Chemometrics and Intelligent Laboratory Systems 146 (2015) 337–346

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Regime change analysis of interval-valued time series with an application to PM10

Carmela Cappelli [a,*], Pierpaolo D'Urso [b], Francesca Di Iorio [a]

[a] Università Federico II di Napoli, Napoli, Italy
[b] Sapienza Università di Roma, Rome, Italy

ARTICLE INFO

Article history:
Received 21 January 2015
Received in revised form 10 June 2015
Accepted 12 June 2015
Available online 19 June 2015

Keywords:
Interval-valued time series
Change point analysis
Atheoretical regression trees (ART)
Urban air pollution
PM10 time series

ABSTRACT

The aim of this paper is to conduct change point analysis of interval-valued time series employing a regression trees approach. In order to deal with such time series we propose to employ a suitable distance measure that takes into account the underlying structure of interval data. Simulation results pertaining to the behavior of the proposed approach as well as an empirical application on a daily sample of air pollutant are provided, that illustrate the practical usefulness of the proposed method.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the literature, several statistical methods have been proposed for analyzing interval-valued data in different empirical fields, such as chemometrics [12–14,25], ecotoxicology [2], meteorology [24,32], medicine [35], physics [22], pattern recognition [18], telecommunications [15], health and retirement [34], economics and finance [20]. As remarked by Manski and Tamer "researchers often have only interval data on variables that can, in principle, be measured more precisely" [34]. For instance, "the interval data on wealth in the Health and Retirement Study (HRS) provide a ready illustration [30]. Let v denote a person's wealth. Under the HRS questionnaire protocol, a respondent is asked to report v. If he does not comply, the respondent is then asked to report if wealth falls within a sequence of brackets. The HRS thus yields a wealth interval $[v_0, v_1]$ for each respondent. The interval is degenerate when a respondent provides a point value of wealth, is an informative interval of positive width when the respondent answers the subsequent bracket questions, and is the uninformative interval $-\infty, +\infty$ otherwise". Notice that, in general, interval-valued data are vague and imprecise data conversely to single-valued data that are precise data. Consequently, interval-valued data are considered as a loss of information. For example, the sentence "the temperature is 14" is more precise and more informative than "the temperature is more or less between 12 and 15". However, there are real cases in which the use of single-valued data (numerical data) may bring about a heavy loss of information. In fact, it is more informative to consider the interval between the minimum and the maximum daily temperature than the average value or the central value of the daily temperature. From a methodological point of view, interesting contributions have been suggested in the main fields of statistics as time series analysis [3,38], cluster analysis [13,16], principal component analysis [12,18,31], multidimensional scaling [17,22], regression analysis [11,20,34], decision trees [32], self-organizing maps [2,15,24], neural networks [9].

In particular, focusing on time series analysis, there are several experimental situations and real applications in which the empirical information is represented by interval-based observations. For example in meteorology we might want to analyze a time series of daily temperatures or air pollutant concentrations, in finance the daily volatility of an asset, in medicine the daily systolic or diastolic pressure of a patience. In cases such as these it is more interesting to take in account of the interval-valued structure of the data considering the minimum and maximum values recorded over the period of interest, or the mean value and the deviation from the extremes, because, for a given time period, they contain more information about the phenomenon of interest than a single-point valued observation which summarise the information in the same time period. For this reason, in recent years, the analysis of interval valued time series (IVTS, hereafter) has attracted the attention of the statistical and econometric communities (see for example Refs. [1,3,37,38]). Indeed, over the last two decades change point detection has emerged as a relevant research topic in several fields of research. In the context of time series analysis the detection

* Corresponding author at: Dipartimento di Scienze Politiche, Via L. Rodinò n. 22, 80138 Italy.
E-mail addresses: carcappe@unina.it (C. Cappelli), pierpaolo.durso@uniroma1.it (P. D'Urso), fdiiorio@unina.it (F. Di Iorio).

