Journal of Agrometeorology 8(2) : 254 - 259 (Dec. 2006)

Nonparametric regression methodology for modelling and forecasting rainfall

K. P. CHANDRAN* and PRAJNESHU

Indian Agricultural Statistics Research Institute, New Delhi -110 012

ABSTRACT

It is well recognized that Indian agriculture is critically dependent on monsoon rainfall. Thus its accurate forecast, particularly at "Smaller" levels, say Meteorological subdivisions level, is extremely important for efficient micro-level planning. Although, a variety of statistical techniques have been employed in the past to achieve the task, newly developing nonparametric regression approach does not seem to have been used so far. Accordingly, in this paper, this methodology is thoroughly discussed. A heartening feature of this technique is that there are only a few underlying assumptions and so this approach is flexible and robust. As an illustration, modelling and forecasting of Kerala meteorological subdivision's rainfall data is carried out. Comparison with traditional approaches, like ARIMA time-series approach and Polynomial modelling, shows superiority of proposed methodology for data set under consideration.

Key words: Polynomial models, ARIMA approach, nonparametric regression, local linear regression smoother, one-step ahead forecast.

In our country, monsoon rainfall plays a very important role in agriculture. Its accurate and timely forecast of distribution and quantum is vital in planning and policy-making. Accordingly, several attempts have been made in the past to develop forecast models for rainfall. These, generally, can be classified into two approaches discussed below. First approach considers rainfall as an "explicit" function of various external factors. However, despite tall claims, none of these models could forecast the severe drought of year 2002 in India. One plausible reason may be that assumption of a specific functional form between rainfall and predictors is always questionable. Further, these models having eight or ten explanatory variables, like Eurasian snow cover, North-West Europe temperature, and East Asia pressure are certainly not appropriate for making forecasts for smaller regions, say at meteorological subdivisions level. In recent times, some attempts have also been made (Guhathakurta *et al*, 1999) to apply

Neural network technique for forecasting rainfall. This is basically a data-

*Present Address : Central Potato Research Institute, Shimla 171 001 driven approach and requires considerable amount of data. Although there is potential in this methodology, a lot of effort still needs to be done before it could be implemented in practice.

In the second approach, temporal variations in rainfall are studied by applying either Polynomial models or Autoregressive integrated moving average (ARIMA) time-series model (Box *et al.*, 1994). Lı the former, trend in response variable is expressed as a "polynomial function" of time whereas in the latter, response variable at any time t is assumed to be expressible as a "linear" function of its values at past epochs t - 1, t - 2, ... Thus, here role of various predictor variables enter into the model "implicitly" through response variable observations at past epochs. One advantage of this type of approach is that data requirements are much less as compared to earlier approach. In this paper, however, we shall confine our attention only to "Implicit modelling" approach.

ARIMA methodology was applied by Rangaswamy (1999) for forecasting rainfall over Coimbatore by considering monthly and annual rainfall data. Borah and Bora (1995) developed Seasonal ARIMA model to predict monthly rainfall around Guwahati. However, one drawback of ARIMA approach is that resultant model is "linear", which may not hold in reality. Recently, Delsole and Shukla (2002) have also studied linear prediction of Indian monsoon rainfall.

Another drawback of "Parametric time-series models" discussed above is that these may not be robust in the sense that slight contamination of data might lead to erroneous conclusions. Further, a time-series may be of the type that there is no suitable "Parametric" model that gives a good fit. Under these circumstances, a newly developing promising approach, viz. Nonparametric regression technique, which is based on fewer assumptions, may be employed to model the system. This methodology is valid for analyzing rainfall data both at country's level as well as for smaller regions. In this paper, proposed methodology is thoroughly discussed and applied, as an illustration, to annual rainfall data of meteorological subdivision of Kerala for the period 1951 to 2003. The data is taken from the website of IITM, Pune (http://www.tropmet.res.in/).

METHODOLOGY

Rainfall (y_t) at time t can be expressed as the sum of response variable m(t) and error term (ε_t):

$$y_t = m(t) + \varepsilon_t \qquad (1)$$

Models discussed so far assume that the form of m(.) is known except for some unknown parameters and shape of the function is extremely dependent on these parameters. Nonparametric regression approach, which does not require strong assumptions about shape of curve, is very useful in situations where functional relationship is not known. Only assumption made here is that m(.) is a continuous function. As discussed in Efromovich

(1999). Smoothing techniques are usually employed to estimate regression function nonparametrically.

