Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 78 (2008) 351–356
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Detecting multiple mean breaks at unknown points in official time series

Carmela Cappelli [a], Richard N. Penny [b], William S. Rea [c], Marco Reale [c,*]

[a] *Dipartimento di Scienze Statistiche, Universitá Federico II, Via Leopoldo Rodinó 22, 80138 Naples, Italy*
[b] *Statistical Methods Division, Statistics New Zealand, Private Bag 4741, Christchurch, New Zealand*
[c] *Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand*

Available online 31 January 2008

Abstract

In this paper, we propose a computationally effective approach to detect multiple structural breaks in the mean occurring at unknown dates. We present a non-parametric approach that exploits, in the framework of least squares regression trees, the contiguity property of data generating processes in time series data. The proposed approach is applied first to simulated data and then to the Quarterly Gross Domestic Product in New Zealand to assess some of anomalous observations indicated by the seasonal adjustment procedure implemented in X12-ARIMA are actually structural breaks.

Keywords: Partitioning; Regression trees; X-12 ARIMA

1. Introduction

In the last 50 years a lot of effort has been devoted to the detection of structural breaks in time series both in the statistic and econometric literature (for a review see Hansen [10]).

Initially, Chow [6] proposed a test for the detection of a structural break at a known date and only much later [1], a test was devised for detecting a structural break at an unknown date. Recently, Bai and Perron [2,3] proposed a test to detect multiple structural breaks at unknown dates, however the procedure is computationally intensive and is not feasible for long time series or for a routine application to many time series.

The detection of structural breaks in time series is of immense interest to anyone that produces and uses time series. National Statistical Offices (NSO), such as Statistics New Zealand, collect, collate and publish data for the use of researchers, policy analysts and the general public. The main concern of NSOs is to release data that reflect the social or economic concept that they are meant to represent, within the budget allocated for this work and, crucially, with little or no revisions after release. As an NSO is supplying time series for a range of users with varying needs and knowledge of statistical analysis much of its output will be descriptive, rather than analytical. Identifying structural breaks is necessary for any NSO.

As the Chief Statistician of the Canadian NSO, Statistics Canada, noted "Credibility plays a basic role in determining the value to users of the special commodity called statistical information. Indeed, few users can validate directly the

* Corresponding author. Tel.: +64 3 364 2987x7681; fax: +64 3 364 2587.
E-mail addresses: carcappe@unina.it (C. Cappelli), richard.penny@stats.govt.nz (R.N. Penny), william.rea@canterbury.ac.nz (W.S. Rea), marco.reale@canterbury.ac.nz (M. Reale).

0378-4754/$32.00
doi:10.1016/j.matcom.2008.01.041

data released by statistical offices" ([7], p. 169). To enable users to use the data supplied by an NSO with confidence considerable work inside the NSO is done to check and report on the quality of the data produced. A key issue is to ensure that the series is consistent throughout its length. By doing this the NSO can assure users that the series is a result of the data generating process, and not the way the NSO has collected, collated and published the data.

Therefore, an NSO is particularly interested in identifying any atypical change in the behaviour of the series. These changes in the series should indicate changes in the data generating mechanism, but may arise through unforeseen effects in the data collection and collation process used by the NSO.

In this paper, the focus is on the problem of detecting multiple breaks in the mean occurring at unknown dates with the proposal of a fast nonparametric procedure based on regression trees. Such a procedure is suitable to NSO's which routinely analyse a large amount of time series.