http://dx.doi.org/10.1016/j.chemolab.2015.06.006
0169-7439/© 2015 Elsevier B.V. All rights reserved.

of changes is useful from several points of view. First it can reveal a behavior of the time series that could otherwise be misunderstood and modeled inadequately; second, in case of long time series, a model estimated on a recent segment of the series might provide more accurate forecasts, eventually the identification of breaks might reveal the presence of outliers (see Ref. [8]). The undiscussed contribution in the field of dating multiple changes occurring at unknown dates is due to Bai and Perron (see among others Refs. [4,5]). In case of multiple changes in mean Cappelli et al. [8] have proposed a method called atheoretical regression trees (ART) that employs least square regression trees (LSRT) to estimate the number and location of changes whereas, recently, an extension based on a fuzzy approach that deals with imprecise, vaguely observed time series has been developed by Cappelli et al. [7]. In chemometrics and environmetrics change point analysis is a useful tool for monitoring and control; in this context an interesting contribution has been proposed by Jansen [28]. In particular, motivated by the application of single molecule detection in a highly dilute solution, Jansen discussed the problem of multiple change point detection in the intensity curve of low-intensive Poisson observations. He explained that the multiple change point detection problem is inherently a multiscale problem and analyzed the data using an extension of the continuous wavelet transform (CWT) (the unbalanced wavelet transform). In this way, the presence of change points in the underlying intensity curve is revealed by a multiscale chain of local maxima in the CWT analysis. Then, he proposed an algorithm for the reconstruction of the chains by linking local maxima across scales; the algorithm is crucial in detecting small changes against intensive noise. Other interesting contributions have been proposed in the literature. Jaurskova [29] proposed change point detection methods for hydrological and meteorological time series. Jandhyala et al. [26] developed methods for detection and estimation of unknown change-point in Weibull parameters and applied the proposed methods to daily minimum temperatures time series. Wu et al. [39] pursued a change-point analysis of rainfall data and global warming data by applying isotonic regression based change-point methodology, as well as a suitable trend detection test for stationary time series. Lund et al. [33] proposed a change-point detection analysis for periodic and autocorrelated time series with application to precipitation data. Dierckx and Teugels [10] suggested a change point analysis of extreme values with applications to earthquake and climatological (i.e. precipitations and temperatures) time series. Jandhyala et al. [27] applied change-point methods for mean annual rainfall time series. Gallagher et al. [19] suggested a change point detection analysis for daily precipitation data. In this paper in order to deal with interval-valued time series we propose to employ, in the framework of ART, a deviation measure based on a suitable squared distance measure [15] that combines the midpoint (centers) and width (radius) of the intervals associated with the interval-valued time varying units. Our methodological approach has important advantages. Firstly, with respect to the other methodological approaches proposed in the literature in environmetrics and chemometrics literature, our change point detection method is able to analyze interval time series by considering all the values contained in the minimum-maximum range of a (e.g. daily) time series conversely to the previous methods which analyze only significant values (e.g. mean or extreme value) of a (e.g. daily) time series. Furthermore our method since it is based on a least-square regression trees approach (in fact is called atheoretical regression trees) is model-free and then it is not conditioned by theoretical restrictions connected to possible distribution assumptions.

The remainder of the paper is organized as follows: in Section 2 we introduce the notion of interval valued data and we illustrate the ART method showing how it can be employed to detect change points in interval-valued time series. In Section 3 we present the results of simulation studies pertaining to the behavior of the proposed approach whereas an empirical application to change point detection in a data set of environmental variables is discussed in Section 4. Concluding remarks are drawn in Section 5.

2. Detecting change points in interval-valued time series

In this section we introduce the notion of interval-valued time series and we briefly describe the issue of change point detection and regression trees showing how these can be employed to detect change points in IVTS.

2.1. Interval-valued time series

An interval-valued time series can be formalized as $\tilde{Y}_t = [l_t, u_t]$, $t = 1, \ldots, T$, such that each time unit has an interval structure characterized by a lower and an upper bound l_t and u_t, respectively.

