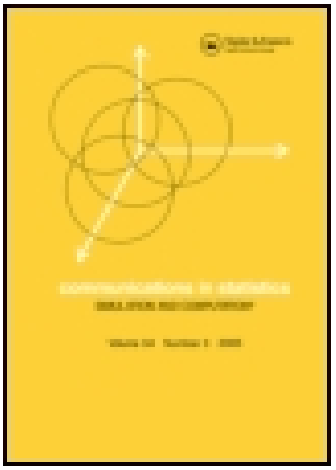

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Trend, Growth Rate, and Change Point Analysis—A Data Driven Approach

C. T. Jose [a] , B. Ismail [b] & S. Jayasekhar [a]

[a] Central Plantation Crops Research Institute , Regional Station , Karnataka, India

[b] Department of Statistics , Mangalore University , Karnataka, India

Published online: 20 Feb 2008.

To cite this article: C. T. Jose , B. Ismail & S. Jayasekhar (2008) Trend, Growth Rate, and Change Point Analysis—A Data Driven Approach, Communications in Statistics - Simulation and Computation, 37:3, 498-506, DOI: 10.1080/03610910701812477

To link to this article: http://dx.doi.org/10.1080/03610910701812477

Communications in Statistics—Simulation and Computation®, 37: 498–506, 2008

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910701812477

Taylor & Francis
Taylor & Francis Group

Inference

Trend, Growth Rate, and Change Point Analysis—A Data Driven Approach

C. T. JOSE[1], B. ISMAIL[2], AND S. JAYASEKHAR[1]

[1]Central Plantation Crops Research Institute, Regional Station, Karnataka, India
[2]Department of Statistics, Mangalore University, Karnataka, India

A data-driven technique is proposed to estimate the trend and relative growth rate of time series data. The method is based on the local linear regression smoother and the only assumption about the form of the trend and growth rate function is that they are smooth functions of time. We also extended the method for handling sudden shifts or changes in the trend or growth rate functions by adding dummy variables for the jumps. Simulation studies are carried out to see the performance of the proposed procedure. The method is applied to study the trend and growth rate of wheat production in India from 1951–2005.

Keywords Change points; Growth analysis; Nonparametric regression; Trend analysis.

Mathematics Subject Classification 62G08.

1. Introduction

In recent years, data-driven or nonparametric techniques have become increasingly popular as tools for data analysis. These techniques impose only few assumptions about the data and therefore, they are more flexible than the usual parametric approaches. In parametric approach, we assume some parametric functional form for the trend or growth rate and estimate the parameters based on the given set of data. Then the results will depend on the data and the assumed functional form. In many situations, we may not know the exact functional form and sometimes there may not be any parametric functional form to represent the data. In such situations, the data-driven approach, which depends entirely on the data will be more suitable. In this article, we have proposed data-driven methods for the estimation of trend, growth rate, and change points of time series data. The method is based on the local linear regression smoother and the only assumption about the form of the trend and

Received July 13, 2006; Accepted August 27, 2007
Address correspondence to C. T. Jose, Central Plantation Crops Research Institute, Regional Station, Vittal, Karnataka 574 243, India; E-mail: ctjos@yahoo.com

growth rate function is that they are smooth functions of time except for the change point. The estimation procedures for trend, growth rate, and change points are given in Sec. 2. Simulation studies to see the performance of the proposed procedures are described in Sec. 3. The trend and growth rate analysis of wheat production in India using the proposed method is provided in Sec. 4. A brief conclusion is given in Sec. 5.

