GI DA JOURNAL OF GEOGRAPHIC INFORMATION AND DECISION ANALYSIS

Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 44-51, 1998

Locally Weighted Polynomial Estimation of Spatial Precipitation

Balaji Rajagopalan
Lamont-Doherty Earth Observatory of Columbia University, P.O. Box, 1000, Rt/9W, Palisades, NY 10964-8000, USA
rbala@rosie.ldgo.columbia.edu

Upmanu Lall
Utah Water Research Lab. and Department of Civil and Environmental Engineering , Utah State University, Logan UT 84322-8200, USA
ulall@cc.usu.edu

ABSTRACT **We demonstrate the utility of locally weighted polynomial regression, a nonparametric technique for surface estimation discussed in Lall *et al*. (1995), for the spatial estimation of precipitation surface, with data related to the Chernobyl nuclear power plant accident. The method uses multivariate, locally weighted polynomial regression with temperature or precipitation as the dependent variable and a feature vector (location, elevation and other attributes) of explanatory variables. Localization of the regression is achieved by using *k* nearest neighbors of the point of estimate and a monotonic distance based weight function. Generalized cross validation is used to pick the order of the polynomial fits, as well as the number of neighbors to use. Pointwise estimates of predictive risk are also obtained.**

KEYWORDS: **locally weighted polynomial regression**
ACKNOWLEDGEMENTS Support for the first author by Grants ICSIO CU 01556601D, NA36GP0074-02 and NA56GP0221 is thankfully acknowledged.

Contents

1. Introduction

There has been renewed interest recently in developing gridded estimates of precipitation and temperature fields to provide inputs to spatially distributed hydrologic, ecologic and

many other management models, and to permit a useful comparison with the output of numerical models of atmospheric circulation. Such estimates are especially difficult to obtain in mountainous environments where there can be large changes in these fields, were data are sparse and are usually restricted to lower elevations.

Many 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 for spatial surface estimation. In the generalized linear regression framework, kriging (Cressie, 1991) is a widely used technique for such estimation problems. Kriging is a linear estimator that interpolates the data using weights based on distance and an inferred parametrically specified spatial correlation structure (also known as the variogram). An advantage of using kriging is that it can provide an estimate of the pointwise mean square error of estimate as a byproduct. Cressie (1991) argues that kriging provides an efficient estimator when the sampling locations are clustered and the residuals of the fit are uncorrelated. This may be true if the correct parametric form for the "trend" function and the "variogram" is specified. Usually, these are not known and are hard to specify properly given a single realization.

Yakowitz and Szidarovsky (1985), and Owosina (1992) compare the properties and performance of kriging to some nonparametric regression estimators and find the nonparametric estimators to perform better than kriging. Satagopan and Rajagopalan (1993) compare kriging with smoothing splines and local polynomial estimators (Cleveland, 1979) for estimation of precipitation surface over a region in the Willamette River basin in the state of Washington and find the nonparametric methods to be better than kriging. Lall *et al*. (1995) developed a nonparametric local polynomial functional estimation technique that is conceptually simple and computationally fast.

The local polynomial estimation scheme is first briefly described. The algorithm is then presented, followed by the application to Chernobyl data. Discussion of the results concludes the paper.

2. Local Polynomial Estimation Technique

Local polynomial estimators for function estimation are one of the many nonparametric estimators that are in practice. Nonparametric or local fitting estimators, with weaker assumptions than the parametric estimators (e.g. kriging), adapt better to heterogeneous and non-stationary data sets. Some attributes of these estimators 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. Consequently, they can approximate a wide class of target, underlying functions.
(3) The nonparametric estimator has parameters that control the local weights and the size of the neighborhood used for estimation.

Some monographs that make this literature accessible are by Silverman (1986), Eubank (1988), Härdle (1989). Owosina (1992) also has a fairly detailed description of various nonparametric function estimation schemes.

