*J. Ind. Soc. Agril. Statist.*
*59(2), 2005 : 126–130*

# Nonparametric Regression with Jump Points Methodology for Describing Country's Oilseed Yield Data

K.P. Chandran and Prajneshu[1]
*Central Potato Research Institute, Shimla*
(Received : February, 2005)

**SUMMARY**

Oilseed production of our country has shown a drastic increase since mid-1980s, primarily due to the setting up of "Technology Mission on Oilseeds". One important issue to be resolved is whether or not increase in yield of oilseed crops is mainly responsible for this transformation. Present paper attempts to examine this statistically using oilseed yield data from years 1950 to 2000. In first step, ARIMA time-series approach is adopted to model the data; however it could not explain the sudden jumps. So nonparametric regression approach, which requires fewer assumptions, is employed. To this end, computer programs are developed in Matlab, Ver. 5.3.1 to carry out estimation of location of jump and computation of critical values of jump size. It is shown that nonparametric regression with jump points provides a good description of data under consideration and gives statistical evidence of jump in productivity of oilseeds. Finally, one-step ahead forecast is carried out with the fitted nonparametric regression model.

*Key words :* Oilseed yield data, ARIMA approach, Nonparametric regression, Jump point, Local polynomial regression.

## 1. INTRODUCTION

India is among largest oil economies in the world occupying a distinct position in terms of diversity in annual oilseed crops. Main oilseed crops are: groundnut, rapeseed, mustard, soybean, sunflower, safflower, sesame, niger, linseed, and castor. Annual production of oilseed crops was virtually stagnating at around 10 million tonnes over a span of more than fifteen years despite considerable increase in area under oilseed crops from 10.73 million hectares in 1950-51 to 19.01 million hectares in 1985-86. Till mid-1980s, supply lagged far behind demand, thus forcing the Government to import large quantities of edible oils. Turning point came in 1986 with setting up of "Technology Mission on Oilseeds". Soon, India attained a status of "Self sufficient and net exporter" during early nineties with an all-India record production of 25 million tonnes during 1996-97. This transformation is rightly termed as the "Yellow Revolution". An excellent account of various aspects of this success story is given in Rai (1999).

In this paper, country's oilseed yield data during the years 1967-68 to 2003-04, taken from DES (2004), is considered for data analysis. Generally, for analyzing this type of time-series data collected over time, Auto Regressive Integrated Moving Average (ARIMA) methodology (Box *et al.* (1994)) is employed. One disadvantage of this methodology is that time-series under consideration should be stationary or should be capable of becoming so by means of differencing or detrending. Accordingly, Structural time-series approach is adopted by Prajneshu *et al.* (2002) for modelling and forecasting country's lac production data. Mukhopadhyay and Sarkar (2001) have carried out detailed trend analysis of agricultural production data of West Bengal during the period 1950-51 to 1992-93 and found no statistical support for acceleration during 1980s, as proposed by some previous studies.

One drawback of above type of research work is that it is based on the assumption of linearity, which does not hold in reality. Another drawback is that models may

*1. Indian Agricultural Statistics Research Institute, New Delhi-110012*

not be robust in the sense that slight contamination of data might lead to erroneous conclusions. Further, a time-series might be of the type that there is no suitable parametric model that gives a good fit (Prakasa Rao (1996)). Under these circumstances, one might take recourse to nonparametric regression approach, which is based on fewer assumptions.

Purpose of present paper is to model oilseed yield data of our country nonparametrically. Details of this approach, when there is presence of jump points in data, are thoroughly discussed. Relevant computer programs for analyzing data, developed in Matlab, Ver. 5.3.1, are appended as Annexure-I. Finally, the methodology is applied to data and its capability to identify jump point is demonstrated.

## 2. ARIMA TIME-SERIES APPROACH

In this section, oilseed yield data is modelled using ARIMA approach, as given in Box *et al.* (1994). These models predict response as a linear combination of its own past values. In first instance, it is noticed that the original data is not stationary but can be made so after first differencing. Subsequently, using SAS, Ver. 8e software package, appropriate model identified to describe the present data is found to be ARIMA (0, 1, 1) with following estimates

| Constant = 11.44, | MA1 = 0.81, | SE = 78.14 |
|---|---|---|
| (2.76) | (0.10) | |

Here figures in ( ) indicate the standard errors of corresponding coefficients. Further, goodness of fit statistics are

AIC = 417.92, SBC = 421.09, MSE = 6106.38

High values of above goodness of fit statistics indicate that ARIMA model is not appropriate to explain sudden jumps present in the data.

