GI DA JOURNAL OF GEOGRAPHIC INFORMATION AND DECISION ANALYSIS

Journal of Geographic Information and Decision Analysis, vol. 2, no. 2, pp. 34-43, 1998

Nonparametric Spatial Rainfall Characterization Using Adaptive Kernel Estimator

Alaa Ali

South Florida Water Management District, 3301 Gun Club Road, West Palm Beach FL 33406

aali@sfwmd.gov

ABSTRACT **A nonparametric statistical tool based on kernel function estimation is developed for spatial rainfall characterization. In this method, observations closer to the point of estimate are weighted higher using kernel function with a prescribed bandwidth. The kernel bandwidth is local and it extends only to the K^{th} Nearest Neighbor, KNN, observation. An optimal value for KNN is selected by cross validation. Unlike Kriging, the underlying stochastic process is not assumed to be stationary. An application of this model using rainfall data is presented.**

KEYWORDS: **nonparametric statistics, K^{th} nearest neighbor, kernel estimator, nonstationarity, cross validation.**

Contents

1. Introduction

This paper presents a nonparametric approach, based on Kernel estimator, for spatial interpolation of rainfall data. Kernel estimators are popular for probability density estimation and regression where prior assumptions of the functional form of the underlying

behavior are not desirable. The estimates are local approximations of the target function that use only information close to the point of estimate. The data are thus allowed to have a larger role in the estimation process than in a parametric model where a particular functional form is assumed *a priori* for the entire data set, and its parameters are estimated from the data.

Kernel estimators are weighted moving averages of a target function, where the weight is prescribed through a kernel function and the moving average is taken over an appropriately determined span or bandwidth. The kernel function is usually chosen to be a symmetric probability density function with finite variance, which has the role of a weight function. The reader is referred to Hardle (1989), Silverman (1986), and Scott (1992) for accessible monographs on kernel estimation and to Lall (1995) for a review of hydrologic applications.

Kernel estimators with fixed bandwidth are usually used. If there is a high degree of irregularity in the geometry of the basin, or of the sampling locations, and/or high anistropy in the rainfall process, a variable bandwidth may be more appropriate than a fixed one. In this study, our focus is to develop an adaptive kernel estimator where the a variable bandwidth is used. A concise formulation of the kernel estimator is provided and application to rainfall data is then presented.

2. Model Formulation

An estimate of rainfall at an unsampled location may be obtained by considering some neighborhood of that location. This can be achieved by averaging over the locations sampled within this neighborhood. A general formulation of such an estimate can be expressed as:

$$\hat{\rho}(x,y) = \frac{\sum_{i \in \Lambda} \rho_i^* \omega_i(x,y)}{\sum_{i \in \Lambda} \omega_i(x,y)} \qquad (1)$$

where: $\hat{\rho}(x,y)$ - an estimate of rainfall at location (x, y); ρ_i^* - rainfall data at location i; $\omega_i(x,y)$ - a prescribed weight function of data point i in the estimate at location x, y; and Λ - a prescribed neighborhood to location (x, y). The above equation represents a general expression of a two dimensional kernel regression estimator with two main parameters: ω_i(x, y), expressed through kernel function, and Λ , expressed through kernel bandwidths. The choice of the kernel function is first discussed and the method for selecting the bandwidth is then presented.

2.1 Choice of Kernel

The selection of a kernel function is not as critical as the selection of its bandwidth. Scott (1992) shows that the shape of the kernel does not significantly affect the mean square error

