ATHEORETICAL REGRESSION TREES FOR DATING BREAKPOINTS IN THE MEAN

Carmela Cappelli[1] and Marco Reale[2*]

Universitá di Napoli Federico II and University of Canterbury

Summary

A non parametric approach is proposed for dating structural breaks whose number and dates of occurrence are *a priori* unknown. In particular, the case of level shifts is considered. For the purpose of locating the breakdates the method exploits, in the framework of least square regression trees, the contiguity property introduced by Fisher for grouping a single real variable. The proposed approach is applied to study the changes in mean water levels of Lake Michigan-Huron.

Key words: Structural breaks detection, Fisher's algorithm, least square regression trees, time series.

1. Introduction

The detection of structural breaks is an important problem in time series analysis that has attracted the attention of both statisticians and econometricians for more than forty years (for a review see Hansen, 2001). In this paper we focus on the problem of detecting multiple breaks occurring at unknown dates. To this aim, recently, Bai & Perron (1998, 2003) have proposed an estimation procedure that makes use of a dynamic programming approach that can be traced back to the Fisher's method of exact optimization (Fisher, 1958) for grouping a single real variable into mutually exclusive and exhaustive subsets having maximum homogeneity, i.e. minimizing the within-group sum of squares.

Fisher introduces the concept of *contiguity* that can be summarized as follows: let i, i' and i'' be three data points such that $i \leq i' \leq i''$, a partition

[1]Dipartimento di Scienze Statistiche, Università di Napoli Federico II, Italy, Via L. Rodinò n.22, Naples I-80138C, Italy.

[2]Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. e-mail:marco.reale@canterbury.ac.nz

∗Author to whom correspondence should be addressed.

Aknowledgements. Carmela Cappelli acknowledges financial support from Dipartimento di Scienze Statistiche.

is said to be contiguous if it consists of groups that satisfy the following condition: if i and i'' are assigned to the same class then i' must be also assigned to that class. For ordered data only contiguous partitions need to be considered to detect the optimal one and Fisher demonstrates that least square partitions, i.e., partitions minimizing the within-group sum of squares, are contiguous. He deals with two subclasses of problems: the *unrestricted* case when the observations are ordered according to their numerical values, and the *restricted* one when an *a priori* ordering is given. Thus the method is applicable to time series where ordering is provided by the time and observations are not exchangeable.

In this paper we propose a non parametric approach for identifying the breakdates that exploits the idea of contiguous partitions within the framework of regression trees providing results comparable to those of Bai & Perron (2003). In particular, our procedure makes use of an artificial covariate and we call it Atheorethical Regression Trees (so forth denoted by ART) 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 a covariate in least square regression trees (Morgan & Sonquist, 1963; Breiman *et al.*, 1984) resorts to a recursive application of Fisher's method to a problem with 2 subperiods producing an open nested partition (Boros, 1996), i.e. a hierarchical structure in the form of a binary tree whose split points correspond to candidate breakdates. The final set of breakdates and corresponding homogeneous subperiods can result either from an automatic procedure, such as pruning, or from the subjective choice of the applied scientist based on *a priori* knowledge. Thus the procedure is data driven since the number of breaks and the times of occurrence are unknown and they are uncovered by the procedure itself.

The remainder of paper is organized as follows. Section 2 describes the model considered and the estimation procedure proposed by Bai & Perron (1998, 2003) based on the Fisher's method of exact optimization. In section 3 we introduce Least Square Regression Trees (LSRT) and we explain how they can be used to define contiguous partitions. Section 4 presents an illustrative application to the changes in the mean water levels of Lake Michigan-Huron. Our procedure is compared to Bai & Perron's method in Section 5 and brief concluding remarks follow in Section 6.

2. Model and estimation procedure

The model considered is:

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

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 a pure structural breaks model because all the model coefficients are subject to change and it has been employed by Bai & Perron (2003) to detect abrupt structural changes in the mean occurring at unknown dates. The problem resorts 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 homogeneous intervals such that $\mu_g \neq \mu_{g+1}$. Bai & Perron (2003) propose an estimation method based on the least squares principle: for each G-partition, the corresponding least square estimates of the μ_g's are obtained by minimizing the within-group sum of squares

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

The estimated breakdates $(\hat{T}_1, \ldots, \hat{T}_g, \ldots, \hat{T}_{G-1})$ are associated with the partition $P^*(G)$ such that $P^*(G) = \arg\min_{P(G)} WSS_{y|P(G)}$. 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's (1958) to find the least squares partition of T *contiguous* objects into G groups. His efficient algorithm exploits the additivity of the sum of squares criterion resorting to a dynamic programming approach (Bellmann & Dreyfus, 1962) that applied to ordered data points finds the global minimum. Despite the computational saving, the method cannot deal with high values of T and G and the same remark holds for the Bai & Perron's procedure, even with today's computing power.