2.1 Local Polynomial Scheme

The description of the technique is kept short and for details the reader is referred to Lall *et al*. (1995). Consider a general regression model given as :

$$\mathbf{y_i} = \boldsymbol{f}(\mathbf{x_i}) + \mathbf{e_i}, \quad i = 1,...,n \qquad (1)$$

where **x** is a vector of M explanatory variables, **y** is the "response" variable, ***f***(.) represents the underlying functional relationship between **y** and **x**, $\mathbf{e_i}$ are noise or measurement errors, that may or may not depend on $\mathbf{x_i}$, and **n** is the number of observations.

An approach for the pointwise estimation of the unknown function ***f*(.)** from data, based on a local Taylor series expansion of ***f*(x)** at the point of estimate **x**, was proposed by Macauley (1931). Cleveland (1979), Cleveland and Devlin (1988) and Cleveland *et al.*(1988) pioneered this idea into a statistical methodology for local approximation of functions from data. Recently (e.g., Hastie and Loader (1993), Fan and Gijbels (1992), Müller(1987)), these methods have been recognized as very useful generalizations of kernel regression (weighted moving averages). Applications to a suite of statistical estimation problems are emerging. The reader is referred to a recent monograph by Wand and Jones (1995) for general background on the methods.

Generally, the strategy is to choose a certain number, k, of nearest neighbors (in terms of Euclidean distance) of the estimation point x, and to form the estimate $\hat{f}(x)$ through a locally weighted, polynomial regression over the (**x**, **y**) data that lie in the neighborhood. Consider the general regression model described in (1). The sampling locations x_i are usually not regularly spaced. We assume the e_i are uncorrelated, mean zero, random variables, assumed to be approximately identically distributed in the k nearest neighborhood of the point of estimate. Then, the locally weighted polynomial regression at each point of estimate x_l*, l =1,□,np, given a (n by M) data matrix x and a (n by 1) response vector y, is obtained through the solution of the weighted least squares problem:

$$\mathbf{Min}\ (\mathbf{y}_l - \mathbf{Z}_l \beta_l)^T \mathbf{W}_l (\mathbf{y}_l - \mathbf{Z}_l \beta_l) \qquad (2)$$

where the subscript l recognizes that the associated element is connected with the point of estimate $\mathbf{x}_l$*; $\beta_l$are estimates of the coefficients of the terms in the basis defined by $\mathbf{Z}_l$; $\mathbf{Z}_l$ is a matrix formed by augmenting **x**, with columns that represent the polynomial expansion of **x** to degree p (including cross product terms if desired); $\mathbf{W}_l$ is a k by k diagonal weight matrix with elements

$$w_{ii,l} = K(u_{i,l}) / \sum_{j=1}^{k} K(u_{j,l})$$

where $u_{i,l} = d_{i,l}/d_{k,l}$; $d_{i,l}$ is the distance from $\mathbf{x}_l$* to $\mathbf{x}_i$ using an appropriate metric, and K(.) is a weight function. We have implemented a bisquare kernel ($K(u)=15/16(1-u^2)^2$). The latter is recommended by Scott (1992) because of its smoothness properties. The matrix $\mathbf{Z}_l$ and vector $\mathbf{y}_l$ are defined over the k nearest neighborhood of $\mathbf{x}_l$*. Singular Value Decomposition (SVD) using algorithms from Press (1989) is used to solve the linear estimation problem resulting from (2).

The reader may note that we are in the familiar territory of linear regression, and will hence have the usual statistical tools available to us. The coefficients β_l are obtained as:

$$\beta_l = (\mathbf{Z}_l^T \mathbf{W}_l \mathbf{Z}_l)^{-1} \mathbf{W}_l \mathbf{y}_l \qquad (3)$$

The resulting estimate of $\hat{f}_{(\mathbf{x}l^*)}$ is then:

$$\hat{f}(\mathbf{x}_l^*) = \mathbf{z}_l \beta_l \qquad (4)$$

where $\mathbf{z}_l$ is the d by 1 vector formed by augmenting $\mathbf{x}_l$ with polynomial terms to order p, and retaining the terms for which βj are found to be significantly different from 0.