Each observation can be represented by a vertical segment defined by the *center* (midpoint) $c_t = \frac{l_t+u_t}{2}$ and *radius* (spread) $r_t = \frac{u_t-l_t}{2}$ of the interval and thus the IVTS can be reformalized as

$$\tilde{Y}_t = [c_t, r_t], \qquad t = 1, \ldots, T.$$

Note that the center-radius representation is simple and convenient because the range is a common measure of variability of a random variable and it is often employed for estimation purposes in various empirical applications.

In order to compare two interval-valued time units the following Euclidean distance can be employed [15]:

$$d(t, t') = \sqrt{(c_t - c_{t'})^2 + (r_t - r_{t'})^2}. \tag{1}$$

This distance measure combines the information on both the center and width of the interval associated to each interval-valued time unit and it satisfies the usual properties of distance measures.

Notice that, other possible distance measures could be considered for comparing interval-valued time units. See, for instance, Refs. [12, 21,23,25].

2.2. Change points detection and regression trees

Consider a standard point-valued time series Y_t characterized by $m + 1$ regimes and m changes so that $t = T_{j-1} + 1, \ldots, T_j$ and $j = 1, \ldots, m + 1$ with $T_0 = 0$ and $T_{m+1} = T$. In order to estimate the set of unknown change points $(T_1, \ldots, T_m)$ the least square principle is used and the estimated break points $(\hat{T}_1, \ldots, \hat{T}_m)$ are such that:

$$\left(\hat{T}_1, \ldots, \hat{T}_m\right) = \underset{(T_1, \ldots, T_m)}{\arg\min}\, SSR(T_1, \ldots, T_m) \tag{2}$$

where $SSR(T_1, \ldots, T_m)$ denotes the sum of squares residuals of the $m + 1$ partition that in case of changes in means is: $SSR(T_1, \ldots, T_m) = \sum_{j=1}^{m+1} \sum_{t=T_{(j-1)}+1}^{T_j} (Y_t - \mu_j)^2$..

Cappelli et al. [8] have proposed a procedure that detects such change points employing LSRT. In general, given a continuous response variable Y and a set of predictors $X_1, \ldots, X_p$, regression trees model the relationship between the response and the covariates using a recursive partitioning approach that results into a partition of Y based upon the values of the predictor variables. In particular, LSRT are piecewise-constant models: a node h i.e. a subsample of statistical units is split into its left and right descendants h_l and h_r to reduce the deviance of the observed dependent variable Y fitting to each node the mean of corresponding Y values. The algorithm selects the split, i.e. the binary division, that minimizes sum of squared residuals:

$$SSR(h_l) + SSR(h_r) = \sum_{i \in \{l,r\}} \sum_{Y \in h_i} \left(Y - \hat{\mu}(h_i)\right)^2 \tag{3}$$

where $\hat{\mu}(h_i)$ is the mean of the observed y values in node $h(i)$ $(i \in \{l, r\})$ thus, the splitting criterion (3) corresponds to the (2) computed for a

binary partition. Once a node is partitioned, the splitting process is recursively applied to each subnode until either they reach a minimum size or no improvement of the criterion (3) can be achieved.

Indeed, LSRT are a practical tool for dating multiple changes in mean occurring at unknown dates in a time series Y_t. At this aim a single artificial covariate is employed in the recursive partition process i.e. a sequence of completely ordered numbers $k = 1, \ldots, T$. Tree regressing the time series Y_t on k provides a partition of the series into homogeneous segments such that $\hat{\mu}_j \neq \hat{\mu}_{j+1}$; the partition is represented as a binary tree whose split points identify candidate change points. Note that a binary tree is a hierarchical structure i.e. a nested sequence of partitions and since the tree tends to be overly large a pruning method is employed to trim it back. Pruning is the process of discarding terminal nodes whose contribute to the reduction in deviance is negligible and it generates a sequence of subtrees that represent nested change point models. In order to select the subtree whose terminal nodes provide the optimal partition corresponding to the actual number of changes and distinct subperiods present in the data, we use classical model selection criteria. Extensive simulation studies, comparison with current methods and applications to various real time series have provided evidence of the usefulness of the approach (see Ref. [36]) that has been called atheoretical regression tree (ART) because it employs a single artificial covariate and no parametric model is assumed and estimated in the subperiods.