We have found that the contiguity property can be used within the framework of LSRT to locate the breakdates in a faster way, sensibly increasing the values of T and G that can be considered, as it will be shown in the next section.

3. Atheoretical regression trees

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 the response variable Y given values of the predictor variables X. The choice of the mean squared error $E(Y - f(X))^2$ as error function of the predictor $f(X)$ leads to LSRT in which $f(X)$ is the conditional expectation $E(Y|X = x)$. 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 a complete discussion on this issue see Breiman *et al.*, 1984 Chap.9).
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

$$SS(h) - [SS(h_l) + SS(h_r)] \tag{3}$$

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 and $SS(h)$ can be thought of as the sum of squares at node h. The splitting criterion consists of minimizing, over all binary partitions of h, the within-group sum of squares:

$$WSS_{y|s}(h) = [SS(h_l) + SS(h_r)]. \tag{4}$$

Thus, the splitting criterion is based on the same objective function as equation (2).

For the purpose of locating the breaksdates, 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 a covariate in least square regression trees leads to creating and checking at any node h, all the $n(h) - 1$ possible binary contiguous partitions of the $y_i \in h$. Thus, the covariate is not to be considered as a predictor variable but rather as an auxiliary variable, a counter, that

allows the tree algorithm to try all admissible splits. In this respect, the procedure is theory-free and for this reason we call it Atheorethical Regression Trees (ART).

The contiguity property ensures that for any node h the best split lies in $\{k\}$ (or in its subintervals after the split of the root has taken place) and the tree algorithm, based on splitting criterion (3), is forced to identify it separating segments with different means. Note that in the original Fisher's method, as well as in the Bai & Perron procedure, optimal partitions for different values of G need not be hierarchically nested. ART method is based on a binary search algorithm and as splitting goes on, the previous partitions are fixed. Thus, after several splits there is no guarantee that the global optimum is reached i.e., that the absolute minimum within-group sum of squares partition is generated. 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. model (1) becomes a piecewise constant model with $\epsilon_t = 0$. Then, there is a partition into G segments for which the within-group sum of squares is zero and it will be identified by a sequential splitting algorithm as the one in ART.

In other words, if the data have a hierarchical structure then ART will find the overall optimum, otherwise it provides a suboptimal solution for which, because the partitions are contiguous, 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.

Given that 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, suboptimality does not appear a high price to pay to obtain full feasibility and indeed, in the application we will show that the partitions provided by ART are comparable to those obtained by the global search procedure.

Another relevant issue is the selection of the final partition and the consequent set of break dates. Indeed, partitioning methods such as Fisher's (and Bai & Perron'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 ART this is not a concern because it produces a hierarchy and the selection of the final set of break-

dates can be handled within the framework of tree methods by means of pruning that is the process of retrospectively discarding branches whose contribution to the reduction of the error is negligible (for details see Breiman *et al.*, 1984, Chap. 3). In this way a nested sequence of partitions and candidate breakdates is created; in order to select the optimal one corresponding to the actual number of break dates and distinct subperiods present in the data, cross validation (CV), the sequential testing procedure of Bai & Perron (1998) and model selection criteria can be employed (for details on the use of model selection criteria in regression trees see Su, Wang & Fan, 2003). Moreover, the inspection of the tree diagram allows an insight into the partitioning process and also provides an ordering of the breakdates based on their position in the tree and the reduction of the error function achieved, thus, manual pruning based on subjective choices of the analyst can be preferred to automatic procedure as pointed out in Zhang & Singer (1998, Chap. 4), where an alternative pruning method is also discussed.

Finally, note that if estimation is not the sole concern and one want 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 Bai & Perron (2003). 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.