Local linear regression smoothers are generally used in order to obtain a smooth fit of regression function. Here, only qualitative information about m(.) is required and the data speaks for itself to find out the actual form of m(.). Kernel weighted local linear smoother (Fan, 1992) is the popular method used in nonparametric estimation. In this method, following local least square function is minimized and estimators of α_0 and α_1 are obtained:

$$\sum_{i=1}^{n} \left[y_i - \alpha_0 - \alpha_1 \left((t - t_i)/h \right) \right]^2 K_h((t - t_i)/h) \qquad (2)$$

Here $K_h(.)$ is a kernel density function and h is bandwidth which decides the degree of smoothing. Most commonly used kernel is Epanechnikov kernel. Thus, estimator of regression function m(t) is given by

$$\hat{m}(t) = \hat{\alpha}_0 = \sum_{j=1}^{n} W_{t_j} \, y_j \bigg/ \sum_{j=1}^{n} W_{t_j} \qquad (3)$$

where

$$W_{t_j} = K[(t - t_j)/h] \{ \sum_{i=1}^{n} K[(t - t_i)/h] \, (t - t_i)^2 - (t - t_j) \sum_{i=1}^{n} K[(t - t_i)/h] \, (t - t_i) \}$$

Choice of an optimum bandwidth is of great importance in nonparametric regression. A large bandwidth will produce oversmoothed curve, while a small value of it produces an undersmoothed curve. Cross validation or leave-one-out method is most commonly used technique for obtaining optimum value of smoothing parameter (h). This is based on regression smoothers, in which jth observation is left out. Thus, resultant modified estimator is

$$\hat{m}_{h,j}(t_j) = n^{-1} \sum_{i \neq j}^{n} W_{h,j}(t_i) \, y_i \qquad (4)$$

and cross validation function, CV (h), is given by

$$CV(h) = n^{-1} \sum_{j=1}^{n} \left[y_j - \hat{m}_{h,j}(t_j) \right]^2 \qquad (5)$$

The optimum value of smoothing parameter (h) is obtained by minimizing CV (h).

In nonparametric regression, depending on the bandwidth used, MSE and standard error of estimator vary. Standard error (SE) of estimator $m(t_k)$ is computed as the square root of kth diagonal element of the matrix

$$W W' \, [n(MSE)/(n - tr(W))] \qquad (6)$$

where W is the smoothing matrix and tr (W) is trace of matrix W.

RESULTS AND DISCUSSION

As already indicated, annual rainfall data of the meteorological subdivision of

Kerala during the period 1951 to 2003, is considered for data analysis. For modelling purposes, only data up to year 2002 is taken while data for 2003 is used for judging forecasting performance of the model for hold-out data.

Linear and quadratic models are fitted to the data by "Method of least squares" and following results are obtained:

$$\hat{y}_t = \underset{(110.06)}{1821.01} + \underset{(3.55)}{0.263}$$

$$\hat{y}_t = \underset{(169.97)}{1731.69} + \underset{(14.52)}{10.01\,t} - \underset{(0.26)}{0.18\,t^2}$$

Figures within brackets ( ) indicate corresponding standard errors. It may be noted that, for some estimates, standard errors are very high. So strictly speaking, polynomial models considered are not appropriate for present data set. However, if we ignore aspect of high standard errors, Mean square errors (MSE) are, respectively, computed as 153001.7 and 151550.3. Further, one-step ahead forecast for year 2003 for linear and quadratic models are found as 1834.9 mm and 1756.3 mm against the actual value of 1461.0 mm, with corresponding standard errors of 394.9 mm and 397.0 mm respectively. Evidently, incorporation of quadratic term has not made any significant improvement and so, in first instance, linear model may be selected.

For present data, analysis is carried out using statistical analysis system, ver. 8e software package. Examination of auto correlation function (ACF) and partial auto correlation function (PACF) reveals stationarity of data series. Model identification in the present case involves finding appropriate values for p and q. With the help of ACF, PACF and Akaike information criterion (AIC), model identified for data under study is AR(1) with following estimates:

$$\text{Constant} = \underset{(67.79)}{1822.60} \qquad \text{AR}(1) = \underset{(0.01)}{0.22}$$

The MSE for this model is computed as 145495. One-step ahead forecast based on the above model is calculated as 1701.5 mm with a standard error of 385.1 mm against the actual value of 1461.0 mm. Thus, ARIMA time-series method is found to be superior to polynomial models for data set under consideration.