## 3. NONPARAMETRIC REGRESSION METHODOLOGY

### 3.1 Basics

A fundamental problem in statistics is to develop models based on sample of observations and making inference using the model so developed. Regression analysis provides information on relationship between response variable and predictor variable as

$$y = m(x) + \varepsilon \qquad (1)$$

A parametric (linear or nonlinear) model assumes that the form of m(.) is known except for some unknown parameters and shape of function is entirely dependent on parameters. Often, it is difficult to guess most appropriate functional form just from the shape of curve and in such situations, nonparametric regression approach, which does not require strong assumptions about shape of curve, is very useful. Only assumption made here is that m(.) belongs to some infinite dimensional collection of functions. Smoothing techniques are usually employed to estimate regression function nonparametrically (Hardle (1990)).

Local linear regression smoothers are generally used in order to obtain a smooth fit of regression function, when no suitable parametric model is available. Kernel Weighted Local Linear Smoother (KWLLS), proposed by Fan (1992), is the popular method used in nonparametric estimation. In this method, estimator of m(.) is given by value of $\alpha_0$ where $\alpha_0$ (and $\alpha_1$) minimizes local least square function

$$\sum_{i=1}^{n} \left[ y_i - \alpha_0 - \alpha_1 (x - x_i) \right]^2 K(x - x_i) \qquad (2)$$

Here, K(x) is a kernel density function. Most commonly used kernel is Epanechnikov kernel given as

$$K(x) = \begin{cases} 0.75\,(1 - x^2) & \text{for } |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, estimator of regression function m(x) is given by

$$\hat{m}(x) = \hat{\alpha}_0 = \sum_{j=1}^{n} W_{x_j} y_j / \sum_{j=1}^{n} W_{x_j} \qquad (3)$$

where

$$W_{x_j} = K[(x - x_j)/h] \{ \sum_{i=1}^{n} K[(x - x_i)/h] (x - x_i)^2 - (x - x_j) \sum_{i=1}^{n} K[(x - x_i)/h] (x - x_i) \}$$

### 3.2 Choice of Smoothing Parameter (Bandwidth)

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 $j^{th}$ observation is left out. Thus, resultant modified estimator is

$$\hat{m}_{h,j}(x_j) = n^{-1} \sum_{i \neq j}^{n} W_{h,j}(x_i)\, y_i$$

where

$$W_{h,j}(x_i) = K[(x_j - x_i)/h]\{\sum_{k \neq j}^{n} K[(x_j - x_k)/h](x_j - x_k)^2 - (x_j - x_i)\sum_{k \neq j}^{n} K[(x_j - x_k)/h](x_j - x_k)\}$$

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

$$CV(h) = n^{-1} \sum_{j=1}^{n} [y_j - \hat{m}_{h,j}(x_j)]^2 \qquad (4)$$

The optimum value of smoothing parameter (h) is obtained by minimizing CV (h). To achieve this task, computer programs are developed in MATLAB, Ver. 5.3.1 (1999), and are given in Annexure-I.

### 3.3 Data Analysis

For the given data, optimum bandwidth (h) is estimated as 0.135. This value of h is used for further estimation of m(.), as given in eq. (3). The MSE value, viz. 3719.76 is found to be significantly lower than that of ARIMA model. However, residuals corresponding to data during late 1980's and early 1990's are seen to be very large, indicating the need for including aspect of jump points in the above methodology.