2. Atheoretical regression trees

Given a continuous response variable Y and a set of p predictors $X_1, \ldots, X_p$, regression trees model the relationship between the response and the covariates employing a recursive partitioning approach that results into a partition of Y based upon the values of the predictor variables. Our procedure makes use of an artificial covariate (so that $p = 1$) that is an arbitrary strictly ascending (or descending) sequence of numbers thus we call it atheoretical regression trees (so forth denoted by ART) [5] because it is theory-free being the covariate not a predictor variable but rather a counter. In what follows we will show that the use of such covariate in least square regression trees [4] resorts to a sequential use of the Fisher's method of exact optimization [9] proposed for grouping n elements into g mutually exclusive and exhaustive subsets having maximum homogeneity i.e., minimizing the within-groups sum of squares. Fisher's algorithm is designed for situations in which the data points are ordered and groups consist of intervals of data. Two subclass of problems are considered: the unrestricted case when the observations can be ordered according to their numerical values, and the restricted one when an a priori ordering is given. Time series data belong to the second case as the ordering is provided by the time and observations are not exchangeable. In this case, seeking the minimum within sum of square partition corresponds to segment the series into homogeneous subperiods that contrast with each other, i.e., it corresponds to detect breaks in the mean.

A drawback of the Fisher's method is that it can deal with moderate-sized values of n and g while ART overcomes these limitations because it corresponds to a sequential application of the Fisher's algorithm to a problem of $g = 2$ subperiods. Furthermore, whereas Fisher's method produces a single partition and it is advisable to create several partitions by varying g, ART produces a hierarchical structure. The final partition and the corresponding set of break dates can result either from automatic procedure such as pruning along with popular model selection criteria or from subjective choice of the applied scientist based on a priori knowledge.

In least square regression trees (LSRT), a node t is split into the left and right descendants t_l and t_r to reduce the deviance of the response variable. Thus, the algorithm selects the split s for which $\mathrm{SS}(t) - [\mathrm{SS}(t_\mathrm{l}) + \mathrm{SS}(t_\mathrm{r})]$ is maximum, where

$$\mathrm{SS}(t) = \sum_{y_i \in t} (y_i - \bar{y}(t)), \quad i = 1, \ldots, n \tag{1}$$

is the sum of squares for node t, and $\mathrm{SS}(t_\mathrm{l})$ and $\mathrm{SS}(t_\mathrm{r})$ are the sums of squares for the left and right descendants, respectively. The splitting criterion is equivalent to maximize the between-groups sum of squares $\mathrm{BSS}(t)$ that can be written as

$$\mathrm{BSS} = \frac{n(t_\mathrm{l})n(t_\mathrm{r})}{n(t)} (\bar{y}(t_\mathrm{l}) - \bar{y}(t_\mathrm{r}))^2 \tag{2}$$

being

$$\bar{y}(t) = \frac{n(t_\mathrm{l})\bar{y}(t_\mathrm{l}) + n(t_\mathrm{r})\bar{y}(t_\mathrm{r})}{n(t)}. \tag{3}$$

Thus, in LSRT the splitting criterion searches for the child nodes consisting of subsets of y values whose means are as far as possible. 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. The resultant tree usually is overly large so that a pruning method is applied to trim it back. Minimizing the within-group sum of squares is a natural criterion for partitioning a single real variable.

This is the case in the Fisher's algorithm of exact optimization whose key aspect is the concept of contiguous partitions. Let i, e and h be three data points that have order $i \leq e \leq h$; a partition is said to be contiguous if it consists of groups that satisfy the following condition: if i and h are assigned to the same class then e must be also assigned to that class. For ordered data only contiguous partitions require to be considered to detect the optimal one minimizing the within-group sum of squares. In the restricted case of time series data the contiguity applies to time i.e., only subsequent intervals in terms of the ordering specified by time are admissible. The number of possible contiguous partitions of n ordered objects into g groups makes a global search unfeasible but 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 that allows to deal with the problem of finding the optimal partition into g groups making use of the results obtained while dealing with the problem of $g - 1$ groups.