2. Estimation of Trend and Growth Rate

Let us first consider the nonparametric regression model with additive error of the form

$$y_i = m(t_i) + \varepsilon_i, \quad t_i = i/n, \ \ i = 1, \ldots, n, \tag{1}$$

where y_i is the observation of the ith time point, m is the trend function, which is assumed to be smooth, and ε_i are random errors with mean zero and finite variance $\sigma^2 < \infty$. The kernel weighted linear regression smoother (Fan, 1992) is used to estimate the trend function nonparametrically. The value of the local linear regression smoother at time t is the solution of a_0 to the following weighted least squares problem:

$$\sum_{t=1}^{n} [y_t - a_0 - a_1(t_i - t)]^2 K\left(\frac{t_i - t}{h}\right), \tag{2}$$

where K is a bounded symmetric kernel density function and h is the bandwidth. Let $\hat{a}_0$ and $\hat{a}_1$ be the solutions to the weighted least squares problem (2). The estimate of the trend function $m(t)$ is given by

$$\hat{m}(t) = \hat{a}_0 = \sum_{j=1}^{n} W_{tj} y_j, \tag{3}$$

where

$$W_{tj} = \frac{K_j[s_2 - (t - t_j)s_1]}{s_0 s_2 - s_1^2}, \quad K_j = K\left[\frac{t - t_j}{h}\right], \quad \text{and} \quad s_l = \sum_{k=1}^{n} K\left(\frac{t - t_k}{h}\right)(t - t_k)^l.$$

The optimum bandwidth h can be obtained by the method of cross validation. The slope $m'(t)$ of $m(t)$ can be considered as the simple linear growth rate at the time point t. The estimate of $m'(t)$ is given by

$$\hat{m}'(t) = \hat{a}_1 = \sum_{j=1}^{n} W'_{tj} y_j \tag{4}$$

where

$$W'_{tj} = \frac{K_j[(t - t_j)s_0 - s_1]}{s_0 s_2 - s_1^2}$$

The following regularity assumptions are made to prove the properties of the estimate.

A1. The regression function $m(t)$ has a bounded and continuous second derivative.
A2. $f(t)$, the marginal density of the covariate is continuous and bounded away from zero.
A3. The kernel function K is a bounded symmetric density function defined in the interval $[-1, 1]$ with $\int u^{2r}K(u)du < \infty$, for $r = 1, 2, \ldots$

The asymptotic properties of the estimators are proved in Theorem 2.1.

Theorem 2.1. *Let* (t_i, y_i), $t_i \in [0, 1]$, $i = 1, \ldots, n$ *are* n *observations of the regression model*(1) *and assume that the conditions* A1–A3 *hold. Then*:

(a) *under the conditions that* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$

$$E[\hat{m}(t)] \to m(t), \quad \sigma_0^2(t) = V[\hat{m}(t)] \to 0, \quad \text{and} \quad \{\hat{m}(t) - E[\hat{m}(t)]\}/\sigma_0(t) \to N(0, 1);$$

(b) *under the conditions that* $h \to 0$ *and* $nh^3 \to \infty$ *as* $n \to \infty$

$$E[\hat{m}'(t)] \to m'(t), \quad \sigma_1^2(t) = V[\hat{m}'(t)] \to 0, \quad \text{and} \quad \{\hat{m}'(t) - E[\hat{m}'(t)]\}/\sigma_1(t) \to N(0, 1).$$

Proof. Since $\sum_{W_{tj}}(t - t_j) = 0$, we can write

$$E[\hat{m}(t)] - m(t) = \sum W_{tj}E[y_j - m(t) - (t - t_j)m'(t)].$$

If we use Taylor expression for $m(t)$, the right-hand side is of the order $O(h^2)$ and therefore

$$\begin{aligned} E[\hat{m}(t)] &\to m(t) \\ V[\hat{m}(t)] &= \sigma^2 \sum W_{tj}^2 \\ \sum_{j=1}^{n} W_{tj}^2 &= \frac{\sum_{j=1}^{n} K_j^2[s_2^2 - 2(t - t_j)s_2s_1 + (t - t_j)^2 s_1^2]}{[s_0s_2 - s_1^2]^2}. \end{aligned}$$

Using the standard result from kernel density estimation, we have:

$$s_l = nh^{l+1}f(t)\int u^l K(u)du[1 + O(h)], \quad l = 0, 1, 2 \tag{5}$$

and

$$\frac{1}{nh^{l+1}}\sum K^2\left[\frac{t - t_j}{h}\right](t - t_j)^l = f(t)\int u^l K^2(u)du + o(1), \quad l = 0, 1, 2. \tag{6}$$