## 4. NONPARAMETRIC REGRESSION WITH JUMP POINTS METHODOLOGY

In this section, focus is on modelling sudden jumps in the data under study. Mc Donald and Owen (1986) used split linear fit of data to estimate jump point. Muller (1992) provided estimators for location and size of change points in nonparametric regression based on left and right one-sided kernel smoothers. The above methods are suited for equally spaced fixed design case. Jose and Ismail (1999), extending the work of Loader (1996), developed generalized estimators for location and size of jump in regression function or its derivatives, based on analysis of residuals from nonparametric kernel regression method. In this section, this approach is followed and is briefly discussed below

The regression function m(.) with change points at $t_j$ of size $\Delta_{t_j}$, j= 1,2,...,p is defined by

$$m(x) = g(x) + \sum_{j=1}^{p} \Delta_{t_j} D_{[t_j,1]}(x) \qquad (5)$$

where g(.) is a continuous function defined on [0, 1] and D is indicator function. The regression function m with change points for $m'$, the first derivative of m at $t_j$ of size $\Lambda_{t_j}$, j = 1, 2,..., p is given by

$$m(x) = g(x) + \sum_{j=1}^{p} \Lambda_{t_j}(x - t_j) D_{[t_j,1]}(x) \qquad (6)$$

Let

$$m_+(t_j) = \lim_{t \downarrow t_j} m(t), \quad m_-(t_j) = \lim_{t \uparrow t_j} m(t)$$

Further, let

$$\Delta_{ij} = m_+(t_j) - m_-(t_j), \quad \Lambda_{ij} = m'_+(t_j) - m'_-(t_j)$$

One-sided estimators of $m_+(x_t)$ and $m_-(x_t)$ are

$$\hat{m}_{\pm}(x_t) = \sum_{i=1}^{n} W_{\pm t_j} Y_j / \sum_{i=1}^{n} W_{\pm t_j} \qquad (7)$$

where

$$W_{\pm t_j} = K_{\pm}[(x_t - x_j)/h]\,[S_{\pm 2} - (x_t - x_j)S_{\pm 1}]$$

and

$$S_{\pm j} = \sum_{i=1}^{n} K_{\pm}[(x_t - x_i)/h](x_t - x_i)^j, \; j = 0,1,2,...$$

If there exists a change point for m(.) at $x_t$ of size, $\Delta_t$, it is possible to take estimate of change point as that point where $|\hat{\Delta}_i|$ is maximum over the set $x_i \in [h, 1-h]$.

For the given data, using cross validation method, optimum bandwidth (h) is found to be 0.135 after transforming x variable (time) into [0, 1]. Using eq. (7), jump location and size is estimated using computer program given in Annexure-I. For given data, only one jump is estimated and that takes place during the year 1987-88 having a size of 163 kg / ha. Since the number of change points is not known in advance, critical value of jump size has to be estimated to decide the number of significant change points.

### 4.1 Estimation of Critical Value

In absence of change points, $\hat{\Delta}_t$ follows normal distribution with mean zero and variance $\sigma^2$. Therefore, to compute critical value for testing significance of jump size, an estimate of $\sigma^2$ is needed. To this end, Wu and Chu (1993) constructed an estimator as a trimmed mean of squared difference of neighbouring observations, i.e.

$$\hat{\sigma}^2 = \sum_{i=2+q}^{n-q} \xi_i / [2(n-1-2q)] \qquad (8)$$

where $\xi_i = (y_i - y_{i-1})^2$, i = 1(1)n and $\xi_i$ are arranged in ascending order. Substituting the expression of $\hat{\sigma}^2$ from eq. (8), Jose and Ismail (1999) derived a modified estimator of $\sigma^2$ as

$$\hat{\sigma}_t^2 = \left[ \sum_{j=1}^{n} W_{+tj}^2 / (\sum_{j=1}^{n} W_{+tj})^2 + \sum_{j=1}^{n} W_{-tj}^2 / (\sum_{j=1}^{n} W_{-tj})^2 \right] \hat{\sigma}^2$$

Critical value is obtained as

$$C_\alpha = \hat{\sigma}_t \, Z_{\alpha^*/2}$$

where $\alpha$ is level of significance and $\alpha^* = 1 - (1-\alpha)^{1/n}$. As critical value obtained for jump at 5% level during year 1987-88, viz. 154.98 is less than estimated jump (163 kg/ha), a significant jump at that point is established.