Despite the saving, Fisher bounds the capacity of the algorithm to $n = 200$ and $g = 6$ and even with today's computers a complete enumeration and search is achievable in a reasonable time only for $g = 2$. The concept of contiguous partitions can be naturally exploited in the framework of least square regression trees. To this aim, let k be an arbitrary ascending (or descending) sequence of completely ordered numbers, for sake of simplicity take $k = 1, 2, \ldots, i, \ldots, n$. The use of such a sequence as covariate into least square regression trees resorts to create and check at any node t for all the admissible binary partitions of the $y_i \in t$ whose number is $n(t) - 1$ and thus it is treatable. Indeed, the contiguity property ensures that at any node t the best split for the given order lays in k and it will be identified by the splitting criterion. Note that in the original Fisher's method optimal partitions for different values of g need not to be hierarchically nested. In the ART method as splitting goes on, the previous partitions are fixed, but for many sets of data this represents a reasonable approximation providing good partitions at a much less expensive computational cost. Indeed, the Fisher's method requires $O(n^2 g)$ steps, whereas ART, at any tree node requires $O(n(t))$ steps to identify the best split. Hartigan [11] provides an excellent justification in favor of the binary division algorithm in the case of time series data: suppose that the time interval consists of g intervals within each of which the values are constant. Then there is a partition into g segments for which the within sum of squares is zero and it will be identified by the tree algorithm. ART generates a hierarchical structure and a nested sequence of partitions corresponding to candidates sets of break dates can be identified by means of pruning, that is the process of discarding terminal nodes whose contribute to reduction in deviance is negligible. In order to find the subtree whose terminal nodes provide the optimal partition corresponding to the actual number of break dates and distinct subperiods present in the data, we use classical model selection criteria (for a discussion on the use of these criteria in tree methods see [15]).

3. Simulation study

To assess the effectiveness of atheoretical regression trees in detecting multiple structural breaks at unknown dates, the methodology is applied to data simulated by the model:

$$y_t = \mu_{r_i} + \epsilon_t \tag{4}$$

where μ_{r_i} is the mean of regime $r_i; i = 1, \ldots, 5$ and ϵ_t are noise terms drawn from an $N(0, 1)$ 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} (break size) started at 2 standard deviations and was decremented to 0.05 in steps of 0.05.

There where 16 regime sizes (regime numbers), 5^2 to 20^2 observations in length, with break sizes ranging from 0.05 to 2 standard deviations in steps of 0.05 standard deviations and 1000 replications of each combination of regime length and break size.

In essence the resultant series are square waves with an amplitude of break size and Gaussian noise of constant variance imposed on them as shown in Fig. 1.

The noisy square wave simulations were analysed by ART with a cost complexity $\alpha = 0$ and the tree size was selected with the Bayesian information criterion (BIC) [13]. The results from these simulations using ART are presented in Fig. 2.

Ideally four breaks should be detected and when the break-size is moderately large and the regime size is long ART consistently finds the real breaks and generating few spurious candidate breaks without even needing to prune. As expected the erroneous pruning of real breaks is confined to short series and small breaks.

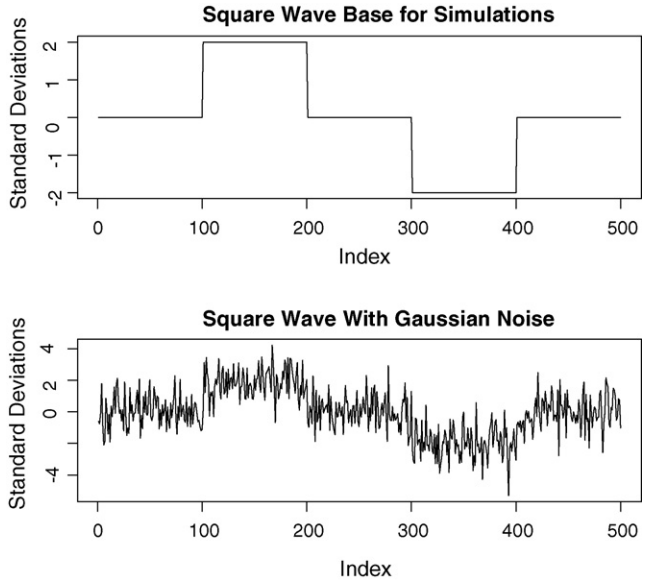

Fig. 1. Noisy square.

