Econometric Reviews, 29(5–6):754–777, 2010

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2010.482001

Taylor & Francis
Taylor & Francis Group

IDENTIFICATION OF CHANGES IN MEAN WITH REGRESSION TREES: AN APPLICATION TO MARKET RESEARCH

William S. Rea[1], Marco Reale[1], Carmela Cappelli[2], and Jennifer A. Brown[1]

[1]*Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand*
[2]*Dipartimento di Scienze Statistiche, Universitá di Napoli Federico II, Naples, Italy*

□ *In this article we present a computationally efficient method for finding multiple structural breaks at unknown dates based on regression trees. We outline the procedure and present the results of a simulation study to assess the performance of the method and to compare it with the procedure proposed by Bai and Perron. We find the tree-based method performs well in long series which are impractical to analyze with current methods. We apply these methods plus the CUSUM test to the market share of Crest toothpaste between 1958 and 1963.*

Keywords Identification of multiple structural breaks at unknown times; Time series analysis.

JEL Classification C01; C14; C32.

1. INTRODUCTION

In the last 50 years considerable effort has been devoted to the detection and location of structural breaks in time series both in the statistical and econometric literature (for recent reviews see Hansen, 2001 and Banerjee and Urga, 2005).

The early cumulative summation, or CUSUM, tests had their origin in industrial quality control as a simple graphical method of detecting small changes in process parameters (Page, 1954). Plotting either the ordinary least-squares residuals or their squares against time is not a sensitive indicator of small or gradual changes in regression parameters. Thus the

Address correspondence to Dr. Marco Reale, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand; E-mail: marco.reale@canterbury.ac.nz

cumulative summation

$$Z_r = \frac{1}{\hat{\sigma}} \sum_{t=1}^{r} z_t; \qquad r = 1, \ldots, T \tag{1}$$

is plotted against time where z_t is the residual at time t and $\hat{\sigma}$ is the estimated standard deviation. The graph is then examined to see if there is a systematic departure of the Z_r from the $E[Z_r] = 0$ line.

Brown et al. (1975) developed formal statistical tests of significance for CUSUM by drawing on the theory of stochastic processes. Under the null hypothesis of no breaks the sum in Equation (1) forms a Brownian bridge.

Chow (1960) proposed a test for the detection of a structural break at a known date and it was only much later that Andrews (1993) proposed a test that was devised for detecting a structural break at an unknown date. Recently Bai and Perron (1998, 2003) proposed a test to detect multiple structural breaks at unknown dates. However their procedure is computationally intensive and is not feasible for long time series or for routine application to many time series.

In this article our focus is on the problem of detecting multiple breaks in the mean occurring at unknown dates. To this extent we propose the use of a fast non-parametric procedure based on regression trees. The remainder of this article is organized as follows. Section 2 outlines the new procedure. Section 3 presents the results of a simulation study which compares the new procedure with the procedure of Bai and Perron (1998, 2003). Section 4 presents an application in market research. Section 5 provides the conclusions.

2. REGRESSION TREES

In the last two decades non-parametric tree based methodologies, or recursive partitioning, have found wide application owing to their computational efficiency which allows them to handle large data sets with relative ease. Probably the best known tree methodology is the Classification and Regression Tree (CART) of Breiman et al. (1993).

Consider the case of a numeric response variable and let (Y, X) be a random vector, with $Y \in R$ and $X \in R^p$. Regression trees seek a function $f(X)$, for predicting Y given values of the predictor variables X.

As error function of the predictor $f(X)$, the mean squared error $E(Y - f(X))^2$ is commonly employed. Use of this measure leads to *least squares regression trees* (LSRT) in which $f(X)$ is the conditional expectation $E(Y \mid X = x)$. Thus, LSRT fit to each tree node the group mean, i.e., the mean of the Y's values falling into the node, because this represents the optimal (or Bayes) prediction minimizing the mean squared error (for complete discussion the reader is referred to Breiman et al., 1993).

Based on a training set $(y_i, x_{i1}, \ldots, x_{ip})_{i=1}^{n}$, the algorithm proceeds by recursively splitting the data into two subsets. Any split is a binary question of the form "Is $x_j \in A$?", so that in the case of a numeric predictor variable, the set of possible splits includes all questions: Is $x_j \leq c$?, for c ranging over the domain of x_j. The split induces a partition of the observations y_i: the left descendant nodes h_l satisfying $\{x_{ij} \leq c\}$ and the right descendant node h_r satisfying $\{x_{ij} > c\}$.

Thus, at any node h the algorithm selects the split s which maximally distinguishes the response variable in the left and the right descendant nodes providing the highest reduction in deviance

$$SS(h) - [SS(h_l) + SS(h_r)] \tag{2}$$

where $SS(h) = \sum_{y_i \in h}(y_i - \bar{y}(h))^2$, $(i = 1, \ldots, n)$, is the sum of squares for node h, and $SS(h_l)$ and $SS(h_r)$ are the sums of squares for the left and right descendants, respectively.

As h_l and h_r are an exhaustive partition of h, $SS(h)$ represents the total sum of squares $TSS_y(h)$ at node h and $[SS(h_l) + SS(h_r)]$ the within-child nodes sum of squares, the splitting criterion stated in Equation (2) consists in minimizing the within-groups sum of squares $WSS_{y|s}(h)$. (We note that Equation (4) is calculated in the same manner.) Once the binary partition of a node is found, the splitting process is applied separately to each subgroup, and so on recursively until the subgroups either reach a minimum size or no improvement of the criterion can be achieved.

We show that LSRT provide a practical tool for locating structural breaks in the mean of long time series data. The recently proposed procedure of Bai and Perron (1998, 2003) (hereafter BP) is also based on the method of Fisher (1958) of exact optimization and produces an optimal partition of a time series. However, it is computationally expensive. A number of financial and geophysical time series such as stock market volatilities, tree-ring indices, mud-varve sequences, and ice core data are very long. Sometimes the geophysical series exceed 10,000 data points with annual resolution. The long compute times and large memory requirements of the BP makes its use impractical on these types of series.

Consider the time series model:

$$y_t = \mu_g + \epsilon_t, \qquad g = 1, \ldots, G, \qquad t = T_{g-1} + 1, \ldots, T_g, \tag{3}$$

where G is the number of regimes (and $G - 1$ the number of breakdates), y_t is the observed response variable and ϵ_t is the error term at time t (we adopt the common convention that $T_0 = 0$ and $T_G = T$ where T is the series length). This is the pure structural breaks model employed by BP to detect abrupt structural changes in the mean occurring at unknown dates.

The problem is to estimate the set of breakdates $(T_1, \ldots, T_g, \ldots, T_{G-1})$ that define a partition of the series

$$P(G) = \{(1, \ldots, T_1), \ldots, (T_{g-1} + 1, \ldots, T_g), \ldots, (T_{G-1} + 1, \ldots, T)\},$$

into maximally homogeneous intervals such that $\mu_g \neq \mu_{g+1}$. The BP estimation method is based on the least squares principle: for each G-partition, the corresponding least squares estimates of the μ_g's are obtained by minimizing the within-group sum of squares

$$WSS_{y \mid P(G)} = \sum_{g=1}^{G} \sum_{t=T_{g-1}+1}^{T_g} (y_t - \mu_g)^2. \tag{4}$$

thus, the objective function is the same as in LSRT. In particular, the estimated breakdates $(\widehat{T}_1, \ldots, \widehat{T}_g, \ldots, \widehat{T}_{G-1})$ are associated with the partition $P^*(G)$ such that $P^*(G) = \text{argmin}_{P(G)} WSS_{y \mid P(G)}$. In this approach, the breakdate estimators are global minimizers since the procedure considers all possible partitions by using the dynamic programming approach proposed by Fisher (1958) to find the least squares partition of T *contiguous* objects into G groups. Fisher shows that the number of computations can be substantially reduced by exploiting the additivity property of the sum of squares criterion by means of a dynamic programming approach (Bellman and Dreyfus, 1962), but, despite the computational saving, the method cannot deal with high values of T and G and the same remark holds for the BP's procedure, even with today's computing power.