The squared distance measure (1) can be used in the framework of ART to detect change points in IVTS.

At this aim we define the deviation over the entire sample period

$$SS\left(\tilde{Y}_t\right) = \sum_{t=1}^{T}\left[(c_t - \bar{c})^2 + (r_t - \bar{r})^2\right]$$

where $\bar{c}$ and $\bar{r}$ are the mean values of the centers and radii, respectively.

Then, in ART's recursive portioning approach, the best split of a generic node h, that identifies a candidate change point, minimizes

$$SSR(h_l) + SSR(h_r) = \sum_{i\in\{l,r\}}\sum_{t=1}^{T(h_i)}[(c_t - \bar{c}(h_i))^2 + (r_t - \bar{r}(h_i))^2] \quad (4)$$

where $\bar{c}(h_i)$ and $\bar{r}(h_i)$ are the mean values of the centers and radii in node h_i ($i \in \{l, r\}$) and $T(h_i)$ is the length of the corresponding subseries.

After a large tree is grown we employ classical cost-complexity pruning [6] to generate the sequence of subtrees corresponding to alternative nested change point models of various dimension (number of changes and regimes). Then, to select the preferable subtree (model) among the competing ones, we consider an information criterion; in particular we use a modified Bayesian Information Criterion (*BIC*) defined as

$$BIC(m) = \ln \hat{\sigma}^2(m) + p \ln(T)/T$$

where $\hat{\sigma}^2(m) = T^{-1} SSR_{\tilde{Y}_t}(\hat{T}_1, \ldots, \hat{T}_m)$ is the sum of squared residuals of the m-partition of the IVTS and $p = (m + 1) \times (k + 1)$ with $k = 2$ because in each regime 2 parameters are estimated i.e. the mean value of the centers and radii, respectively.

3. Simulation experiments

In this section we present the results of simulation experiments carried out to evaluate the proposed approach considering as performance indicators the number of structural change points detected (cp) and the rate of correct identification of the change points (ci) (either exact identifications or short intervals around the true value). Three basic scenarios have been analyzed:

1. changes only in the centers;
2. changes only in the radii;
3. changes in both centers and radii.

The data generating process (DGP) of each interval-valued time unit is: $Y_t \sim N(\mu_t, \sigma_t^2)$. Throughout the simulations, it is $m = 2$ thus two change points occur in the data at times $T_1 = 80$ and $T_2 = 150$, the length of the series is $T = 200$ and 1000 Monte Carlo replications are generated.

For the third scenario, i.e. changes in both centers and radii, a further simulation experiments has been carried on considering a shorter series ($T = 130$) with $m = 4$ changes not equally spaced.

3.1. Scenario I

We start with the case where changes occur only in the centers and thus their identification should not be affected by the presence of the radii. Indeed this is a base case to assess the behavior of the method that we expect to perform as standard ART for single-valued data providing similar results.

The parameter values of the Normal distributions in the different segments

$$\begin{cases} Y_t \sim N(0,1) & if \quad t \leq 100 \\ Y_t \sim N(0.5,1) & if \quad 100 < t \leq 200 \\ Y_t \sim N(1,1) & if \quad t > 200 \end{cases}.$$

For illustrative purposes one of the simulated series (centers and radii) and the corresponding error bar series are plotted in Fig. 1. Note that since the change in the centers is quite small and the radii are not subject to change, on the whole the plot of the error bar does not suggest the presence of changes. It is also worth noticing that increasing steps are difficult cases where most procedures fail to select both the right number of changes and their location.

We have applied ART employing the deviation measure defined in the (4) and to avoid the identification of changes in the tails, we have set a minimum segment length of 15 observations.