4. Application to Lake Michigan-Huron water levels

As an illustrative application we apply the ART procedure to analyze the changes in water levels of the Lake Michigan-Huron, one of the largest bodies of water worldwide. Parts of the shorelines are urbanized and industrialized, others are intensively farmed and some others are wild and represent the ideal habitat of many rare plants and animals. All these reasons explain the importance of Lake Michigan-Huron and the attention it receives from governmental agencies of US and Canada, the scientific community and the media. The same application was considered by Fisher (1958).

The data-set reports the lake yearly mean water levels over a time span of 140 years running from 1860 (the first year of systematic water level recording) to 1999. The series has been partitioned by ART, setting a small minimum segment size (5 years) in such a way that if short intervals of extreme high or low water levels are present in the series, they are revealed and isolated by the procedure. Figure 1 depicts the resulting maximal tree of 10 terminal nodes in which the breakdates can be retrieved simply by adding the split point to the initial year of the series.

The tree may suggest to a scientist that one partition is preferable to

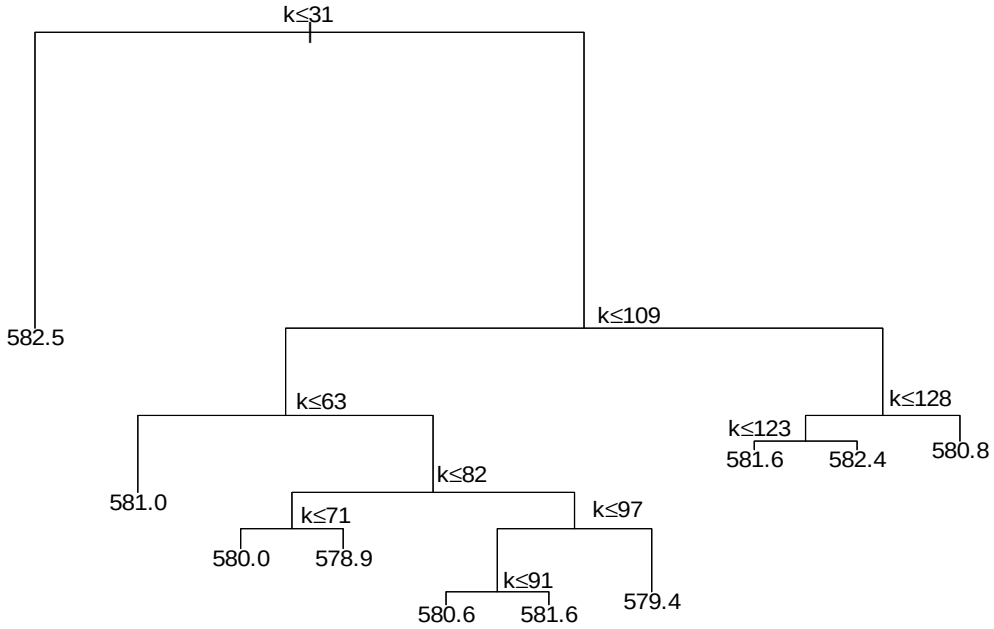

Figure 1: Large initial tree grown by ART. Values above the internal nodes report the split points, values beneath the terminal nodes indicate the mean water levels (feet) of the corresponding subperiods.

another. Here, we have considered the CART (Breiman *et al.*, 1984) pruning method to create a sequence of subtrees and then, in order to select the final set of break dates identifying periods of fairly constant mean water levels, we have applied CV, and model selection criteria. Concerning the latter, since none of them can be unconditionally recommended as a default procedure, here we consider the most popular ones, namely AIC (Akaike, 1970), AICC (Hurvich & Tsai, 1989) and BIC (Schwarz, 1978); note that the BIC has been proposed by Yao (1988) for selecting the number of change-points and it is also suggested by Bai & Perron (2003) based on a previous simulation study (Perron, 1997).

Both AIC and AICC select as final partition the maximal tree whereas CV and BIC select the 8-segment partition. Figure 2 depicts the optimal 8-segment partition and table 1 reports the corresponding 95% confidence intervals computed using the asymptotic distribution derived in Bai and Perron (1998) and assuming constant error variance across the segments .