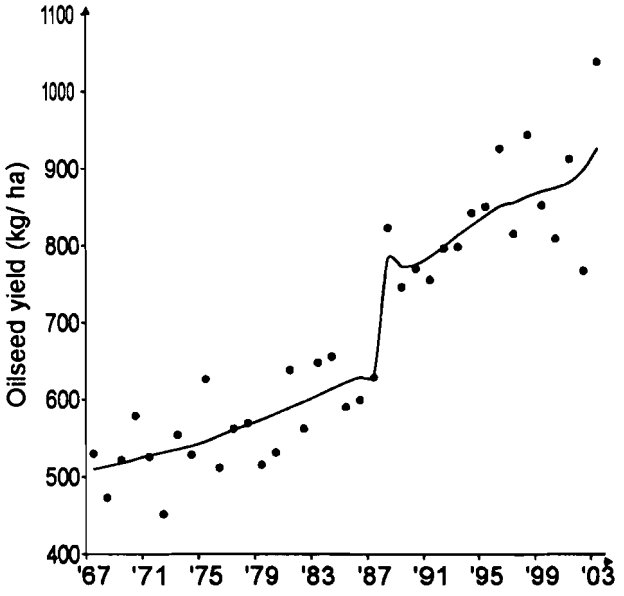


**Fig.1.** Fitted nonparametric regression with jump points model to all India oil seed yield data

Evidently, oilseed yield estimates corresponding to data during late 1980's and early 1990's are closer to actual values, resulting in a reduction in MSE value from 3719.76 to 2387.19. This shows that sudden boom in oilseed productivity during late 1980s has, indeed, contributed to the success of "Yellow Revolution". A graph of fitted model along with data is exhibited in Fig. 1. Finally, one-step ahead forecast for the year 2004-05 is computed as 1053.67 kg/ha.


### ACKNOWLEDGEMENTS

Authors are grateful to the referee for valuable comments.


### ANNEXURE-I

### PROGRAM TO ESTIMATE JUMP POINTS AND JUMP SIZES

```
fpi=fopen('C:\data\oilyield.txt', 'r');
fpo=fopen('c:\data\oilyieldjump.xls', 'w');
n=37;
```

### DATA GENERATION

```
[z]=fscanf(fpi, '%f',[2,37]);
for i=1:n
   x(i)=z(1,i)/z(1,n); y(i)=z(2,i);
end;
```

### CROSS VALIDATION FOR ESTIMATION OF BANDWIDTH

```
for nh = 3:n/2
h = nh/n; mse=0.0;
for i = 1:n
a1(i)=0.0;a2(i)=0.0;a3(i)=0.0;ak1=0.0;ak2=0.0;j=i;
r1=(x(j)-x(i))/h;
while r1<1,
k1=(1-r1*r1)*0.75; if r1=0; k1=0; end;
a1(i)=a1(i)+k1; a2(i)=a2(i)+k1*r1*h;
a3(i)=a3(i)+k1*r1*r1*h*h; ak1=ak1+k1*y(j);
ak2=ak2+k1*y(j)*r1*h;
j=j+1; if j<=n; r1=(x(j)-x(i))/h; else r1=1; end; end;
j=i-1; if j<1; r1=-1; else r1=(x(j)-x(i))/h; end;
while r1>-1,
k1=(1-r1*r1)*0.75; a1(i)=a1(i)+k1;
a2(i)=a2(i)+k1*r1*h; a3(i)=a3(i)+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(i))/h; end; end;
dm1(i)=a1(i)*a3(i)-a2(i)*a2(i);
```

```
    m(i)=ak1*a3(i)-a2(i)*ak2; m(i)=m(i)/dm1(i)
    mse=mse+(y(i)-m(i))*(y(i)-m(i)); end;
mse=mse/n; fprintf(fpo, \n%5d %10d\n',nh,mse); end;
end;
```

