J. Amer. Statist. Assoc. 82, pp231-238, 1987.

Scott, D.W., Multivariate Density Estimation, John Wiley and Sons, New York, 1992.

Silverman, B.W., Density estimation for statistics and data analysis, Chapman and Hall, London, 1986.