By using (5) and (6), we can easily see that under the condition $h \to 0$ and $nh \to \infty$ as $n \to \infty$:

$$\begin{aligned} \sum W_{tj}^2 \to 0 \quad \text{and} \quad \sigma^2 \sum W_{tj}^2 \to 0 \\ \hat{m}(t) - E[\hat{m}(t)] = \sum_{j=1}^{n} W_{tj}\varepsilon_j. \end{aligned}$$

To prove the normality of $\hat{m}(t)$, it is sufficient to show that the Lindberg condition

$$\frac{\max|W_{tj}|}{[\sum W_{tj}^2]^{1/2}} \to 0 \quad \text{as } n \to \infty$$

$$\max|W_{tj}| \leq \frac{|s_2| + |s_1|h}{s_0 s_2 - s_1^2}.$$

Using Eqs. (5) and (6), it can be shown that under the condition $h \to 0$ and $nh \to \infty$ as $n \to \infty$

$$\frac{\max|W_{tj}|}{[\sum W_{tj}^2]^{1/2}} \to 0.$$

This completes the proof of part (a) and part (b) can be proved in a similar way.

Under the assumption that the trend function m is smooth and $m(t) \neq 0$ for all $t \in [0, 1]$, the value of the relative growth rate at time t can be taken as:

$$r_t = \frac{m'(t)}{m(t)}.$$

Since $\hat{m}(t) \to m(t)$ and $\hat{m}'(t) \to m'(t)$, a consistent estimate of the relative growth rate r_t is given by:

$$\hat{r}_t = \frac{\hat{m}'(t)}{\hat{m}(t)} \to r_t.$$

Let us consider the nonparametric regression model with multiplicative error of the form

$$y_t = m(t)e^{\varepsilon_i}, \tag{7}$$

where y_t is the observation of the tth time point, $m(t)$ is the trend function, which is assumed to be smooth, and ε_t are random errors with mean zero and finite variance. The model (7) can be transformed to the simple regression model of the form

$$\log(y_t) = \log[m(t)] + \varepsilon_t \quad \text{or}$$

$$z_t = m_1(t) + \varepsilon_t$$

The transformed regression function m_1 and its slope m_1' can be estimated using the same method given in the case of additive errors. In the case of multiplicative model, the estimate for the relative growth rate r_t is given by:

$$r_t = \frac{m'(t)}{m(t)} = m_1'(t).$$

And the estimate of r_t is obtained by:

$$\hat{r}_t = \hat{m}_1'(t) \to r_t.$$

2.1. *Estimation of Change Points*

To estimate the trend and growth rate functions, we assume that the functions are smooth. But in many situations there exists sudden changes or jumps in the trend and/or growth rates. Local linear regression smoother is used to obtain a smooth fit of a regression or trend function. The estimators of the location and size of change points in a function are obtained by fitting local linear regression with dummy variables for the jumps. This method does not require that the number and order of jumps to be known in advance as do most other existing methods (Loader, 1996; Muller, 1992). Other related works can be found in Jose and Ismail (1999, 2001) and Bowman et al. (2006).

A jump point for the regression model (1) at $\tau \in [h, 1-h]$ with jump sizes Δ_0 and Δ_1 for the function m and its slope m', respectively, is defined in the following sense:

$$m(\tau+) - m(\tau) = \Delta_0 \quad \text{and} \quad m'(\tau+) - m'(\tau) = \Delta_1.$$

For any change point τ, we have some i, $1 \le i \le n$, such that $t_i \le \tau \le t_{i+1}$. However, the data cannot be used to distinguish possible changes in this interval. Therefore we assume that the change points occur at any of the design points in the interval $[h, 1-h]$ and the distance between any two change points is greater than h.

Let there exist a jump point for the regression model (1) at $t_\tau \in [h, 1-h]$ with jump sizes Δ_0 and Δ_1 for the function m and its slope m', respectively, then the minimization problem (2) becomes

$$\text{Minimize} \sum_{j=1}^{n} \{y_j - a_0 - a_1(t_j - t) - [\Delta_0 - \Delta_1(t_j - t_\tau)] I_{(t_\tau,1)}(t_j)\}^2 K\left(\frac{t - t_j}{h}\right),$$

where I is the indicator function and $b = [a_0\ a_1\ \Delta_0\ \Delta_1]$ be the coefficient vector.