2.2 Parameter selection **k** *&* **p**

A variety of estimators of Predictive Mean Square Error of $\hat{f}$ have been proposed in the literature. Cleveland and Devlin (1988) considered Mallows Cp in their work on local polynomial regression.Li(1985) discusses the theoretical foundations of this and other measures such as Ordinary and Generalized Cross Validation, the Finite Prediction Error, the AIC and the BIC. Of these the Generalized Cross Validation (GCV) statistic proposed by Craven and Wahba (1979) is of particular interest since it has performed well (Härdle 1984, 1989) in practical applications. It is defined as :

$$GCV(\hat{f}) = MSE(\hat{f}) / (n^{-1} tr[\mathbf{I}-\mathbf{H}])^2 \quad (5)$$

where the Mean Square Error MSE

$$(\hat{f}) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{f}(\mathbf{x}_i))^2 \quad (6)$$

H is the influence matrix defined through

$$\hat{f} = \mathbf{H}\mathbf{y} \quad (7)$$

I is the identity matrix, and tr[.] represents the trace of the matrix.

Note that (7) represents a linear estimator, and the i^{th} diagonal element of **H** can be thought of as the "weight" of that data point on the estimate at that point. Eubank (1988, p. 406) states that for linear regression (local or global, and on the raw variable or a polynomial in it) it is easy to show that $0 \leq h_{ii} \leq 1$. Thus if h_{ii} is 1, and the other h_{ij} are 0, we see that we have 0 degrees of freedom, and the estimate at each point is simply the original data, i.e. the model completely overfits or undersmooths. The corresponding MSE is zero, and the GCV is infinity. On the other hand, if all the h_{ij} are equal, the estimate at every point is the sample average of the y_i. The degrees of freedom are (n-1), since we fit one parameter, and the MSE may be large if $f(.)$is not a constant. For the situation were n is large, MSE and GCV will approach each other in magnitude. Consider also the case where the h_{ij} are equal for the k nearest neighbors of a point and 0 elsewhere. In this case we may approximate $f(.)$ better since we form a moving average of y values, and hence have a lower MSE. However, the degrees of freedom will only be (k -1), and the GCV may be larger. The denominator in equation (5) consequently has the role of a penalty for the effective number of parameters used in fitting the model. The effective number of parameters is determined by the number of neighbors and the number of terms in the local polynomial.

The motivation for using $GCV(\hat{f})$ comes from a theorem proved by Craven and Wahba (1979). They showed that $GCV(\hat{f})$ is a nearly unbiased estimator of the predictive risk, as long as the degrees of freedom are sufficiently large.

We shall consider the whole data set for parameter selection. The global GCV (GGCV) can be estimated after performing n local regressions at each data point $\mathbf{x}_i$ (i = 1,..,n), as:

$$GGCV(\hat{f}) = (\sum_{i=1}^{n} e_i^2 / n) / (1 - \sum_{i=1}^{n} h_{ii} / n)^2 \quad (8)$$

where h_{ii} is estimated from equation (7), and where $e_i = y_i - \hat{f}(\mathbf{x}_i)$

One can select appropriate values of k and p, as the minimizers of the GGCV value computed in equation (8) for each combination of k and p. These would be the values of k and p that would do well on the average. However, in certain situations (e.g., where the curvature of the target function varies over the data, and where the variance of the noise varies over the range of the data), one may wish to make such choices locally at the point of estimate.

Lall *et al.* (1995) introduced the use of a local GCV score that uses data directly from

the local regression at the point of estimate. In this case the errors $e_{i,l}$ are the residues of the model fitted over the k nearest neighbors of the point $\mathbf{x}_l$*, and $\mathbf{W}_l$ is the corresponding weight matrix. The trace of the matrix **H** in this case is simply d, the number of coefficients fitted. The local GCV (LGCV) score is then given as:

$$LGCV_l(\hat{f}) = (e_l^T \mathbf{W}_l e_l) / ((k - d)/k)^2 \quad (9)$$

The appropriate values of k and p can then be obtained as the ones that minimize the local GCV score for the local regression. The $LGCV_l$ value also provides insight into the local predictive error variance.