Table 1 reports the results averaged over the 1000 Monte Carlo replications. The method identifies the right number of change point ($cp = 2.03$) and only occasionally detects an additional spurious break. The rates of exact correct identifications (ci) are pretty high for both changes and they became notably high for larger intervals around the true date, despite the fact that the changes are quite small.

Eventually, since in this case only the centers are subject to change, for comparison purposes we have applied the standard ART with deviation measure (3) to the series of the centers, and found very similar results (see Table 2). In other words, although the data could be treated as point-valued, the use of a deviation measure for IVTS does not affect the analysis and the standard method does not outperform the proposed approach.

3.2. Scenario II

In the second simulation experiments only the radii are subject to change. This is a case particularly relevant because, when no changes occur in the midpoints, any standard method (classical ART or Bai and Perron's) that does not take into account the structure of the data, applied to single-valued time series would detect no changes. In order to render this situation we have simulated the time unit from non-homoschedastic normal distributions with constant mean. The DGP are:

$$\begin{cases} Y_t \sim N(0,1) & if \quad t \leq 100 \\ Y_t \sim N(0,1.5) & if \quad 100 < t \leq 200 \\ Y_t \sim N(0,2) & if \quad t > 200 \end{cases}.$$

Again, the error bar plot of one simulated series (Fig. 2) shows a very slight evidence of changes. Despite the weak graphical evidence, the

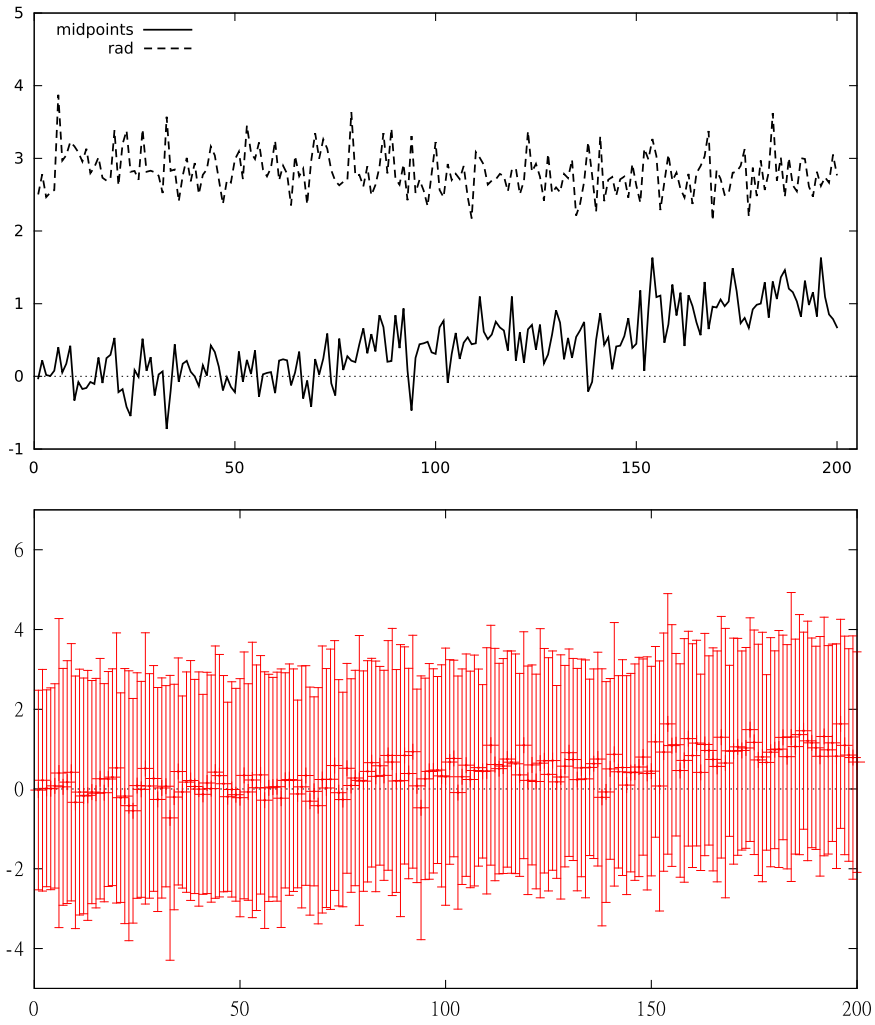

Fig. 1. Changes only in the centers, one simulated series.