To estimate the change point, fit the following weighted least squares regression corresponding to all $t_k \in [h, (1-h)]$:

$$\text{Minimize} \sum_{j=1}^{n} \{y_j - a_0 - a_1(t_j - t_k) - [\Delta_0 - \Delta_1(t_j - t_k)] I_{(t_k,1)}(t_j)\}^2 K\left(\frac{t_k - t_j}{h}\right).$$

The solution to the above weighted least squares problem is given by:

$$\hat{b}(k) = \left[\hat{a}_0 \quad \hat{a}_1 \quad \widehat{\Delta}_0 \quad \widehat{\Delta}_1\right] = [T_k W_k T_k']^{-1} T_k W_k Y$$

where

$$T_k = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ (t_1 - t_k) & (t_2 - t_k) & \cdots & (t_k - t_k) & (t_{k+1} - t_k) & \cdots & (t_n - t_k) \\ 0 & 0 & 0 & 0 & 1 & \cdots & 1 \\ 0 & 0 & 0 & 0 & (t_{k+1} - t_k) & \cdots & (t_n - t_k) \end{bmatrix}$$

$$W_k = \text{diag}\left\{K\left(\frac{t_k - t_1}{h}\right) \cdots K\left(\frac{t_k - t_n}{h}\right)\right\} \quad \text{and} \quad Y' = \left[y_1\ y_2 \cdots y_n\right].$$

The regression sum of squares due to $\widehat{\Delta} = [\widehat{\Delta}_0 \quad \widehat{\Delta}_1]$ is given by

$$SSR_k(\Delta) = \begin{bmatrix} 0 & 0 & \widehat{\Delta}_0 & \widehat{\Delta}_1 \end{bmatrix}' T_k W_k Y.$$

The residual sum of squares is given by:

$$SSE_k = Y'\{I - (T_k W_k)'[(T_k W_k)(T_k W_k)']^{-1}(T_k W_k)\}Y.$$

The ratio of the mean regression sum of squares of Δ to the mean residual sum of squares with $t = t_k$ is given by:

$$s_{t_k} = \frac{SSR_k(\Delta)/2}{SSE_k/(2nh-3)}.$$

The estimate of the jump point is given by:

$$t_{\hat{\tau}} = \underset{t \in [h, 1-h]}{\arg\max}(s_t)$$

and the corresponding estimates of the coefficient vector $\Delta = [\Delta_0 \quad \Delta_1]$ be the estimates of the jump sizes. The above procedure can easily be extended to the case of more than one jump point. Let there be q jump points for the regression function m and/or its derivative at t_{τ_j}, $j = 1, \ldots, q$, then the estimates of the change points are given by:

$$t_{\hat{\tau}_j} = \underset{t \in A_j}{\arg\max}(s_t), \quad j = 1, \ldots, q,$$

where

$$A_j = [h, (1-h)] - \bigcup_{k=1}^{j-1} [(t_{\hat{\tau}_k} - h), (t_{\hat{\tau}_k} + h)]$$

and the corresponding estimates of the coefficient vector $\Delta = [\Delta_0 \quad \Delta_1]$ be the estimates of the jump sizes. If the number of jump points is not known in advance, the above sequential procedure continues for $j = 1, \ldots, \rho$ (say), where ρ is fixed in such a way that the $\max(s_t)$, $t \in A_\rho$ is greater than or equal to its critical value $C_\alpha(\rho)$ and $\max(s_t)$, $t \in A_{\rho+1}$ is less than its critical value $C_\alpha(\rho+1)$. The computation of critical value is given below.