In the case of LSRT time assumes the role of the predictor variable when, in fact, it is merely a counter. The absence of a true predictor variable and lack of distributional assumptions leads us to call this application Atheoretical Regression Trees (ART).

Let k be an arbitrary ascending (or descending) sequence of completely ordered numbers, for sake of simplicity take $k = 1, 2, \ldots, i, \ldots, T$. Tree regressing the series y_t – whose breaks are to be located – on k yields nested partitions of y_t whose split points represent candidate break dates.

Note that while in the original Fisher's method (as in BP's) optimal partitions for different values of G need not be hierarchically nested, in ART, as the binary search algorithm goes on, the previous partitions are fixed. Thus, after several splits there's no guarantee that the global optimum i.e., the absolute minimum within-groups sum of square partition is reached. It is so only after a single split but, as noticed in Gordon (1973) for many sets of data binary divisions represents a reasonable approximation providing good partitions (see also Edwards and Cavalli-Sforza, 1965).

In the case of time series data Hartigan (1975) provides an excellent justification in favor of the (faster) binary division algorithm: suppose that the observed time series consists of G segments within each of which the values are constant, i.e., the model in Equation (3) becomes a piecewise constant model with $\epsilon_t = 0$. The series can be partitioned into G segments where for each segment the within-group sum of squares is zero. This partitioning can be identified by a sequential splitting algorithm such as the one in LSRT.

Also, because the observations are ordered by time, misplacements can occur only on the boundaries. As discussed in Hansen (2001), although structural breaks are treated as immediate, it is more reasonable to think that they take a period of time to become effective, thus misplacements on the boundaries are not a concern.

Despite the potential suboptimal solutions, LSRT has a further advantage over global search algorithms as used in the BP method of being computationally fast. The global search algorithm requires $O(n^2)$ steps, whereas ART, at any tree node requires $O(n(h))$ steps to identify the best split, where $n(h)$ is the number of values in node h.

Another distinction between ART and the global search algorithm is in the selection of the final partition and the consequent set of break dates. Indeed, methods such as Fisher's (and BP's) have the drawback of producing a single partition for a prespecified value of G and, in general, it is advisable to produce and compare more partitions by varying G. In the case of LSRT this is not a concern because the method produces a hierarchical tree structure associated with the breaks. The selection of the final set of breakdates can be handled within the framework of tree methods by pruning. Pruning is the process of retrospectively discarding branches whose contribution to the reduction of the error is negligible, for details see Breiman et al. (1993, Ch. 3). In this way a nested sequence of partitions and candidate breakdates is created. In order to select the optimal sequence corresponding to the actual number of break dates and distinct subperiods present in the data, cross validation (CV), the sequential testing procedure of BP and model selection criteria can be employed, for details on the use of model selection criteria in regression trees see Su et al. (2004). Moreover, the inspection of the tree structure allows an insight into the partitioning process. Breakdates can be ordered based on their position in the tree and the reduction of the error function achieved. For this reason manual pruning based on subjective choices of the analyst can be preferred to an automatic procedure, see Zhang and Singer (1999, Ch. 4).

Finally, note that if estimation is not the sole concern and one wants to test for structural breaks or model the observations in the segments, it can be appropriate to consider restrictions on the possible values of the breakpoints as suggested by BP. Indeed, extra conditions on the reduction

in deviance and/or on the length of the subperiods are easily handled within the tree growing recursive partitioning approach of ART.

3. SIMULATION EXPERIMENTS

There are several questions to be addressed in applying regression trees to time series. These are:

1. As ART fits piecewise constant functions to data, does ART discover or impose breaks on a time series?
2. What is the effect of serial correlations on ART's performance in detecting structural breaks?
3. Given that observations in time series are, in general, non-interchangeable, can cross-validation be used in tree selection?
4. Is ART robust to non-Gaussian noise structures?
5. Is ART robust to outliers in the data?
6. Can confidence intervals be established for the breaks?

In this section we present the results of simulation experiments using ART to detect structural breaks to answer the above questions. For comparison purposes we used the estimation procedure proposed by BP based on Fisher's method of exact optimization.

For ART we used tree growing and pruning procedures as implemented in `tree` (Ripley, 2005) as a contributed package in the `R` software. For the BP method we used the contributed package `strucchange` (Zeileis et al., 2002) in `R` (R Development Core Team, 2005). In the implementation of the BP method in `strucchange` the user selects the minimum segment size rather than the parameter G discussed above. In all simulations we left the minimum segment size at 0.1 times the length of the series. The implementation we used did not incorporate later work published by Bai and Perron (2006) consequently did not include later corrections for autocorrelation or heteroskedasticity.

3.1. Uncorrelated Series with a Single Break

A set of simulations were run with series of uncorrelated observations drawn from standard Normal, geometric and gamma distributed populations with a single break point at the midpoint of the series giving two equal length regimes.

The gamma distribution is given by

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

The choice of gamma parameters was $\alpha = 2$ and $\beta = 1$. The mean is $\mu = \alpha\beta = 2$ and the variance $\sigma^2 = \alpha\beta^2 = 2$.

For a geometric distribution

$$g(x;\theta) = \theta(1-\theta)^{x-1} \qquad \text{for } x = 1, 2, 3, \ldots$$

$$\mu = \frac{1}{\theta}$$

and

$$\sigma^2 = \frac{1-\theta}{\theta}.$$

The choice of parameter for the geometric distribution was $\theta = 0.5$. Thus $\mu = 2$, and $\sigma^2 = 1$.

One of the beautiful aspects of the standard normal distribution cannot easily be reproduced by the gamma or geometric distributions. In the standard normal the standard deviation and the variance are equal in magnitude. This difference affects the interpretation of the break size parameter. The variance was two for the gamma distribution we used in the simulations. We chose to keep the break size numerically the same for these simulations though it could no longer be interpreted as being in standard deviations of the noise term.

In most simulations there were 16 regime sizes, 5^2 to 20^2 observations in length. Thus in the graphs of the results the axis labeled "Regime Number" is non-linear in scale. The break sizes ranged from 0.05 to 2 standard deviations in steps of 0.05 standard deviations. Unless otherwised stated the `mincut` parameter, which determines the small allowable terminal node, was left at the default value of 5 and 1000 replications of each combination of regime length and break size were run in each of the simulations.

For the comparable results from the BP the break sizes ranged from 0.1 to 2 standard deviations in steps of 0.1 standard deviations. The longest series were composed of 256 data points per regime (regime 16). 100 replications were run of each parameter combination.

Figure 1 presents the results for the ART and BP for a single break at the midpoint of the series. Note that direction of the regime number axis is reversed in the BP results compared to the ART results. When the series were short ART was very prone to over-fitting but this tendency gradually disappeared by a series length of approximately 700 data points (regime 18 or 19). ART had a tendency to over-fit for smaller breaks and for shorter regimes. BP tended to under-fit for smaller breaks and for shorter regimes.

We tested ART's ability to find the location of the break when the break was not at the midpoint but was within the first half of the series.