Confidence intervals covering small periods, such as those for break dates 31, 82, 97 and 109 indicate that the breakdates are precisely estimated, whereas overlapping confidence intervals suggest that the break dates might

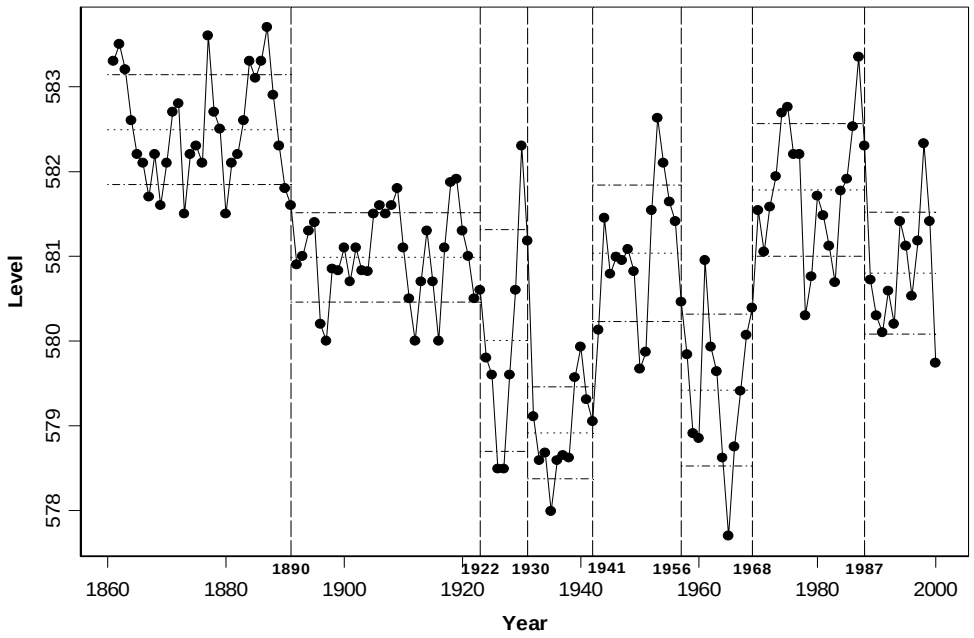

Figure 2: Final partitions for $G = 8$. The horizontal lines indicate μ and $\mu \pm \sigma$.

G	Confidence intervals						
8	31	63	71	82	97	109	128
	28 : 34	57 : 69	66 : 76	80 : 84	94 : 100	108 : 110	122 : 134

Table 1: 95% Confidence intervals for breakpoints of optimal 8 segment partitions.

not correspond to distinct subperiods. This is the case of confidence intervals for breakpoints 63 and 71. Indeed, these breakdates isolate short intervals of unstable low water levels that can be connected to the particularly dry hot years in the 1930's and to dredging in the St. Clair-Detroit system, the main outflow for Lake Michigan-Huron. Given the hierarchical structure of ART, it easy to identify from Figure 1 which breakdate is to be discarded, i.e. breakdate 71.

The entire segmented series shows that Lake Michigan-Huron has experienced a long period of stable high water levels till the last decade of the 19th century when they dropped to lower but still stable levels lasting for more than thirty years. Starting from the mid 1920's lake levels have began to alternate irregularly between shorter periods of lower and higher levels.

G	$SS(P_G)$	P_G								
2	155.76	31								
3	131.76	31	109							
4	110.14	30	63	109						
5	100.25	30	63	82	109					
6	82.53	30	63	83	97	108				
7	75.48	30	63	83	97	109	128			
8	70.31	30	63	71	82	97	109	128		
9	66.13	30	63	71	82	91	96	109	128	
10	63.69	30	62	67	72	82	91	96	109	128

Table 2: Partitions with global search procedure.

Extreme low water levels occurred in the mid 1960's with the record low in 1964, likely in response to a further navigation channel dredged in the St. Clair River. A long period of sustained high levels has been recorded during the 1970's and most of the 1980's with the record high level of 1986.

As a matter of fact, there are concerns that global warming and the so called greenhouse effect might cause an irreversible lowering of lake water levels. Our study gives an insight into the phenomenon showing that up to now no long term lowering is apparent and that lake mean water levels are rather characterized by a high instability over the years, alternating between higher and lower phases.
Concerns for the future are certainly justified based on projected scenarios of climate shift (see for example Morsch and Quinn, 1996; Chao, 1999) towards warmer temperatures and decreased precipitations.