(MSE) of estimate. Different kernels can be made equivalent in terms of the MSE by appropriately varying the bandwidth. The kernel function is usually symmetric to yield unbiased estimates by using a symmetric distribution of the weights on both sides of the point of estimate. It is also positive everywhere. Scott (1992) shows that the optimal positive kernel, which minimizes the MSE, is the Epanechinkov kernel, $\{\omega(t) = .75 * (1-t^2);\ |t| \leq 1\}$, whose MSE efficiency is said to be 1. Although different kernels can have nearly the same efficiency by adjusting their bandwidths, the Epanechinkov kernels is selected for this study. A two dimensional kernel as a product of two Epanechinkov kernel can be expressed as:

$$\omega_i(x,y) = \frac{9}{16} * (1 - t_{i,x}^2) * (1 - t_{i,y}^2) \tag{2}$$

where: $t_{i,x} = \left| \frac{x - x_i}{h_x} \right| \leq 1 \quad t_{i,y} = \left| \frac{y - y_i}{h_y} \right| \leq 1$;

h_x and Λ are bandwidths in x and y directions, respectively; and x_i and y_i are Easting and Northing coordinates of point i .

The above kernel has two different global bandwidths h_x, and h_y. Note that such a kernel may be needed if there is a high degree of global asymmetry in the geometry of the basin, the sampling locations, and/or the variation in the rainfall across the area. Another way to deal with such irregularities is to use an adaptive kernel estimator with a local bandwidth that varies to accommodate local irregularities. This kernel can be expressed in radial coordinate as:

$$\omega_i(x,y) = \frac{3}{4} * (1 - t_{i,r}^2) \tag{3}$$

where:

$t_{i,r} = \left| \frac{dr_i}{h_{r,i}} \right| \leq 1$,

$dr_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$, and

$h_{r,i}$ = radial bandwidth at point *i*.

Given *n* observations, the regression function at location *x*, *y* using a radial Kernel estimator is then formed as:

$$\hat{\rho}(x,y) = \frac{\sum_{i=1}^{n} \rho_i * (1 - t_{i,r}^2)}{\sum_{i=1}^{n} (1 - t_{i,r}^2)}, \quad (4)$$

where

$$t_{i,r} = \left| \frac{dr_i}{h_{r,i}} \right| \leq 1$$

2.2 Bandwidth selection

Equation 4 requires a proper selection for $h_{r,i}$. For a situation where ρ" (the second derivative of the rainfall process with respect to x and y) is zero, (i.e., no variation in ρ), the best estimate is obtained by taking the whole domain as the averaging neighborhood. On the other hand if ρ" is high, (i.e., a high variation in ρ), smaller neighborhoods are desirable. A very small neighborhood increases the estimation variability and a very large neighborhood increases the estimation bias. Therefore, the bandwidth selection represents a trade-off between estimation bias and estimation variance.

A global bandwidth, h_r, is usually selected using the method of cross validation. This method judges the adequacy of fit of the model under consideration. An appropriate cross validation method for kernel regression estimation is the Least Square Cross Validation (LSCV) (Hardle, 1989). At each observation point i, we estimate the rainfall data, $\hat{\rho}_{-i}$, using all the available observations except the observation at point i. A cross validated sum of squares of errors is then formed between the $\hat{\rho}_{-i}$ and ρ_i as shown in Equation 5.

$$LSCV = \sum_{i=1}^{n} \left(\frac{\sum_{j=1 \neq i}^{n} \rho_j * (1 - t_{j,r}^2)}{\sum_{j=1 \neq i}^{n} (1 - t_{j,r}^2)} - \rho_i \right)^2 \quad (5)$$

The term $t_{j,r}$ in Equation 5 is a function of the global and local radial bandwidth. A global optimal bandwidth is the one that corresponds to the lowest LSCV score. A global bandwidth will cover a variable number of observations depending on the location of the estimate. A variable (local) bandwidth may be chosen to cover a certain number of observations. Given K nearby observations, the local bandwidth at point i extend to the K^{th} Nearest Neighbor, KNN, of point i. The optimal KNN is then the one that minimizes the LSCV score in the above equation. The resulting estimator is called the adaptive kernel where its bandwidth is dependent on: 1) the process spatial variation; and 2) the observations density around the point of estimate; it is independent of global irregularity of the site geometry, sampling locations, and/or anistropy.