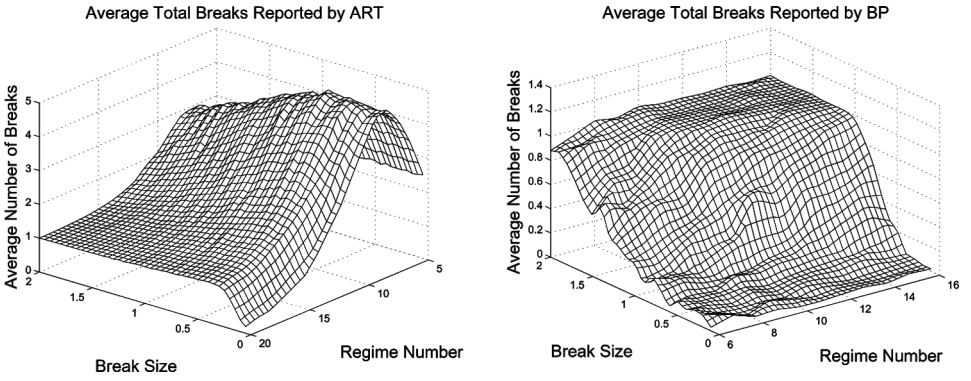

FIGURE 1 Left panel: Average total number of breaks reported by ART. Right panel: Average total number of breaks reported by BP. Simulated series of uncorrelated observations with Gaussian noise and a single break. Pruning based on cost complexity using deviance. Break size is measured in terms of standard deviations. Regime number refers to the length of the series, regime 5 is length 5^2 and the series is 2×5^2 data points long, regime 20 is length 20^2 and the series is 2×20^2 data points long.

We examined series with 100, 400, and 1600 observations. The BP was not run for comparison. We present the results for the 400 and 1600 data point series, the remainder are available on request from the authors.

The results for the 400 and 1600 observation series are presented in Figure 2. At these series lengths there was no evidence of over fitting. The dominant factor in locating a break was its size rather than its location. Unsurprisingly, it was more difficult for ART to locate the break when it was close to the start of the series.

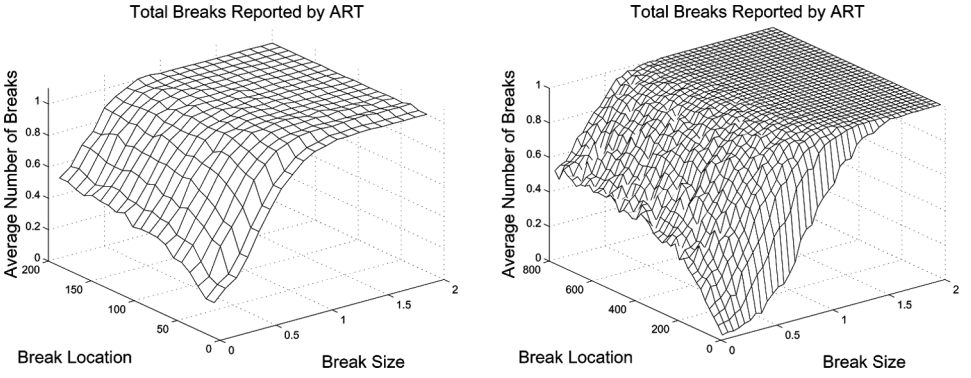

FIGURE 2 Average total number of breaks reported by ART for simulated series of 400 (left panel) and 1600 (right panel) uncorrelated observations, Gaussian noise and a single break at different locations and BIC pruning.

3.2. Uncorrelated Series with Multiple Breaks

An anonymous referee suggested we examine ART's ability to correctly locate breaks when there are two offsetting breaks of equal size, having described this setup as "notoriously difficult". The results of these simulations are presented in Figure 3.

The left panel of Figure 3 presents the total number of breaks reported by ART for these two break series. This graph is similar to a number of others presented here in that the regression tree reported a number of spurious breaks when the series were short. Once the series became long enough, here about regime 16 (i.e., 256 data points per regime), this tendency disappeared.

The right panel of Figure 3 presents the locations of the breaks for the regime 16 series as this was the shortest series for which ART did not over fit. As can be seen the tree reported the location of the two breaks as being near the correct locations at data points 256 and 512 respectively. As the break size decreased the breath of the interval increased but there were no genuinely incorrect breaks reported.

To further investigate the performance of ART for series with multiple breaks we simulated series with 4 breaks:

$$y_t = \mu_{r_i} + \epsilon_t$$

where

μ_{r_i} = the mean of regime r_i; $i = 1, \ldots, 5$
$\epsilon_t =$ noise terms drawn from an $N(0, 1)$, gamma, or geometric distribution.

In all simulations $\mu_{r_i} = 0$ for $i = 1, 3, 5$ and $\mu_{r_4} = -\mu_{r_2}$. The value of μ_{r_2} started at 2 standard deviations and was decremented to 0.05 in steps

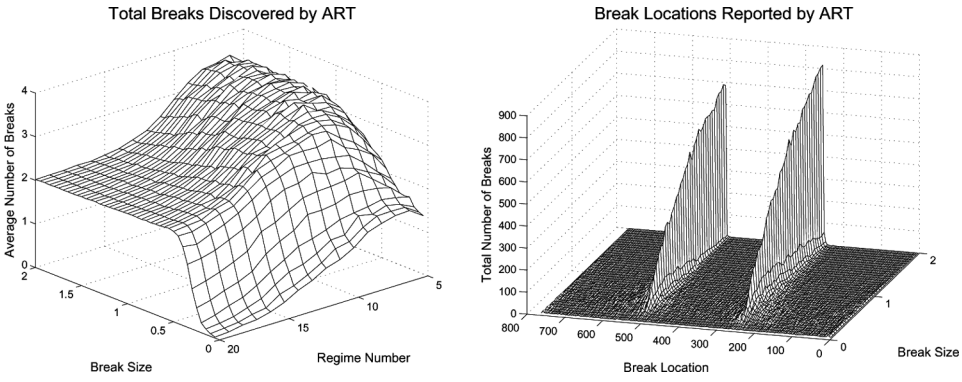

FIGURE 3 Left panel: Total number of breaks reported by ART in the two offsetting break simulations with BIC pruning. Right panel: The locations of the breaks reported for the series with regime length 256 data points, i.e., regime 16 in the left panel. The series had Gaussian noise.

of 0.05. When the BP was used to detect breaks in the series, because the amount of computation required, the value of μ_{n_2} was sometimes decremented to 0.1 in steps of 0.1.

The resultant series were square waves with an amplitude of break size with Gaussian (or other) noise of constant variance imposed on them. The results for the series with Gaussian noise and cost-complexity pruning is presented in the left panel of Figure 4, the results for the BP are presented in the right panel of the same figure. The results for series with gamma and geometrically distributed noise and BIC pruning are presented in the left and right panels of Figure 5 respectively.

As can be seen ART tended to seriously overfit breaks in short series for all three noise structures. The BP, on the other hand, never overfitted breaks but underfitted in a large portion of the parameter regions examined. The series with gamma and geometric noise and BIC pruning are very similar to the results for Gaussian noise and BIC pruning in Figure 6. (BIC pruning is discussed below.) The results for gamma and Gaussian noise were almost indistinguishable. For series with geometric noise ART reported an average of approximately 0.2 breaks per series more than the other two noise structures and the plateau region in the foreground of the figures was correspondingly smaller.

It is well-known that tree-based procedures over-fit small data sets (Cooper, 1998; da Rosa et al., 2008). This can be seen in several figures, for example the left panel of Figure 4. However, as the series lengthened the problem of overfitting reduced and was not evident by regime 15 (length of about 1000 data points). This was where the compute times of the BP began to become excessive (see Section 3.8 below for comparisons of compute times). The BP method underfitted for small breaks particularly for short series.