4. Application to official statistics

Much of the reporting on NSO outputs focuses on the movements in the time series, rather than the values. Often for series the largest part of the short term movement is caused by annually recurring variation, termed the seasonal variation. For this reason most statistical agencies provide the measured figure along with the seasonally adjusted value (where appropriate), and increasingly the trend estimates, and direct users to the latter series rather than the unadjusted figures.

To seasonally adjust series Statistics New Zealand attempts to separately identify and estimate a seasonal, calendar, trend and irregular component for the time series it releases. The components are generically termed unobserved components. While state space modelling will provide estimates of these unobserved components, many NSO, including Statistics New Zealand, use variants of the Census II Method of the U.S. Bureau of the Census [14,12]. Statistics New

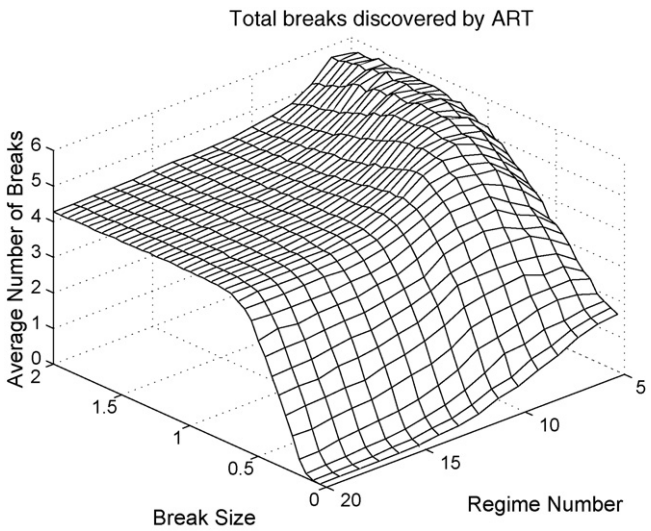

Fig. 2. Total number of breaks found by ART in the noisy square wave simulations.

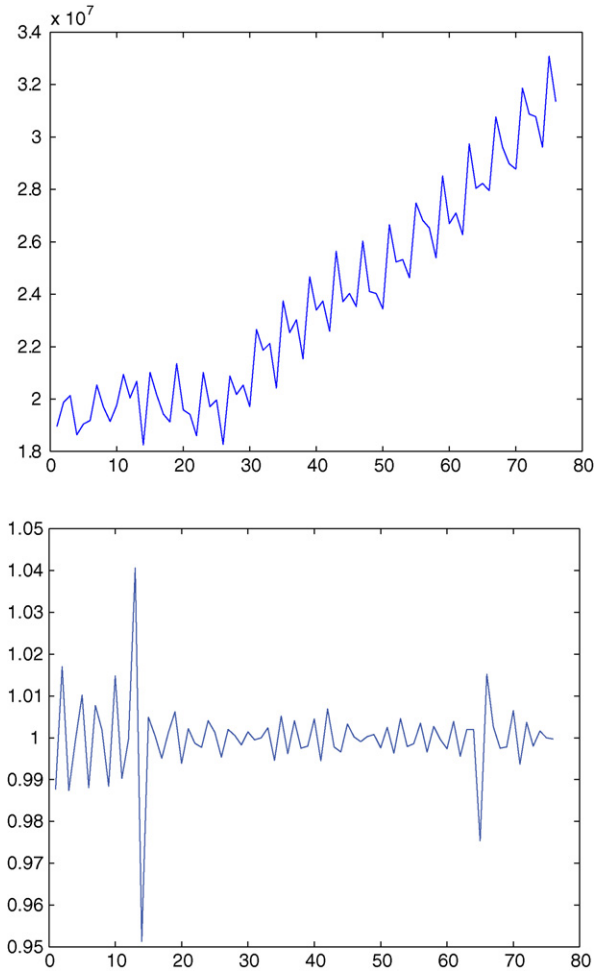

Fig. 3. QGDE (top); irregular component of QGDE (bottom).