In this study, a direct application of Equation 5 did not produce an optimal solution for the value of KNN. Therefore, a modified LSCV algorithm has been adopted. In this algorithm, the data are split into two equal data sets where we use one set to predict the other. There are ${}^{n}C_{n/2}$ possible scenarios of splitting a data set of size n into two equal data sets. To avoid intensive computation, a prescribed number, N, of scenarios is randomly selected. For each scenario the LSCV is computed as a function of KNN. The resulting LSCV scores are averaged over the N scenarios. The minimum average score is then evaluated.

2.3 Estimation near the site boundary

A problem with forming a kernel estimate is encountered near the boundaries of the basin. Near the boundary, the averaging intervals extend across the boundary where there are no observations. Thus, we do not have a symmetric weighted moving average, leading to the effective center of the average not being at the point of estimate. Several solutions to this problem are offered in the literature (Muller 1991; Silverman 1986). It is noteworthy that this problem is associated with any spatial interpolator. To overcome this problem, we shall limit the spatial domain for estimation to be within the observed data.

3. Application

The model presented above has been applied and evaluated using 467 observations of rainfall data irregularly scattered over space. These data are divided into two data sets. The first data set is used for estimation and it consists of 100 points. The second data set is used for evaluation and it consists of 367 points. Layouts of the two data sets and the estimation region are presented in Figures 1a, and b. Our interest is to apply the kernel estimator, Equation 1 to 4) spatially characterize the rainfall over the site, and 2) evaluate the model performance. In this section, we start with parameter specifications followed by results presentation.

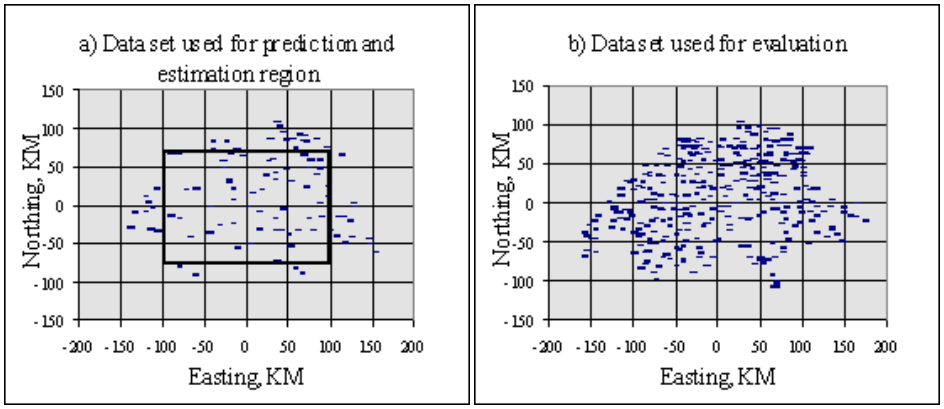

Figure 1 Layouts for the two data sets and the estimation region used in this study.

3.1 Parameter Specifications

As mentioned earlier, Epanechenkov kernel is used in Equation 4. In the modified LSCV algorithm, 10000 data splitting scenarios were randomly selected. The average LSCV scores as a function of KNN is presented in Figure 2.

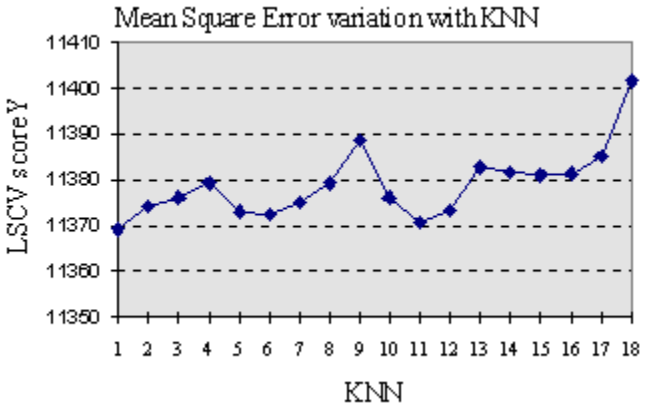

Figure 2 KNN variation with the average LSCV based on 10000 data splitting scenarios.