FIGURE 4 Left panel: Total number of breaks reported by ART in the noisy square wave simulations. Deviance based cost-complexity pruning. Right panel: Total number of breaks reported by BP. The series had four breaks and Gaussian noise.

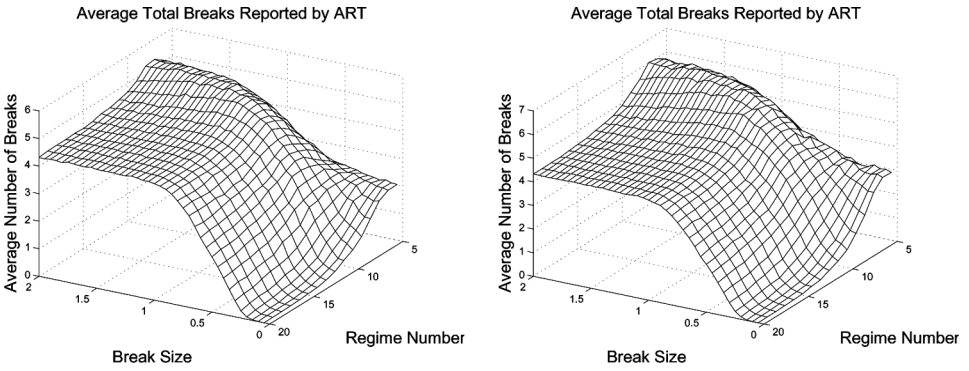

FIGURE 5 The average total number of breaks reported by ART when using BIC pruning for the noisy square wave with gamma distributed noise (left panel) and geometrically distributed noise (right panel). The series had four breaks.

3.3. Alternative Pruning Methods

We examined three tree-pruning methods; deviance-based cost complexity pruning, the default method in the `R` package, the Bayesian Information Criterion (BIC) (Schwarz, 1978) and cross-validation (see Breiman et al., 1993, pp. 306–309) to see if the problem of overfitting in the ART method could be reduced by a more aggressive pruning criteria.

Of the range of information criteria available we selected the BIC on the basis of the results reported by Su et al. (2004) and because it is known to be more robust to non-Gaussian error structures than the AIC (Akaike, 1973).

In time series data observations are usually not interchangeable. Thus the common 10-fold cross-validation cannot be used. The alternative we considered was leave-one-out cross-validation. This minimizes the disturbance to any correlation structure in the data but it is much more computationally expensive than the BIC, requiring N trees to be constructed where N is the number of data points.

The results of BIC and leave-one-out cross-validation are presented in the left and right panels of Figure 6 respectively.

If we compare the left panel of Figure 4 with left and right panels of Figure 6 we see that with BIC tree selection ART pruned away many breaks in the short series resulting in a considerably lower "ridge" feature in those parameter regions. The leave-one-out cross-validation selected even smaller trees in the short series than did the BIC. The high sharp "ridge" of Figure 4 was reduced to a small rise in the right panel of Figure 6.

As observed earlier, no additional pruning was required for long series. The broad plateau area in the parameter regions on the left hand side of the left panel of Figure 4, and both panels of Figure 6 showed that in long

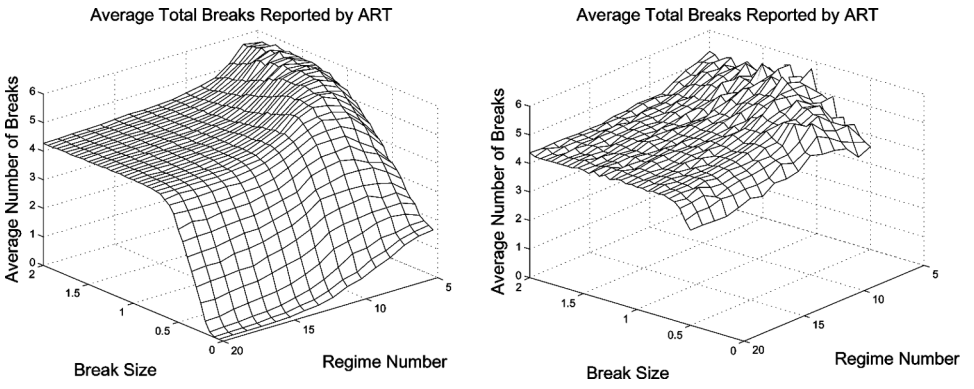

FIGURE 6 The average total number of breaks reported by ART when using BIC pruning (left panel) and leave-one-out cross-validation (right panel). The series had four breaks and Gaussian noise.

series the BIC and leave-one-out cross-validation methods usually did not prune additional nodes from these trees.

For routine tree selection we recommend the BIC for its robustness to non-Normality, but note that for series with more than approximately 600 data points the BIC became indistinguishable from the default cost-complexity pruning.

3.4. Series with Correlated Data

To investigate the ability of ART and BP to detect structural breaks in correlated data we analyzed series with AR(1), AR(2), AR(5), and MA(1) correlations. The results for the AR(1), AR(2) and MA(1) series are presented here and the remainder are available on request from the authors.

An autoregressive process of order one, AR(1) can be expressed as

$$X_t = \phi X_{t-1} + Z_t$$

where ϕ is a constant (the AR(1) parameter) and Z_t is a noise term drawn from a standard normal, gamma or geometric distribution.

The left panel of Figure 7 presents the results for ART for an AR(1) series of 1024 observations in length with varying AR(1) parameter and break size. We considered values of the AR(1) parameter in the range $|\phi| \leq 0.95$. The right panel presents the results of the BP for a break size of two standard deviations. It should be noted that the series standard deviation changes with the magnitude of the AR(1) parameter. The break size was stated in terms of the input noise series.

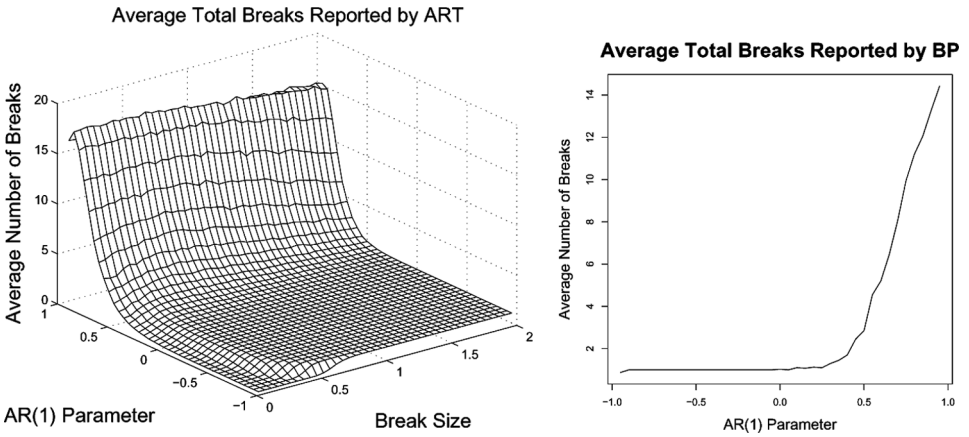

FIGURE 7 Left panel: Average total number of candidate breaks reported by ART in series with AR(1) correlations and a single break. Default pruning. Right panel: Comparable results from the BP for a break size of two.

Both ART and the BP were robust to negative values of the AR(1) parameter and to small positive values (less than approximately 0.25). However, neither break detection method was robust to larger positive values, each reported increasing numbers of spurious breaks as the AR parameter approached unity.