Zealand currently uses Census Method II Variant X-12, commonly called X-12 [8]. X-12 uses Henderson moving averages to decompose the original series into the set of unobserved components.

As outliers can cause problems in time series analysis X-12 identifies outliers and records this information in a table, termed the C17 table. If a value is acceptable it will have a weight of 1 in the C17. Values that are unusual have a weight less than one, with a zero weight meaning the value is not used for estimation purposes. Within Statistics New Zealand it has been observed that breaks often result in adjacent zero weights. Is the C17 component a useful indicator for possible structural changes of the data generating process?

Our strategy to answer this question is to identify breakpoints in a time series using atheoretical regression trees and verifying if they were indicated as anomalous by the C17 component (values different from 1). In particular, we focus our attention on the series of the irregulars, which is computed as residual from the other smoothed components. As an illustration the real case of the Quarterly Gross Domestic Expenditure in New Zealand from June 1986 to March 2005 is considered. Fig. 3 shows time plots of the original series and the irregular component.

The C17 table indicates the presence of five anomalous observations that might be either outliers or structural break-dates; these are: June 1989, September 1989, September 1994, September 1997 and December 1999. We have run ART on the irregulars series without setting a minimum segment length for the regimes since we want to use ART as a benchmark for C17. The greedy algorithm grows a large tree with 11 terminal nodes corresponding to 10 candidates break dates. By means of both Cross Validation and BIC we have found a structural break corresponding to March 1989. This observation precedes June 1989, indicating that June 1989 is not to be considered an outlier but rather the beginning of a new regime.

Atheoretical regression trees can be extended to detect also breaks in trend and even volatility and because of their speed they are an ideal instrument for NSO's: once an anomalous value is signalled by the seasonal adjustment procedure they can quickly assess whether a structural break has occurred.

References

[1] D.W.K. Andrews, Tests for parameter instability and structural change with unknown change point, Econometrica 61 (1993) 821–856.
[2] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1998) 47–78.
[3] J. Bai, P. Perron, Computation and analysis of multiple structural change models, Journal of Applied Econometrics 18 (2003) 1–22.
[4] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, 1984.
[5] C. Cappelli, M. Reale, Detecting changes in mean with atheoretical regression trees, University of Canterbury Mathematics and Statistics Department Research Report 2005/02, 2005.
[6] G.C. Chow, Tests of equality between sets of coefficients in two linear regressions, Econometrica 28 (1960) 591–605.
[7] I.P. Fellegi, Characteristics of an Effective Statistical System, International Statistical Review 64 (1996) 165–187.
[8] D.F. Findley, B.C. Monsell, W.R. Bell, M.C. Otto, B.-C. Chen, New capabilities and methods of the X-12 ARIMA seasonal-adjustment program, Journal of Business & Economic Statistics 16 (1998) 127–177.
[9] W.D. Fisher, On grouping for maximum homogeneity, Journal of the American Statistical Association 53 (1958) 789–798.
[10] B. Hansen, The new econometrics of structural change: dating breaks in U.S. labor productivity, Journal of Economic Perspectives 15 (2001) 117–128.
[11] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York, 1975.
[12] D. Ladiray, B. Quenneville, Seasonal Adjustment with the X-11 Method, Spinger-Verlag, New York, 2001.
[13] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[14] J. Shiskin, A.H. Young, J.C. Musgrave, The X-11 Variant of the Census Method II Seasonal Adjustment Program, Technical Paper 15 (revised), U.S. Bureau of the Census, 1967.
[15] X.G. Su, M. Wang, J.J. Fan, Maximum likelihood regression trees, Journal of Computational and Graphical Statistics 13 (2004) 586–598.