Figure 2 shows 3 comparable local minima at KNN=1, 6, and 11. The selection of KNN=1 will give a very noisy estimate (high variance), while KNN=11 may provide an overly smoothed estimate. Therefore, a KNN value of 6 seems a good choice. Using Epanechenkov Kernel and KNN=6, the results for this application are presented below.

3.2 Spatial Characterization of Rainfall Distribution

Spatial rainfall characterization was performed using the first data set over a domain restricted within the observed data (see Figure 1a). The resulting image (Figure 3) reflects isolines extending in the North-East direction with principal variation in the North-West direction. In the middle of the image, we notice a band of low rainfall ($\leq$10 mm) extending in the North-East with width ranging from 25 to 50 KM. Two zones of high rainfall ($\geq$40 mm) are observed South-East and North-West to this band respectively. Estimated values range from 6.6 to 41.2 mm.

3.3 Model Performance

To test the model predictability, spatial estimation was performed using the first data set over the point locations of the second data set. Figure 4a shows a scatter plot of observed data versus estimated values. The figure reflects a high estimation variance and, hence, a relatively low correlation (R = .82). The equation of the fitted straight line indicates a relatively small estimation bias. However, we notice, from Figure 4b, that the model tends to underestimate observations of high values ($\geq$40 mm) and to overestimate observations

of low values ($\leq$20 mm) significantly. Most of the extreme values of the estimated data do not correspond to the extreme values of the observed data. The model does not reasonably reproduce the 10 extreme points. The average overestimation for the lowest 5 observations is 15.6 mm, and the average underestimation of the highest 5 observations is 17.5 mm.

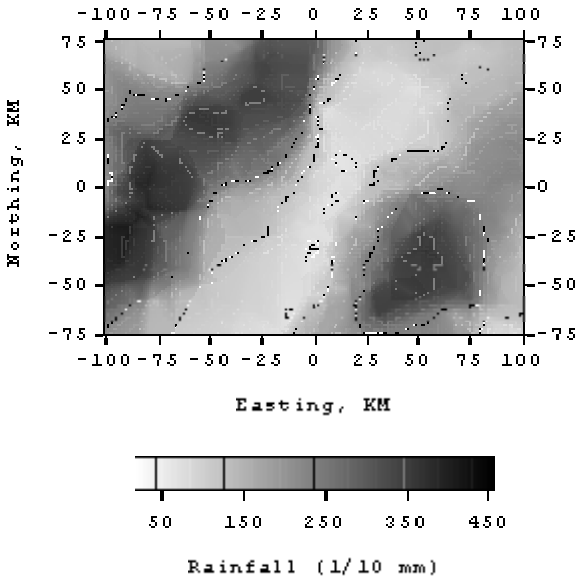

Figure 3 Contour image of the estimated rainfall based on 100 data points.

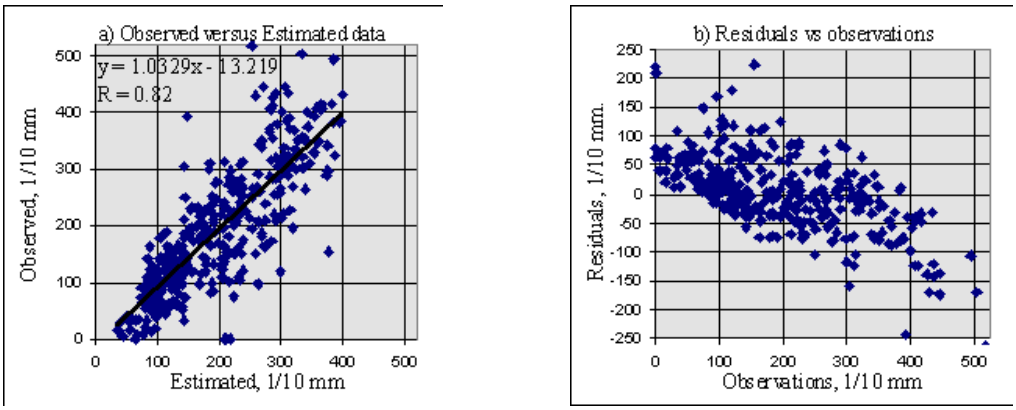

Figure 4 Two comparisons: a) between observed and simulated points; and b) between residuals and observed points for the second data set.