It can be shown (see Chatfield, 2004, p. 115) that the spectral density of an AR(1) process is given by

$$f(\omega) = \frac{1}{\pi(1 - 2\phi\cos\omega + \phi^2)}. \tag{5}$$

Thus for $\phi < 0$ the power is concentrated in the high frequency regions in which the series oscillates rapidly about the mean. For $\phi > 0$ the power is concentrated in the low frequency regions giving the series long excursions away from the mean. It is in these low frequency regions that ART and BP report spurious breaks. We comment further on this in Section 3.5.

An autoregressive process order two, (AR(2)), X_t, can be expressed as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

where ϕ_1, ϕ_2 are constants and Z_t is a noise term as before.

In the AR(2) simulations the series lengths were 1024 data points, all break sizes were two standard deviations in terms the input noise series and a single break at the mid-point of the series. Values of $|\phi_1|, |\phi_2| > 0.5$ were not considered. The variance of an AR(2) process depends on the values of ϕ_1 and ϕ_2 but this does not appear to have been an issue in this set of simulations.

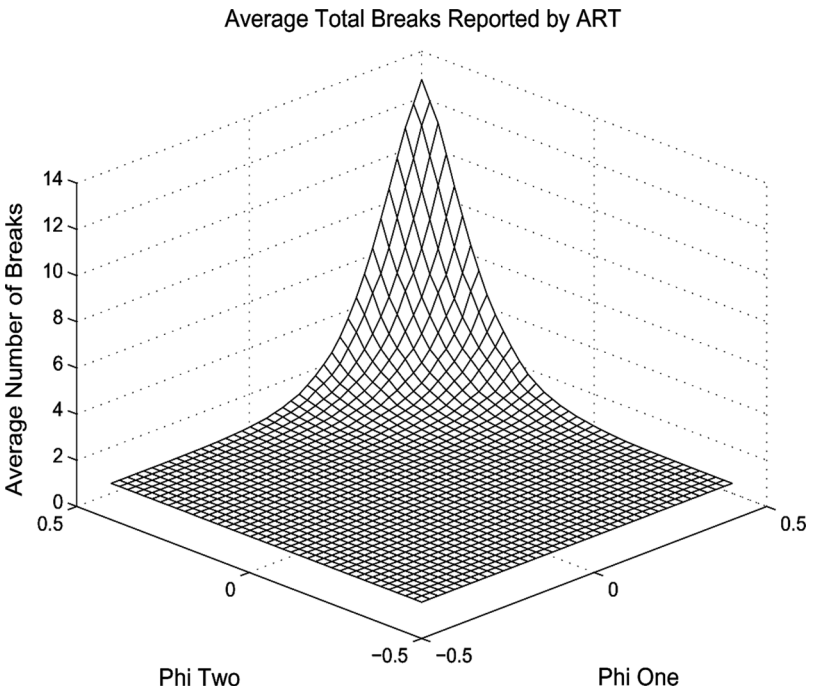

FIGURE 8 Average total number of breaks reported by ART as a function of the two AR(2) parameters with break size two, Gaussian noise and BIC tree selection.

The results are presented in Figure 8. These results can be understood by considering the spectrum of an AR(2) process which can be shown (see Diggle, 1990, p. 77) to be given by

$$f(\omega) = \frac{1}{\left(1 - \sum_{l=1}^{2} \phi_l \cos l\omega\right)^2 + \left(\sum_{l=1}^{2} \phi_l \sin l\omega\right)^2}.$$

In the range of values for ϕ_1 and ϕ_2 which we considered, the power was concentrated in the lower frequencies when both ϕ_1 and ϕ_2 were close to 0.5. As with the AR(1) simulations, the presence of low frequencies in the data was interpreted by ART as structural breaks. This resulted in a decreased ability locate the actual break in a series. This can be seen in the parameter region at the back of Figure 8 in which many spurious breaks were reported. The BP was not run for comparison.

We ran a set of simulations with AR(5) models to compare the effectiveness of leave-one-out cross-validation with cost-complexity and the BIC as a pruning techniques in series with varying levels of serial correlation. The results for each of the three tree selection methods were close to indistinguishable showing there was no preferred method of

pruning in the presence of serial correlation. Graphical presentation of the results are omitted.

3.5. Series with MA(1) Correlations and Non-Gaussian Noise

To assess ART's robustness to non-Gaussian noise we ran some simulations with the geometric and gamma noise. The results for the Gaussian and gamma distributions and MA(1) serial correlations are presented here, the graphic presentation of the geometric results are available on request from the authors.

All series were 1024 observations in length with a single break at the midpoint. The break was added after the full series was generated.

The results for series with Gaussian noise are presented in left panel of Figure 9, the gamma noise in the right panel. There were minimal differences between these two sets of results. There were slightly more breaks reported when gamma noise was used and the plateau region in the foreground of the graphs was slightly smaller. Similar results were obtained with the geometricly distributed noise. Thus while the non-Gaussian noise structures examined did induce larger numbers of reported breaks than Gaussian noise these were not excessive.

It can be shown (see Diggle, 1990, p. 74) that the spectrum of an MA(1) process is given by

$$f(\omega) = 1 + 2\theta \cos \omega + \theta^2.$$

This has some features in common with the AR(1) spectrum, Equation (5), as when $\theta < 0$ the power is concentrated in the higher frequencies. For $\theta > 0$ it is concentrated in the lower frequencies. It differs in that the

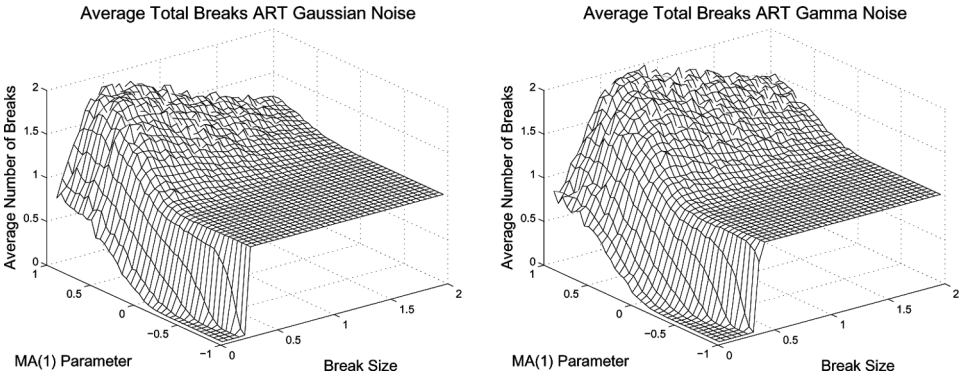

FIGURE 9 Left panel: Total number of breaks reported by ART as a function of MA(1) parameter and break size with Gaussian noise. Right panel: Gamma distributed noise. BIC tree selection.

MA(1) spectrum is bounded as $\theta \to -1$ or $\theta \to 1$. The AR(1) spectrum tends to infinity as $\omega \to 0$ and $\phi \to 1$.

As in the AR(1) case ART tended to report spurious breaks when low frequencies were present in the data. Low frequencies were present in the parameter region toward the back of Figure 9.

For large break sizes and values of θ close to one, ART's tendency to overfit (i.e., report spurious breaks) was still a problem which became progressively more serious as the break size decreased.

3.6. Series with Heteroskedasticity

We examined ART's robustness to heteroskedasticity by simulating series with a break at the midpoint and different standard deviations in the two halves. The first half always had a standard deviation of one. The break size is stated in standard deviations of the first half. The second half had a standard deviation ranging from one to 2.95. We examined two lengths of series, 800 and 1800 data points. We did not run BP for comparison due to the excessive computational times it would require.