3. Application

We applied the technique described above to the precipitation data of 100 sample points. We searched for the k between 8 to 100 in steps of 2 and p in 1 to 2 on GGCV criteria. The optimal choice of k came out to be k = 32 and p = 2. We then estimated the precipitation at the 367 locations. The auto correlation function (ACF) of the residuals is shown in Figure 1(a). The residuals are not correlated significantly. The histogram of the residuals in Figure 1(b) shows they are normally distributed. These are two main assumptions (i.e., normality of the residuals and independence) in the regression context, and they both seem to be satisfied well. Plot of estimated versus the observed values is shown in Figure 1(c). It can be seen that the scatter falls pretty much around a straight line. The correlation between the observed and estimated values is 0.8, indicative of a very good fit. In Figure 1(d) we show the plots of k vs GGCV for p = 1 and 2 to show the nature of the GGCV function for this data set.

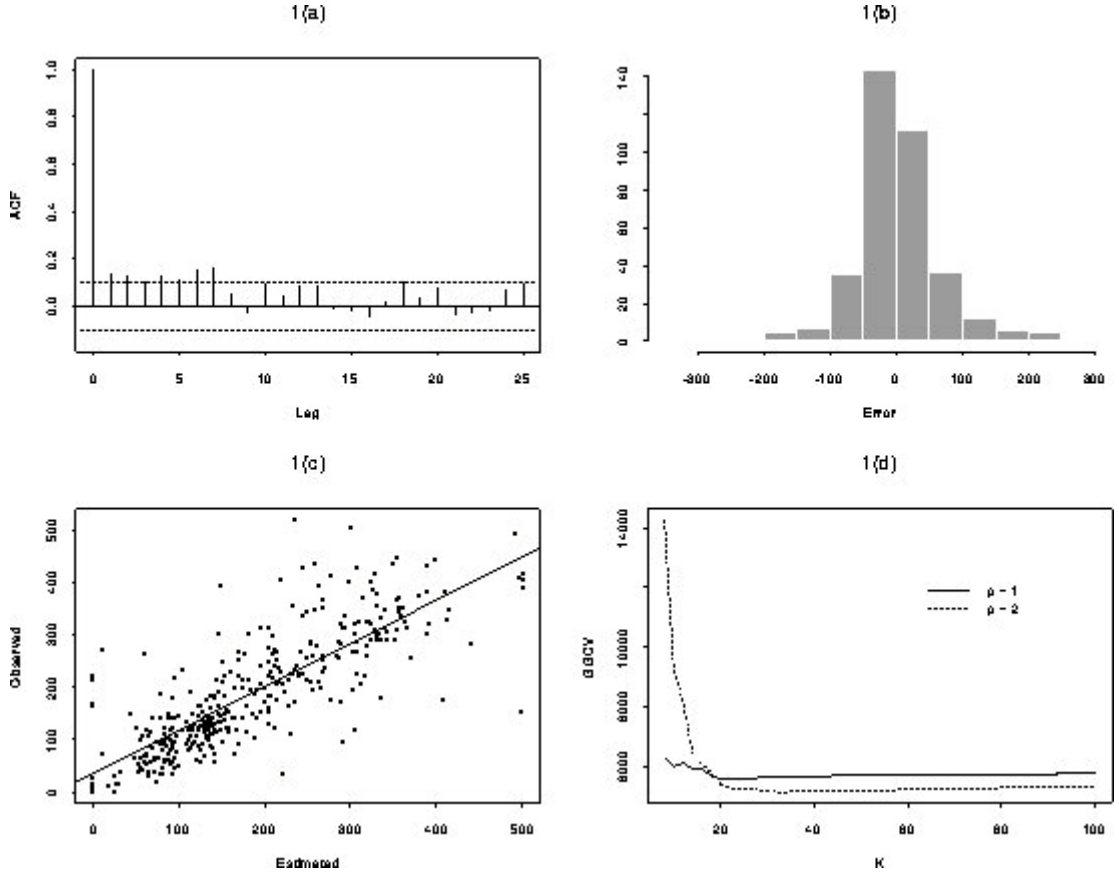

Figure 1(a): ACF of errors; (b) Histogram of errors; (c): Scatter plot of Observed vs Estimated values. The solid line is the best fit between them; (d): GGCV vs k plot for p = 1 and p = 2.