method is effective in detecting the presence of changes in the radii as well as their number and location, as we can conclude from Table 3 that reports the results averaged over the 1000 MC replications.

Note that the rates of correct identifications are slightly lower with respect to the previous case (changes only in the centers), this is due to the fact that the changes in the radii have been rendered by the variance of the DGP of the time units and in terms of square roots, they are extremely mild.

Table 1
Changes only in the centers—simulation results averaged over the 1000 MC replications.

cp = 2.03			
Break 1 ($T_1 = 80$)			
ci	$ci \pm 1$	$ci \pm 2$	$ci \pm 4$
.55	.76	.85	.93
Break 2 ($T_2 = 150$)			
ci	$ci \pm 1$	$ci \pm 2$	*ci*
.56	.80	.87	.95

Table 2
Standard ART applied to the center series—simulation results averaged over the 1000 MC replications.

cp = 2.08			
ci	$ci \pm 1$	$ci \pm 2$	$ci \pm 4$
Break 1 ($T_1 = 80$)			
.53	.74	.83	.91
Break 2 ($T_2 = 150$)			
.56	.78	.86	.95

3.3. Scenario III

Eventually we have analyzed the case where both centers and radii are subject to changes.

In the first setting the means of the normal distributions of the centers have been set as in case study I whereas the changes in radii have been generated using the variances of case study II. One simulated series is depicted in Fig. 3.

We see that the presence of changes both in the centers and radii, although not very strong, leads to a graphical evidence of two changes of increasing magnitude.

Table 4 reports the results of the simulation experiment.

Not surprisingly in presence of changes in both centers and radii the proposed method achieves the highest rates of correct identification.

It's worth reminding that the procedure is based on a recursive binary partitioning algorithm thus, once a split is performed and the corresponding change point is identified, the search is repeated separately on the subsegments defined by the change point. For this reason, as it will be shown by the next experiment, the method is able to handle the case of multiple (more than two) change points.

In order to provide more evidence of the effectiveness of the proposed approach we have carried on a further simulation