The results are presented in Figure 10. ART was more robust to heteroskedasticity in the longer series than in the shorter series. This was consistent with the other observations presented in this article that the problem of over-fitting declined with increasing series length.

3.7. Robustness to Outliers

A set of four simulations were run to assess ART's robustness to outliers. A series of 400 data points was generated from standard normal noise and tested to ensure ART did not report a break in the series. In

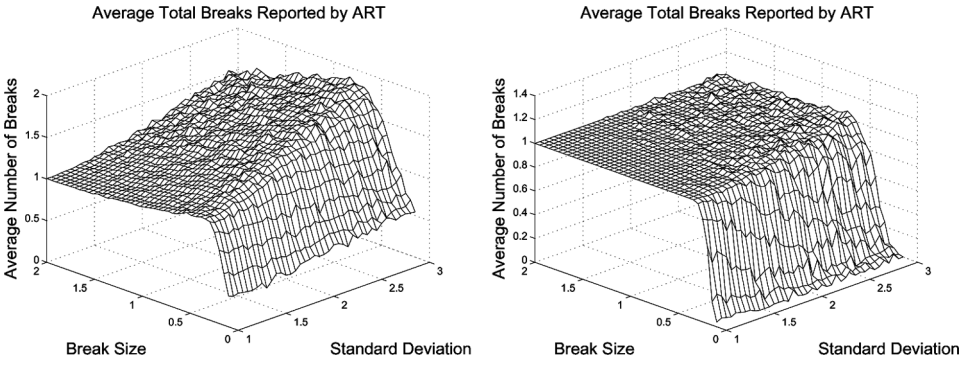

FIGURE 10 Average total number of breaks reported by ART in series with heteroskedasticity. Left panel: 800 data point series. Right panel: 1800 data point series. Deviance-based cost-complexity pruning. The series had one true break.

the four sets of simulations the data point at locations 20, 40, 60, and 80 respectively were then gradually moved up in steps of 0.05 standard deviations until ART reported a break. The distance of the outlier from the mean was recorded in standard deviations. It was clear from these initial four sets of results that ART was becoming more robust to outliers the further the outlier was from the beginning of the series. So we did not investigate this further.

The results are presented in Figure 11. As can be seen, the closer the outlier was to start of a series the easier it was to induce ART to report a spurious break. A small number of trees were visually examined and in most cases ART's solution to the problem of an outlier was to isolate it into a short segment of `mincut` in length.

3.8. Computation Times

We ran some comparisons of compute times between the BP and ART. These are presented in Table 1. As can be seen the compute times for the BP increased rapidly with series length while the compute times for

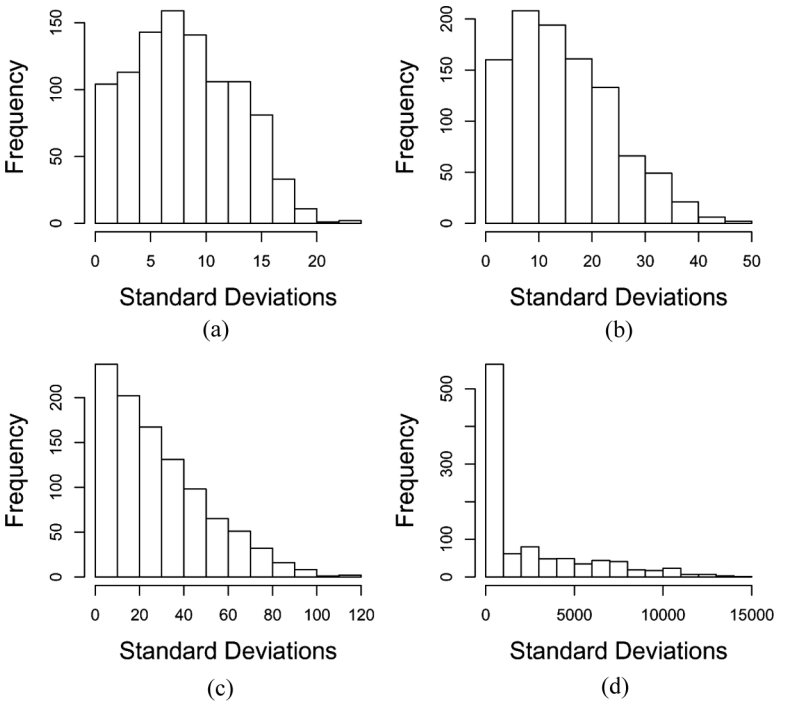

FIGURE 11 Histograms of the size of an outlier required to induce ART to report a break in a series of length 400 at four different positions. (a) Data point 20. (b) Data point 40. (c) Data point 60. (d) Data point 80.

TABLE 1 Comparison of processor times (hours:minutes:seconds) required to run the BP with minimum segment length set to 0.05 and ART for series of various lengths. The times for the ART were so short it was difficult to get accurate timings. All times were run on a SunBlade 1000 with 750Mhz UltraSPARC-III CPU and 2Gb RAM

Series length	Bai and Perron	Regression tree
500	0:03:10.05	0:00:00.03
1000	0:14:08.24	0:00:00.04
1500	2:04:13.25	0:00:00.05
2000	3:41:56.75	0:00:00.07
2500	6:18:34.35	0:00:00.08

ART are short and increased only slowly. Given the disparity between the compute times ART is a useful addition to the practitioner's toolbox when analyzing long time series.

3.9. Nonlinear and Long Memory Time Series

An anonymous referee asked what could be said about long memory and nonlinearity. This is a deep and complex field. In terms of structural break methodologies, it is well known that they may report breaks when only long memory is present. For example, Wright (1998) proved that the probability that the standard CUSUM test would report a break in a long memory series converged to one with increasing series length, see also Kuan and Hsu (1998) and Krämer and Sibbertsen (2002). In addition there is a debate in the literature whether the long memory in financial and economic series is due to structural breaks, see, for example, Diebold and Inoue (2001) and Granger and Hyung (2004). If this question is to be resolved for real series exhibiting long memory it seems likely that structural break methodologies will play a role in the data analysis of such series.

We simulated series with long memory using the function `farimaSim` in the contributed package `fSeries` Wuertz (2005) in `R` (R Development Core Team, 2005). The values of the fractional integration parameter, d, ranged from 0.02 to 0.48 in steps of 0.02 units. We present the results for series lengths 1000 and 8000 data points in Figure 12.

In contrast to the CUSUM test, ART gained robustness to long memory as the series length increased. We did not run comparisons with the BP because of the long compute times required. However, the evidence from Figure 7 suggested that the BP would perform similarly.

Research on the application of ART to long memory time series is on going and will be the subject of a subsequent article.

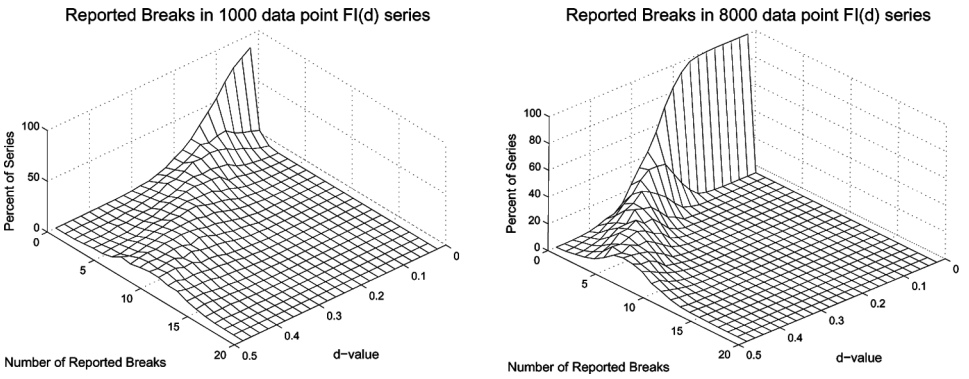

FIGURE 12 Distribution of the numbers of breaks reported by ART in series with long memory. Left panel: 1000 data point series. Right panel: 8000 data point series. Deviance-based cost-complexity pruning. The series had no true breaks.