Table 1 shows some key statistics for this application. The results show that except for the extreme value statistics, the model performance is reasonable. On one hand, the extreme values, (maximum and minimum) of the data are not well reproduced. On the other hand the mean and the median are reasonably reproduced. The standard deviation of the simulated points is smaller than that of the observed ones due to, mainly, the over smoothing.

Table 1 Observed and simulated key statistics for the second data set (367 points)

Statistics	Minimum	Maximum	Mean	Median	Standard Deviation
Observed	0	517	185	162	111
Simulated	35.9	400	192	179	88

3.4 Estimation errors

Figure 5 shows the spatial distribution of the estimation error. It is noticed that the errors are mainly negative (light color) in the areas of high rainfall shown in Figure 3, and mainly positive in the areas of low rainfall. This results in correlated errors as noticed in the experimental variogram in Figure 6. The average width of zones with similar errors and the variogram range show that these errors are correlated up to a distance of 20 km.

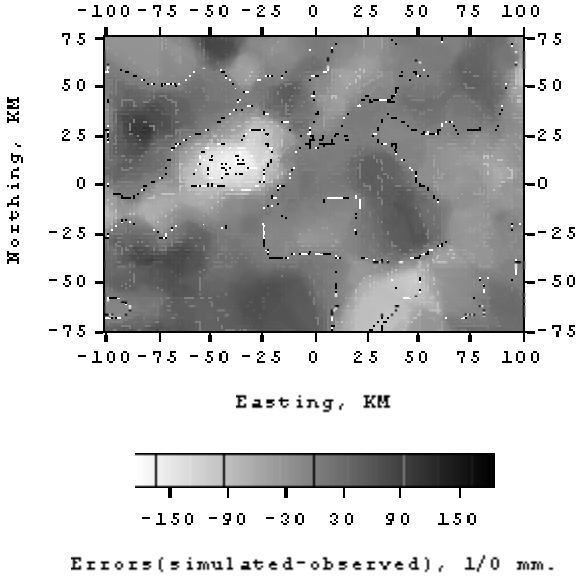

Figure 5 Spatial distribution of the estimation error.

Table 2 shows some error measures for the results presented above. While the Maximum Negative and Positive Errors are high, the Root Mean Squared Error, the Bias, Relative Error, and Mean Absolute Errors are relatively low.

Table 2. Observed and simulated error measures.

Note: RMSE=Root mean squared errors; RE = Relative Error; MAE = Mean Absolute Error; MNE = Maximum Negative Error; MPE = Maximum Positive Error

Statistics	RMSE	BIAS	RE	MAE	MNE	MPE
Errors	64	+7.0	5.8	47.0	-264	223

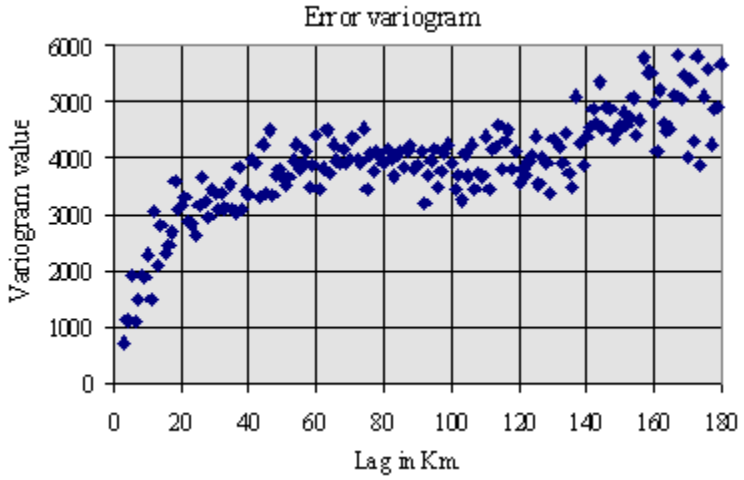

Figure 6 Experimental variogram of the estimation errors.