2.2. *Computation of Critical Values*

It can be proved that s_t asymptotically follows an F distribution with $(2, 2nh-3)$ degrees of freedom. Using the confidence bounds for the maximum of the variance ratios (David, 1981), the critical value $C_\alpha(\rho)$ of $\max(s_t)$, $t \in A_\rho$ with α level of significance is given by:

$$C_\alpha(\rho) = F_{\alpha^*}(2, 2nh-3),$$

where $F_{\alpha^*}(2, 2nh-3)$ is the value of the F distribution with $(2, 2nh-3)$ degrees of freedom and $\alpha^* = 1 - (1-\alpha)^{1/n^*(\rho)}$ level of significance and $n^*(\rho)$ is the number of

observations in the set A_ρ. Note that if there is a jump point in the trend function m or its slope m', then there will be a jump in the growth rate function also. Therefore, estimating jump or shift point(s) in the growth rate function is equivalent for estimating jump point(s) in the trend function and/or its derivative. The trend or growth rate function with jump points can be estimated by fitting piece-wise local linear regression in between the estimated jump points.

3. Simulation Studies

Simulation studies are carried out to see the performance of the proposed estimators of the trend and growth rate functions. The following two regression models correspond to the additive and multiplicative error terms are considered for the simulation study:

$$y_t = 2[1 + \sin(3t/n)] + \varepsilon_t, \quad t = 1, \ldots, n \tag{8}$$

$$y_t = 2e^{(1+2\sin 2t/n)+\varepsilon_t}, \quad t = 1, \ldots, n, \tag{9}$$

where y_t is the value of the observation at time t, the error process ε is taken as $N(0, \sigma^2)$, with $\sigma = 0.50$. Based on the above 100 independent sets of observations are generated for both the models with different values of n ranging from 50–1,000. The estimates for trend and growth rate functions for the simulated data are obtained using the method given in Sec. 2. The average mean square errors (MSE) of the estimated values with their true values for both the regression models are given in Table 1. Note that the estimated values converges to the true values as n increases.

To see the performance of the estimators of the change points, the following discontinuous regression model is considered:

$$y_t = \begin{cases} 2 + 4\sin(2t/n) + \varepsilon_t, & t = 1, \ldots, n/2 \\ y_{n/2} + 1 + \sin(t/n) + \varepsilon_t, & t = n/2 + 1, \ldots, n \end{cases}, \tag{10}$$

where y_t is the observed value at time t, ε is taken as $N(0, \sigma^2)$ and $\sigma = 0.50$. Based on the above 100 independent sets of observations are generated for different values of n ranging from 50–1,000. The average mean square errors (MSE) of the

Table 1
Average MSE of the estimates of the trend (m) and growth rate (r) of 100 sets of simulated data of model (8) and model (9)

	Model (8)		Model (9)	
n	MSE (m)	MSE (r)	MSE (m)	MSE (r)
50	0.026	5.16×10^{-5}	15.73	3.60×10^{-4}
100	0.013	1.78×10^{-5}	7.55	4.36×10^{-5}
200	0.008	1.05×10^{-5}	5.21	2.03×10^{-5}
500	0.003	6.02×10^{-6}	4.77	1.08×10^{-5}
1000	0.002	2.72×10^{-6}	4.52	8.45×10^{-6}

Table 2
Average MSE of the estimates of the trend (m), growth rate (r), and change point (τ) of 100 sets of simulated data of the regression model (10)

n	MSE (m)	MSE (r)	MSE (τ)
50	4.25×10^{-2}	8.59×10^{-5}	4.40×10^{-3}
100	2.09×10^{-2}	1.53×10^{-5}	2.30×10^{-3}
200	1.36×10^{-2}	3.46×10^{-6}	1.70×10^{-3}
500	0.67×10^{-2}	2.16×10^{-7}	7.01×10^{-4}
1000	0.26×10^{-2}	2.22×10^{-8}	1.83×10^{-4}

estimated values with the true values of the regression function, growth rate function and change points are given in Table 2. Note that MSE of all the estimates are approaching to zero as n increases.

4. Trend and Growth Rate of Wheat Production in India

The proposed procedure is applied to the data of wheat production in India from 1951–2005. The analysis indicated a sudden shift in the trend and growth rate in the year 1967 (Figs. 1 and 2). This sudden shift in trend and growth rate of wheat production may be the impact of the green revolution started in India during the middle of the 1960's. The data-driven technique shows that the growth rate of wheat production in India was decreasing from 1951 to 1967 and there was sudden increase in growth rate during 1967 and again the growth rate started declining (Fig. 2). But these changes can not be observed in the case of growth rate estimated using the traditional parametric approach.