3.10. Confidence Intervals

The standard method of establishing confidence intervals in non-parametric methods such as ART is via the bootstrap. However, the general solution to bootstrapping a time series is an unsolved problem. In certain artificial circumstances, such as the simulated series presented here, a confidence interval could be established via bootstrapping. But bootstrapping in these circumstances did not offer any advantages over direct simulation. This remains an area worthy of future research.

4. CREST TOOTHPASTE DATA

In this section we briefly review past literature and then apply ART, CUSUM, and BP to the Crest toothpaste data set. These data originate from the Market Research Council of America's panel of household purchase records for the period January 1958 to April 1963. It has been studied by a number of authors such as Jones (1970), Palda and Blair (1970), Montgomery (1971), Shuchman and Riesz (1975), Wichern and Jones (1977), and Tsay (1987).

Crest was introduced to the American market in 1956 and was the second most popular toothpaste after Colgate-Palmolive's Colgate brand for the next four years. On 1 August 1960 the Council on Dental Therapeutics of the American Dental Association (ADA) made an endorsement of Crest as an aid to dental hygiene after determining that Crest's active ingredient, stannous fluoride, did reduce dental caries. Crest's manufacturer, Proctor and Gamble, "reintroduced" the product and in the next few weeks Crest's market share rose dramatically. Despite

vigorous advertising campaigns and special promotions by Colgate and other brands, Crest's market share continued to rise.

Advertising expenditure data are available on a annual basis. Shuchman and Riesz (1975) note that Crest's market share continued to climb in 1962 even though Crest's advertising expenditure declined from \$10,545,000 to \$9,220,000 between 1961 and 1962 after an increase from \$6,242,000 in 1960.

Shuchman and Riesz (1975) state that after the ADA endorsement, the market share of Crest rose to about 35%. After this initial dramatic rise in market share in week 138, Crest's market share did not return to this level again until week 168, more than half a year later. This appears to correspond to the period of intensive counter-advertising studied by Shuchman and Riesz (1975), and "deals" offered by the other manufacturers studied by Montgomery (1971).

Montgomery (1971) and Shuchman and Riesz (1975) studied a number of social, demographic and psychological factors of those who switched to Crest in the post-endorsement period. However, they did not study differences between early and late adopters even though both sets of authors noted the continued rise in Crest's market share after the initial jump following the ADA endorsement.

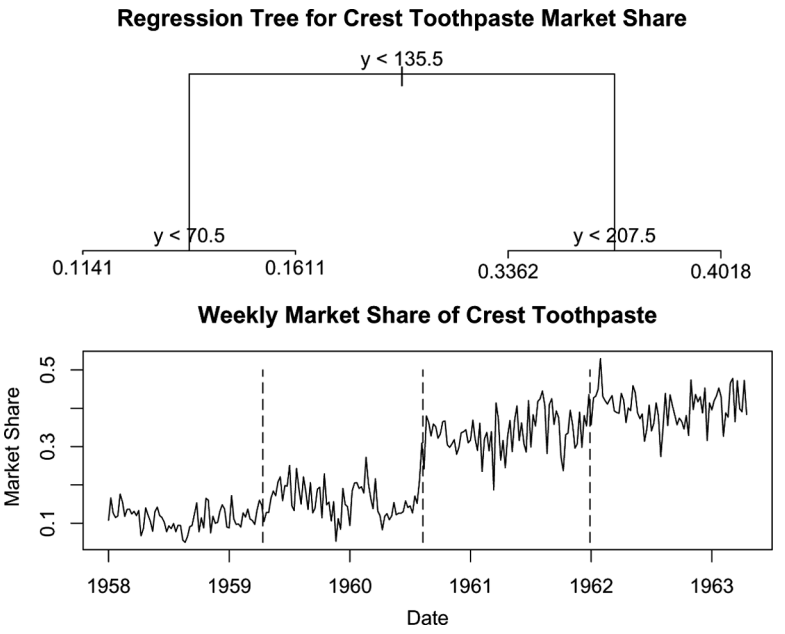

FIGURE 13 The top panel is the regression tree for Crest data. The bottom panel is the plot of Crest data with ART breaks marked by dashed vertical lines.

When we analysed the data with ART we found that the increase in market share reported by these authors occurred in two steps (see Figure 13). The first occurred in early August 1960 immediately after the ADA endorsement to a market share of about 34%. The second was an increase to about 40% market share from the beginning of 1962 (week 207). This two-step rise is also seen in the recursive use of the established CUSUM test (see Figure 14). This is essentially a manual application of recursive partitioning using the CUSUM test as the splitting criteria.

However, the BP reported only two breaks (see Table 2) which exactly match the first two breaks reported by ART. The failure to find a break at or near the 207th week is consistent with our simulation experiments in which the BP never reported too many, but occasionally missed, breaks.

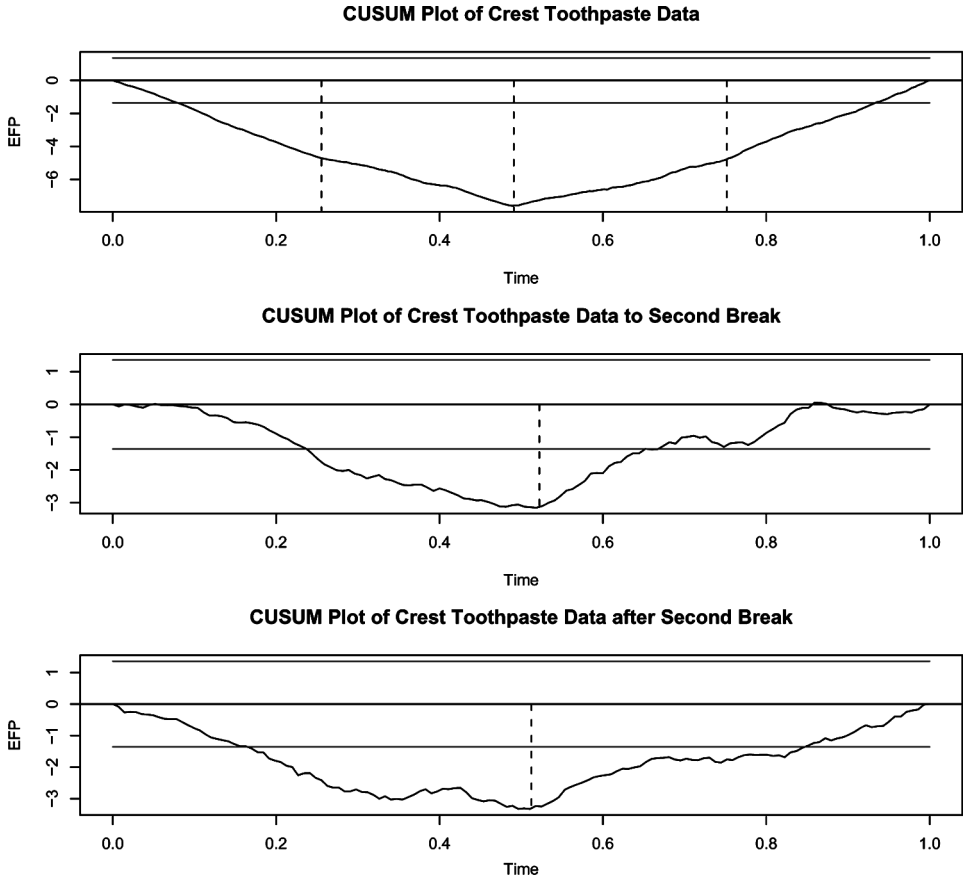

FIGURE 14 The top panel is the CUSUM plot for whole series. The middle panel is the CUSUM plot to the break in early August 1960. The bottom panel is the CUSUM plot from the break in August 1960 to the end of the series. The dashed lines are the ART break points. The thin lines about the heavier zero line are the 5% significance level.