As no software package is available for applying the above methodology, computer programs are developed in MATLAB, Ver. 5.3.1 software package which is appended as annexure - I. In the first step, time-interval pertaining to the data under consideration is transformed into the interval [0, 1]. For the present data, optimum bandwidth is computed as 0.096, using cross validation method. This is used for further nonparametric estimation of rainfall at different time epochs. MSE is computed as 88887.2, which is found to be considerably lower than that for ARIMA time-series model. Further, one-step ahead forecast for the year 2003 is obtained as 1388.0 mm with a standard error of 115.8 mm. Thus, nonparametric regression methodology, applied to annual rainfall data of Kerala subdivision, provides a forecast closer to the actual value and with a lower MSE as compared to ARIMA time-series model. Therefore, proposed approach is found to be better than polynomial as well as ARIMA time-series models for modelling as well as forecasting of rainfall data under consideration. Evidently, nonparametric regression model is able to describe cyclical variations in rainfall data of Kerala. Finally, using this methodology for data from 1951 to 2003, rainfall for the year 2004 is forecast as 1317.6 mm.

To sum up, in present communication, utility of employing Nonparametric regression approach is highlighted for modelling and forecasting of rainfall data at meteorological subdivisions level.

CONCLUSIONS

Nonparametric regression methodology developed and discussed here revealed that this technique has a few underlying assumptions and so this approach is flexible and robust. As an successful illustration, modelling and forecasting of annual rainfall data for Kerala meteorological subdivision suggests that this methodology can be applied to other subdivisions of India.

Annexure- I

Program for estimation of nonparametric regression function and standard error

(i) Estimation of nonparametric regression function:

```
mse=0.0;     n= 26;
for i=1:n
h=0.12;      a0=0; a1=0; a2=0;
for j=1:n
 u1=(x(j)-x(i));    u=u1/h;
kr(j)=0.75*(1-u*u);
if abs(u)>1    kr(j)=0; end;
a0=a0+kr(j); a1=a1+u1*kr(j);
a2=a2+u1*u1*kr(j);
end;
for k=1:n
w(i,k)=kr(k)*(a2-(x(k)-x(i))*a1)/(a0*a2-
a1*a1);
end;
end;
tr=trace(w);     m=w*y';      e=y-m';
for i=1:n
 mse= mse+e(i)*e(i)/n;
fprintf(fpo,'%5d
%10.4f%10.4f\n',i+1975,y(i),m(i));
end
```

(ii) One-step ahead forecast:

```
 a1(n+1)=0.0;a2(n+1)=0.0;a3(n+1)=0.0;ak1=0.0;ak2=0.0;
j=n;
if j<1
r1=-1;
else r1=(x(j)-x(n))/h;    end;
while r1>-1,
k1=(1-r1*r1)*0.75; a1(n+1)=a1(n+1)+k1;
a2(n+1)=a2(n+1)+k1*r1*h;
a3(n+1)=a3(n+1)+k1*r1*r1*h*h;
ak1=ak1+k1*y(j); ak2=ak2+k1*y(j)*r1*h;
j=j-1;
if j<1
r1=-1;
else r1=(x(j)-x(n))/h;    end;      end;
```

```
dm1(n+1)=a1(n+1)*a3(n+1)-
a2(n+1)*a2(n+1);
m(n+1)=ak1*a3(n+1)-a2(n+1)*ak2;
m(n+1)=m(n+1)/dm1(n+1)
fprintf(fpo,'\n%5d %10.4f
\n',n+1976,m(n+1));
end;
```

(iii) Estimation of standard error:

```
v=w*w'*mse*n/(n-tr);
for i=1:n
s1(i)=v(i,i);
s(i)=s1(i)^.5;
fprintf(fpo,'SE=%8.2f\n',s(i));
end;
st=fclose(fpo);
```

REFERENCES

Borah, D. K. and P. K. Bora 1995. Predicting the monthly rainfall around Guwahati using a Seasonal ARIMA model. *J. Ind. Soc. Ag. Stat.*, 47: 278-287.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel 1994. Time-series analysis: Forecasting and control. 3rd edn., Prentice Hall, U. S. A.

Delsole, T. and J. Shukla 2002. Linear prediction of Indian monsoon rainfall. *J. Climate*, 15: 3645-3658.

Efromovich, S. 1999. Nonparametric curve estimation. Springer-Verlag, U. S. A.

Fan, J. 1992. Design adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, 87: 998-1004.

Guhathakurta, P., M. Rajeevan, and V. Thapliyal. 1999. Long range forecasting Indian summer monsoon rainfall by a principal component neural network model. *Met. Atmos. Phys.*, 71: 255-266.

Rangaswamy, R. 1999. An ARMA model for long range forecasting of rainfall over Coimbatore. *Mausam*, 50: 299-310.