Some general statistics of the observed and estimates are presented in Table 1. The measures of central tendency (mean and median) are very well reproduced by the estimates. However, the variance in the estimates is very low and consequently the maximum observations are not well estimated. Also notice that the mean of the estimates of the highest ten values is low and variance high.

Table 1 Statistics of observed and estimates

Statistic Observed /Estimate	
Minimum	0. 0.
Maximum	517 501
Mean	185 181
Median	162 152
Variance	12358 11975
Mean of ten highest values	456 333*
Var. of ten highest values	1193 6400*
RMSE	67 85
Bias in errors	14 78
Absolute Mean errors	45

*Obtained from the estimates of the same locations of the observed ten highest values

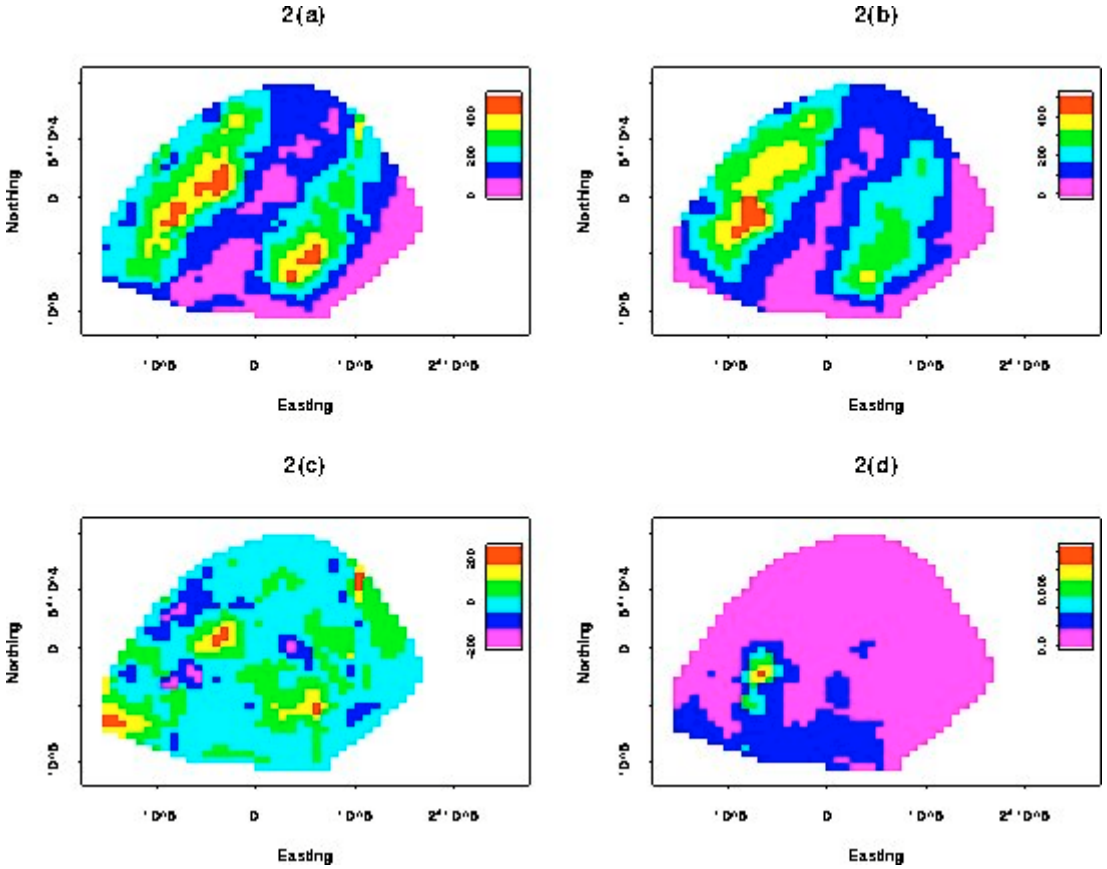

Figure 2(a): Image of observed values; (b) Image of estimates; (c) Image of errors; (d) Image of LGCV estimates