TABLE 2 Break dates reported by ART, the Bai and Perron procedure (BP) and the 95% confidence interval for the BP

ART	BP	BP CI
70	70	62–71
135	135	134–137
207	–	–

Although all three tests find a break at or near week 70 the literature does not suggest an underlying reason for this. The break at week 135 is clearly attributable to the ADA endorsement. The break at week 207 reported by ART and CUSUM was commented on by various researchers only as a subjective assessment of a visual inspection of the data that Crest's market share was continuing to rise. This shows the need for structural break tests such as ART to indicate when substantial changes have taken place in a time series and to allow a fuller investigation of the underlying reasons to be undertaken.

5. CONCLUSIONS AND FUTURE RESEARCH

In this article we presented a very fast procedure, Atheoretical Regression Trees (ART) to detect multiple changes in mean at unknown times. Such a procedure is suitable for practitioners who routinely who deal with long series which may be impractical to analyze with current methods.

The simulations have provided some answers to our six original questions. They are:

1. ART did impose spurious breaks when the series was short but this tendency disappeared as the series becomes longer. This was seen in both single and multiple break, heteroskedastic and long memory simulations. On the basis of our simulations we suggest a rule of thumb that ART can be worthwhile for series exceeding 600 data points when BIC pruning is used and 1000 data points if cost-complexity pruning is used. The strength of the serial correlation also needs to be checked. For AR(1) models with $\phi > 0.25$ ART is likely to report spurious breaks. Robustifing ART to serial correlation is the subject of on-going research.
2. ART was robust to negative serial correlation and a small amount of positive correlation, but in this regard ART was no worse than the BP.
3. Leave-one-out cross-validation can be used for tree selection but was computationally expensive.

4. ART reported higher numbers of breaks for series with non-Gaussian noise structures but not excessively so.
5. ART's robustness to outliers depended on the outlier location. For an outlier 20 data points from the end of the series a outlier needed to be a mean of over 8 standard deviations away from the series mean to cause ART to spuriously report a break when none existed. This increased to a mean of over 2300 standard deviations when the outlier was at data point 80.
6. It is possible to obtain a confidence interval for the breakpoints using the bootstrap. However, this is currently limited to series without serial correlations and may thus be of little practical use.

The ART procedure can be easily implemented in any software that provides the classification and regression trees methodology.

There are some clear differences between ART and the established Bai and Perron procedure. ART overfitted short time series, i.e., it reported breaks where none existed in the data generating process. It performed well in the simulations for long series for which the compute times for the BP may be prohibitive. The BP procedure on the other hand had the tendency to report less breaks than it should when the shift in mean was moderate or small.

ART was computationally faster than the CUSUM procedure as implemented in the `R` package `strucchange` Zeileis et al. (2002). It combined both detection and location of structural breaks in single algorithm.

Section 3.9 only touched on the problem of the duality between structural breaks and long memory in the briefest manner. Investigating the statistical properties of structural break methods in general, not just ART, when applied to long memory time series requires further research.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F., eds. *2nd International Symposium on Information Theory*. Budapest: Akademia Kadio, pp. 267–281.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* 61(4):821–856.

Bai, J., Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* 66(1):47–78.

Bai, J., Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18:1–22.

Bai, J., Perron, P. (2006). Econometric theory and practice: frontiers of analysis and applied research. In: Corbea, D., Durlauf, S., Hansen, B. E., eds. *Theory and Applications of Long-Range Dependence*. Cambridge, UK: Cambridge University Press.

Banerjee, A., Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics* 129:1–34.

Bellman, R. E., Dreyfus, S. E. (1962). *Applied Dynamic Programming*. Princeton, NJ: Princeton University Press.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1993). *Classification and Regression Trees*. Chapman & Hall/CRC.

Brown, R. L., Durbin, J., Evans, J. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society Series B* 37(2):149–192.

Chatfield, C. (2004). *The Analysis of Time Series 6th Edition*. Chapman & Hall/CRC Press.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28(3):591–605.

Cooper, S. J. (1998). Multiple regimes in US output fluctuations. *Journal of Business and Economic Statistics* 16(1):92–100.

da Rosa, J. C., Veiga, A., Medeiros, M. C. (2008). Tree-structured smooth transition regression models based on CART algorithm. *Journal of Computational Statistics and Data Analysis* 52:2469–2488.

Diebold, F. X., Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics* 105:131–159.

Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction*. Oxford, UK: Clarendon Press.

Edwards, A. W. F., Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics* 21:362–375.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53(284):789–798.

Gordon, A. D. (1973). Classification in the presence of constraints. *Biometrics* 29:821–827.

Granger, C. W. J., Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* 11:213–228.

Hansen, B. (2001). The new econometrics of structural change: dating breaks in the US labor productivity. *Journal of Economic Perspectives* 15(4):117–128.

Hartigan, J. (1975). *Clustering Algorithms*. New Jersey: John Wiley and Sons.

Jones, J. M. (1970). A comparison of three models of brand choice. *Journal of Marketing Research* 7:466–473.

Krämer, W., Sibbertsen, P. (2002). Testing for structural change in the presence of long memory. *International Journal of Business and Economics* 1(3):235–242.

Kuan, C.-M., Hsu, C.-C. (1998). Change-Point Estimation of Fractionally Integrated Processes. *Journal of Time Series Analysis* 19(6):693–708.

Montgomery, D. B. (1971). Consumer characteristics associated with dealing: an empirical example. *Journal of Marketing Research* 8:118–120.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* 41(1/2):100–115.

Palda, K. S., Blair, L. M. (1970). A moving cross-section analysis of demand for toothpaste. *Journal of Marketing Research* 7:439–449.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripley, B. (2005). *Tree: Classification and Regression Trees*. R package version 1.0-19.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464.

Shuchman, A., Riesz, P. C. (1975). Correlates of persuability: the crest case. *Journal of Marketing Research* 12(1):7–11.

Su, X., Wang, M., Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics* 13(3):586–598.

Tsay, R. S. (1987). Non-Gaussian state-space modeling of nonstationary time series: comment: detecting and modeling changes in time series. *Journal of the American Statistical Association* 82(400):1056–1059.

Wichern, D. W., Jones, R. H. (1977). Assesing the import of marketing disturbances using intervention analysis. *Management Science* 24(3):329–337.

Wright, J. H. (1998). Testing for a structural break at unknown date with long-memory disturbances. *Journal of Time Series Analysis* 19(3):369–376.

Wuertz, D. (2005). *fSeries: Financial Software Collection*. R package version 220.10063.

Zeileis, A., Leisch, F., Hornik, K., Kleiber, C. (2002). Strucchange: an R package for testing for structural change in linear regression models. *Journal of Statistical Software* 7(2):1–38.

Zhang, H., Singer, B. (1999). *Recursive Partitioning in the Health Sciences*. New York: Springer.