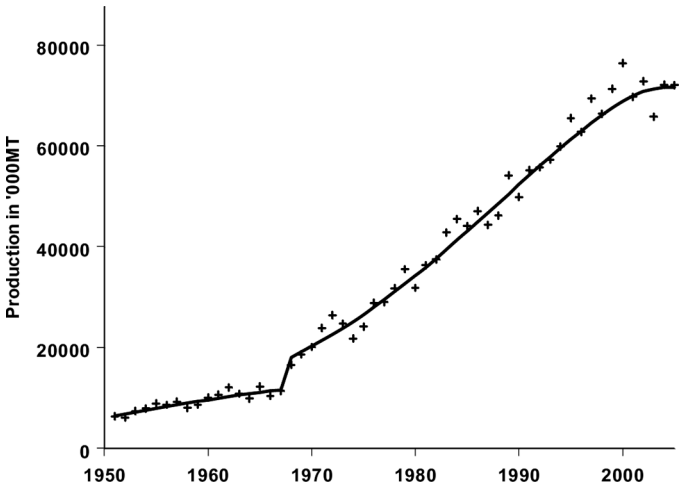

Figure 1. The wheat production values are represented by the symbol '+' and the estimated trend function is represented by the solid line.

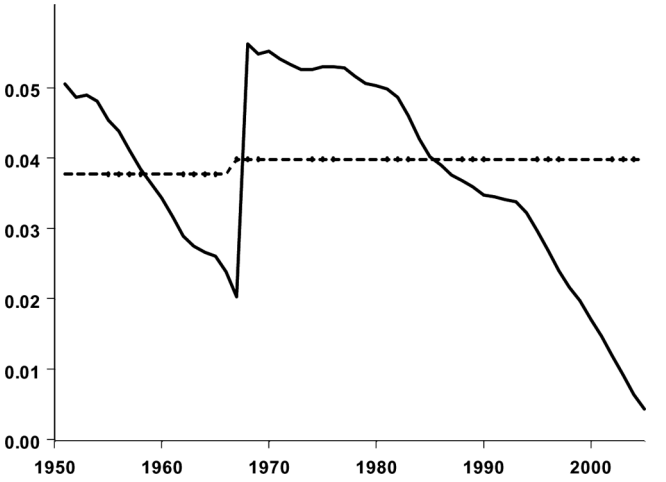

Figure 2. The dotted line represents the exponential growth rate function fitted before and after the shift point 1967 and the solid line represent the growth rate function estimated using the data-driven approach.

5. Conclusion

The nonparametric or data-driven approach used in this article is more flexible than the usual parametric methods. In parametric approaches to study the trend and growth rate, we assume some parametric functional form and draw inferences based on the estimated parameters. In the present nonparametric approach, the only assumption about the form of the trend and growth rate functions are that they are smooth. The method is also extended to the case of sudden shifts or changes in the trend or growth rate functions. The local changes in trend and growth rate can be studied using the present approach whereas, in the parametric approach, we are estimating the paramaters for an interval and that will not reflect the local changes within the interval.

References

Bowman, A. W., Pope, A., Ismail, B. (2006). Detecting discontinuties in nonparametric regression curves and surfaces. *Statistics and Computing* 16:377–390.

David, H. A. (1981). *Order Statistics*. New York: Wiley.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87:998–1004.

Jose, C. T., Ismail, B. (1999). Change points in regression functions. *Communications in Statistics—Theory and Methods* 28:1883–1902.

Jose, C. T., Ismail, B. (2001). Nonparametric inference on jump regression surface. *Journal of Nonparametric Statistics* 13:791–813.

Loader, C. R. (1996). Change point estimation using nonparametric regression. *Annals of Statistics* 24:1667–1678.

Muller, H. G. (1992). Change points in nonparametric regression analysis. *Annals of Statistics* 20:737–761.