The images of observed values and those of the estimated values are shown in Figure 2(a) and (b) respectively. The gross features are very well reproduced. The areas with very high precipitation tend to be smoothed out in the estimates. This is

understandable - the method constructs the estimate from a small neighborhood and in regions of high precipitation (typically at higher elevations) the neighborhood includes low precipitation (typically at lower elevations) regions as well, this leads to a significant smoothing and hence reduced precipitation at locations of high precipitation. This can be easily rectified by including elevation as the third independent variable if data exists.

The image of the errors (Figure 2c) also shows that the errors are high in regions of high precipitation. We also compute LGCV (Equation 9), a measure of local error of the estimate. It is shown in Figure 2(d). The higher the LGCV values the higher the estimation error and this too indicates that the LGCV is high in regions of high precipitation. This may suggest the use of a log transform prior to regression. However, given that many values of precipitation are zero, we chose not to do this for the current application.

References

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots, *Journal of American Statistical Association*, 74 (368), 829-836.

Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: an approach to regression analysis by local fitting, *Journal of American Statistical Association.*, *83* (403), 596-610.

Cleveland, W. S., Devlin, S. J, and Grosse, E. (1988) Regression by local fitting, *Journal of Econometrics*, *37*, 87-114.

Craven, P. and Wahba, G. (1979) Smoothing noise data with spline functions, *Numerische Mathematik*, 31, 377-403.

Cressie, N. (1991) *Statistics for spatial data*, New York, N.Y.: John Wiley & Sons Inc.

Eubank, R. (1992) *Spline smoothing and nonparametric regression*, New York, N.Y.: Marcel Dekker.

Fan, J. and Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, 20, 196-216.

Härdle, W. (1984) *How to determine the bandwidth of some nonlinear smoothers in practice*, Lecture Notes in Statistics 26, pp163-184, Springer-Verlag.

Härdle, W. (1989) Applied nonparametric regression, in *Econometric Society Monographs*, Cambridge: Cambridge University Press, , pp. 333

Hastie, T. J. and Loader, C. (1993) Local regression: automatic kernel carpentry (with discussion), *Statistical Science,* 8, 120-43.

Lall, U., Bosworth, K. and Owosina, A. (1995) Local polynomial estimation of spatial surfaces, *Computational Statistics*, (to appear)

Li, K.-C. (1985) From stein's unbiased risk estimates to the method of generalized cross validation, *Annals Statistics.* 13 (4), 1352-1377.

Macauley, F. R. (1931) *The smoothing of time series*, New York: National Bureau of Economic Research.

Müller, H.-G. (1987) Weighted local regression and kernel methods for nonparametric curve fitting, *Journal of American Statistical Association* 82, 231-238.

Owosina, A. (1992) Methods for assessing the space and time variability of groundwater data, MS thesis, Utah State University, Logan, UT.

Press W. (1989) *Numerical recipes* : *the art of scientific computing*. Cambridge: Cambridge University Press.

Satagopan J. and Rajagopalan B. (1994) Comparing spatial estimation techniques for precipitation analysis, In: K. Hipel, A. I. McLeod, U.S. Panu and V.P. Singh, *Time Series Analysis and Forecasting*, Amsterdam: Kluwer, Vol. 3, 317-330.

Scott, D.W. (1992) *Multivariate density estimation*, New York, N.Y.: John Wiley and Sons.
Silverman, B. W. (1986) *Density estimation for statistics and data analysis*, London: Chapman and Hall,
Stuart, A. and Ord, J. K. (1991) *Kendall's advanced theory of statistics,* New York: Oxford University Press.
Yakowitz, S. J. and Szidarovsky, F. (1985) A comparison of kriging with nonparametric regression methods, *Journal of Multivariate Analysis*, 16(1), 21-53.
Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing*, London: Chapman and Hall.

JGIDA vol. 2, no. 2
JGIDA Home