## ESTIMATION OF JUMP LOCATIONS AND JUMP SIZES

```
for i=1:h+1
    k(i)=0.75*(1-(i-1)/h*(i-1)/h);
end;
    a(1,1)=0.0; a(1,2)=0.0; a(2,2)=0.0;
for r=1:h
    a(1,1)=a(1,1)+k(r+1); a(1,2)=a(1,2)+x(r)*k(r+1);
    a(2,2)=a(2,2)+x(r)*x(r)*k(r+1);
end;
    a(1,3)=a(1,1); a(1,4)=a(1,2); a(2,3)=a(1,2);
    a(2,4)=a(2,2); a(1,1)=2*a(1,1)+k(1); a(2,1)=0;
    a(3,1)=a(1,3); a(3,2)=a(2,3); a(3,3)=a(1,3);
    a(3,4)=a(1,4); a(4,1)=a(1,4); a(4,2)=a(2,4);
    a(4,3)=a(3,4); a(4,4)=a(4,2); a(2,2)=2*a(2,2);
    a(1,2)=0;
c(1,1)=a(1,1);c(1,2)=a(1,2);c(2,1)=a(2,1);c(2,2)=a(2,2);
ic=inv(c);ia=inv(a);
for i=h+1:n-h,
    b(1)=0.0; b(2)=0.0; b(3)=0.0; b(4)=0.0;ty=0;
    for r=1:h
    ty=ty+k(r)*y(r+i-1)*y(r+i-1)+k(r+1)*y(i-r)*y(i-r);
    b(1)=b(1)+k(r)*y(r+i-1)+k(r+1)*y(i-r);
    b(2)=b(2)+x(r)*k(r+1)*y(i+r)-x(r)*k(r+1)*y(i-r);
    b(3)=b(3)+k(r+1)*y(i+r);
    b(4)=b(4)+x(r)*k(r+1)*y(i+r);
    end;
    d(1)=b(1);    d(2)=b(2);    sol2=ia* b'
    sol1=ic* d';   sr1=b*sol2-d*sol1;
    ey=ty-b*sol2;   sr(i)=sr1*(2*n*h-2)/(ey*2);
    sol(i,1)=sol2(1);    sol(i,2)=sol2(2);
    sol(i,3)=sol2(3);    sol(i,4)=sol2(4);
    end;
    [q,p]=max(sr);
for j=1:4
    sol(p,j);
end;
fprintf(fpo, '\n%8.4f\n',p);
```

## ESTIMATION OF NONPARAMETRIC REGRESSION FUNCTION

```
h=0.135;
for i=1:n
    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=y*w';
st=fclose(fpo);
```

## REFERENCES


Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control.* 3rd edn. Prentice Hall, U.S.A.

DES (2004). *Agricultural Statistics at a Glance.* Directorate of Economics & Statistics, Ministry of Agriculture, India.

Fan, J. (1992). Design adaptive non-parametric regression. *J. Amer. Statist. Assoc.*, **87**, 998-1004.

Hardle, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press, U.S.A.

Jose, C.T. and Ismail, B. (1999). Change points in regression functions. *Comm. Stat. - Theory Methods*, **28**, 1883-1902.

Loader, C.R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.*, **24**, 1667-1678.

Matlab (1999). *Version 5.3.1.* The Math Works, Inc., U.S.A.

Mc Donald, J.A. and Owen, A.B. (1986). Smoothing with split linear fits. *Technometrics*, **28**, 195-208.

Mukhopadhyay, D. and Sarkar, N. (2001). Has there been any acceleration in the growth of agriculture in West Bengal?: A fresh look using modern time series techniques. *Sankhya,* **B63**, 89-107.

Muller, H.G. (1992). Change points in regression analysis. *Ann. Statist.*, **20**, 737-761.

Prajneshu, Ravichandran, S. and Wadhwa, S. (2002). Structural time series models for describing cyclical fluctuations. *J. Ind. Soc. Agril. Statist.*, **55**, 70-78.

Prakasa Rao, B.L.S. (1996). Nonparametric approach to time series analysis. *In : Stochast. Process and Statist. Inf.* Eds. B.L.S. Prakasa Rao and B.R. Bhat, New Age International Ltd., 73-89.

Rai, M. (1999). *Oilseeds in India- A Success Story in Mission Mode.* Asia-Pacific Association of Agricultural Research Institutions, Thailand.

SAS (1999). *SAS User's Guide, Version 8e.* SAS Institute Inc., U.S.A.

Wu, J.S. and Chu, C.K. (1993). Kernel type estimators of jump points and values of a regression function. *Ann. Statist.*, **21**, 1545-1566.