5. Comparison with the global search procedure

In order to compare our procedure with the global minimizer of the sum of squares, we have applied to the Lake Michigan-Huron data the procedure of Bai & Perron as implemented in the R package `strucchange` (Zeileis *et al.*, 2002). Table 2 gives the partitions for G ranging from 2 to 10 with the indication of the corresponding within-group sum of squares. The partitions are obtained setting the same minimum segment size as in the application of ART.

As expected these partitions are not completely nested, but, comparing the breakpoints identifying the partitions with those in the tree grown by ART (see Figure 1) we see that the tree contains the same partitions with most of the breakpoints being identical and single misplacements occurring

on the boundaries. These misplacements cause the within-group sum of squares to be higher in the case of ART, indeed, for $G = 8$ the within sum of squares of the ART partitions is 70.78, thus, the loss is not remarkable. The only partition that is not in the tree is for $G = 10$, in fact, the nested nature of ART partitions, together with the constraint on the segment length makes it more difficult to have a further break between observations 63 and 71 as in the global search procedure. For this reason the global search algorithm tends to create a larger number of candidate partitions with respect to ART, highly sensitive to the chosen minimum segment size.

A further aspect to be compared concerns the computing cost associated with the two procedures. On an Intel Pentium 4 1.8GHz the CPU time of executing the global search for the Huron-Michigan analysis is 6.11s against 0.03s of ART, i.e. comparable results can be obtained in a much shorter time. Indeed, when it comes to long series, the global search procedure is computationally demanding and may be unfeasible unless the minimal length of the segments is set to a value that strongly reduces the number of admissible partitions to be enumerated and evaluated.

6. Concluding remarks

We have proposed a data driven nonparametric procedure for detecting multiple structural breaks in the mean occurring at unknown dates. The method, called ART, exploits the concept of contiguous partitions introduced by Fisher (1958) within the framework of least squares regression trees resorting to a sequential use of his algorithm of exact optimization.

Although our procedure does not find the global minimum, its results are comparable to those obtainable by applying Bai and Perron's procedure, coinciding for most of the breakdates. The main advantages of the proposed approach are:

1. *simplicity* - it can be easily implemented or run with packages containing routines to grow and prune least square regression trees;

2. *feasibility* - it can be used to find the least squares partition of an ordered sequence with no limitations on either the number of observations or the number of segments;

3. *visualization*- it results in a nested hierarchy represented by a tree diagram that displays the whole partitioning process and allows the scientist to interact and make use of *a priori* knowledge.

In this paper the method has been applied to explore changes in lake mean water levels that are known to cause social, economic and ecosystem dis-

ruption. In particular, declining lake levels are of concern to the Great lakes community but the application to Lake Michigan Huron data has shown that, although water levels have varied considerably, no long-term lowering is yet apparent .
Current research is addressing the problem of modifying the splitting criterion in ART to account for breaks-dates in models with trending regressors.

References

Akaike, H. (1970). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, eds. B.N. Petrov & F. Csaki F., Budapest: Akademia Kadio, 267–281.

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

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

Bellmann, R.E. & Dreyfus, S.E. (1962). *Applied Dynamic programming*. Princeton: Princeton University Press.

Boros, E. & Hwang, F.K. (1996). On the optimality of nested partitions. *SIAM Journal on Optimization* **6**, 1153–1162.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*. Monterey (CA): Wadsworth & Brooks.

Chao, P. (1999). Great Lakes water resources: climate change impact analysis with transient GCM scenarios. *Journal of American Water Resource* **35**, 1499–1508.

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**, 789–798.

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

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

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hurvich, C. & Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrics* **76**, 297–307.

Morgan, J.N. & Sonquist, J.A. (1963). Problems in the analysis of survey data and a proposal. *Journal of American Statistical Association* **58**, 415–434.

Mortsch, L.D. & Quinn, F.H. (1996). Climate change scenarios for Great Lakes Basin ecosystem studies. *Limnology and Oceanography* **41**, 903–911.

Perron, P. (1997). L'estimation de modèles avec changements structurels multiples. *Actualité Économique* **73**, 457–505.

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

Su, X.G., Wang, M. & Fan, J.J. (2004). Maximum Likelihood Regression Trees. *Journal of Computational and Graphical Statistics* **13**, 586–598.

Yao, Y.C. (1988). Estimating the number of change-points via Schwarz's criterion. *Statistics and Probability Letters* **6**, 181–189.

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

Zhang, H. & Singer, B. (1998). *Recursive partitioning in the health science*. New York: Springer-Verlag.