The same scheme was applied to the second data set to test the model consistency in prediction when a different data set from the same population and with different sizes is used. Figure 7 shows a similar trend to that observed in Figure 3. Increasing the size of the data provides a more resolved estimate with, roughly, the same Mean Square Error.

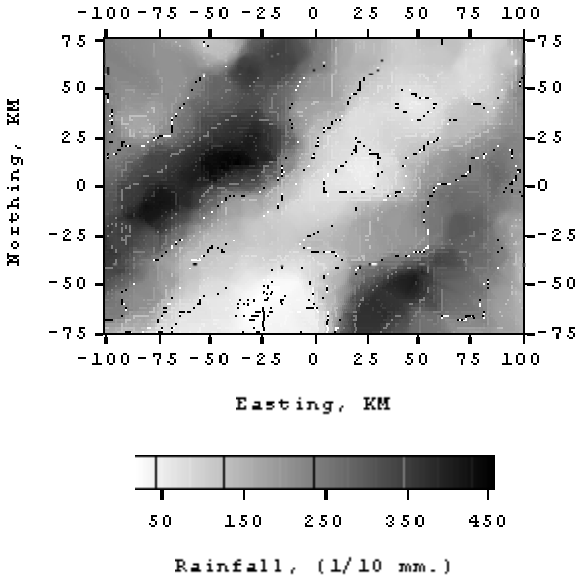

Figure 7 Contour image of the estimated rainfall based on 367 data points.

4. Conclusions

A new method for rainfall characterization using an adaptive kernel estimator was presented. The kernel function and a variable averaging span bandwidth represent the main parameters of this estimator. A proper selection for a variable bandwidth is crucial. The

variable bandwidth is selected based on the K^{th} Nearest Neighbor value that minimizes the LSCV score. A new method for the LSCV is adopted where half of the data set is used to predict the other. While this method is objective in selecting an optimal KNN, it is not robust and it may not converge. Furthermore, the LSCV score shows abrupt variation with KNN with several local minima while the selection is subjective and may not be clear.

The model presented in this study appears attractive in dealing with situations of non-stationary environments, irregular data, and site geometry. While the KNN based bandwidth kernel is flexible and adaptive; it failed to estimate the extreme values due to over smoothing. The resulting statistics show a low estimation bias and a high variance. Also, the errors are correlated up to a distance of 20 km. Such a correlation is a result of heavy over smoothing in zones of extreme values.

In general, this method is nearly fully automatic and can be utilized in connection with other decision support systems. It tends to reproduce the average behavior of a process but fails to reproduce the extreme values. A stochastic simulation scheme based on this methodology may be an attractive tool in reproducing the extreme events. More research efforts are pending to make a better selection for the KNN/bandwidth and to develop a simulation scheme.

References

Hardle, W. (1989) *Applied Nonparametric Regression*. Cambridge, Mass.: Cambridge University Press.

Lall, U. (1995) Nonparametric function estimation: Recent hydrologic applications, *Reviews of Geophysics*, US National Report 1991-1994, pp. 1093-1102.

Muller, H. G. (1991) Smooth optimum kernel estimators near endpoints, *Biometrika* 78 (3); 521-530.

Scott, D. W. (1992) *Multivariate Density Estimation, Theory, Practice, and Visualization*. New York: John Wiley and Sons.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

JGIDA vol. 2, no. 2

JGIDA Home