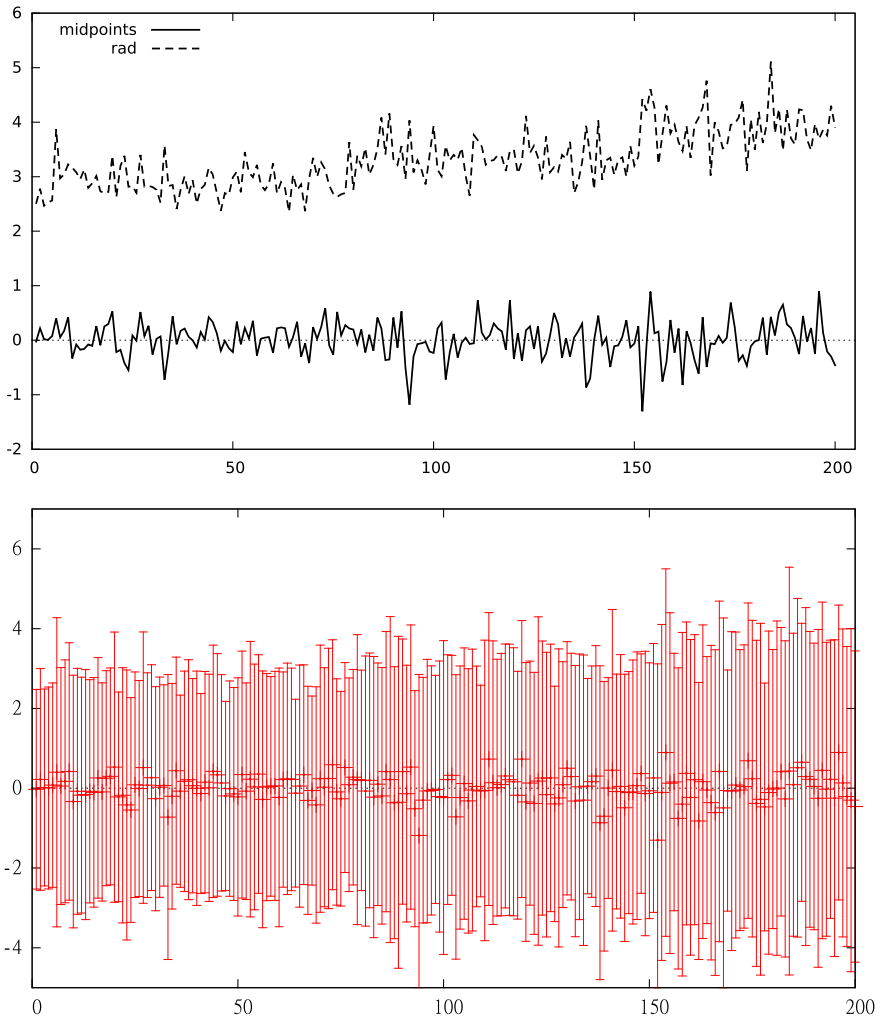

Fig. 2. Changes only in the radii, one simulated series.

experiment considering a more complex setting. In particular a short time series ($T = 130$) has been generated that contains four changes both in the centers and radii, not equally spaced. The DGP of the subseries are:

$$\begin{cases} Y_t \sim N(0,1) & \text{if} \quad t \le 25 \\ Y_t \sim N(0.3,0.8) & \text{if} \quad 25 < t \le 60 \\ Y_t \sim N(0.7,1.1) & \text{if} \quad 60 < t \le 80 \\ Y_t \sim N(0.5,1.3) & \text{if} \quad 80 < t \le 110 \\ Y_t \sim N(0.7,1.1) & \text{if} \quad t > 110 \end{cases}.$$

As we can see the changes in the parameters of the normal distributions are extremely mild and their values fluctuate; as a consequence both the series of the centers and radii show an irregular behavior as we can see from Fig. 4. Note that since the length of the subseries is very short we have set a minimum segment length of five observations.

The results averaged over the 1000 MC replications (see Table 5) are consistent with the previous simulations. In particular, since the proposed approach is based on a deviance measure that combines centers and radii, this simulation confirms that in case of changes in both, the method provides the best performances with very high percentage of correct identification that reach almost 100% for intervals of ± 4 observations around the true date. Moreover the method is able to handle multiple (more than two) change points that are quite close to each other and unequally spaced. Eventually note that the mean number of change points detected (3.87) is slightly underestimated, this is likely due to the change at observation 80 that concerns a very short period (20 observations) and whose percentage of correct identifications are lower compared to the other changes.

Table 3
Changes only in the radii—simulation results averaged over the 1000 MC replications.

cp = 2.04			
ci	$ci \pm 1$	$ci \pm 2$	$ci \pm 4$
Break 1 ($T_1 = 80$)			
.43	.63	.78	.90
Break 2 ($T_2 = 150$)			
.37	.63	.75	.87

4. An air pollution application: The change point analysis of the PM10 time series in Rome

In the empirical application we have considered a data set of environmental variables collected in Rome in 1999 in the monitoring station installed in Via Arenula, close to the historical center of the city. We have focused on concentration of particulate matter 10 (PM10) that is among the most harmful of all air pollutants. In general particulate matter is the term for solid or liquid particles (such as dust, fume, mist, smog, smoke) found in the air that cause air pollution. It may vary greatly in color,

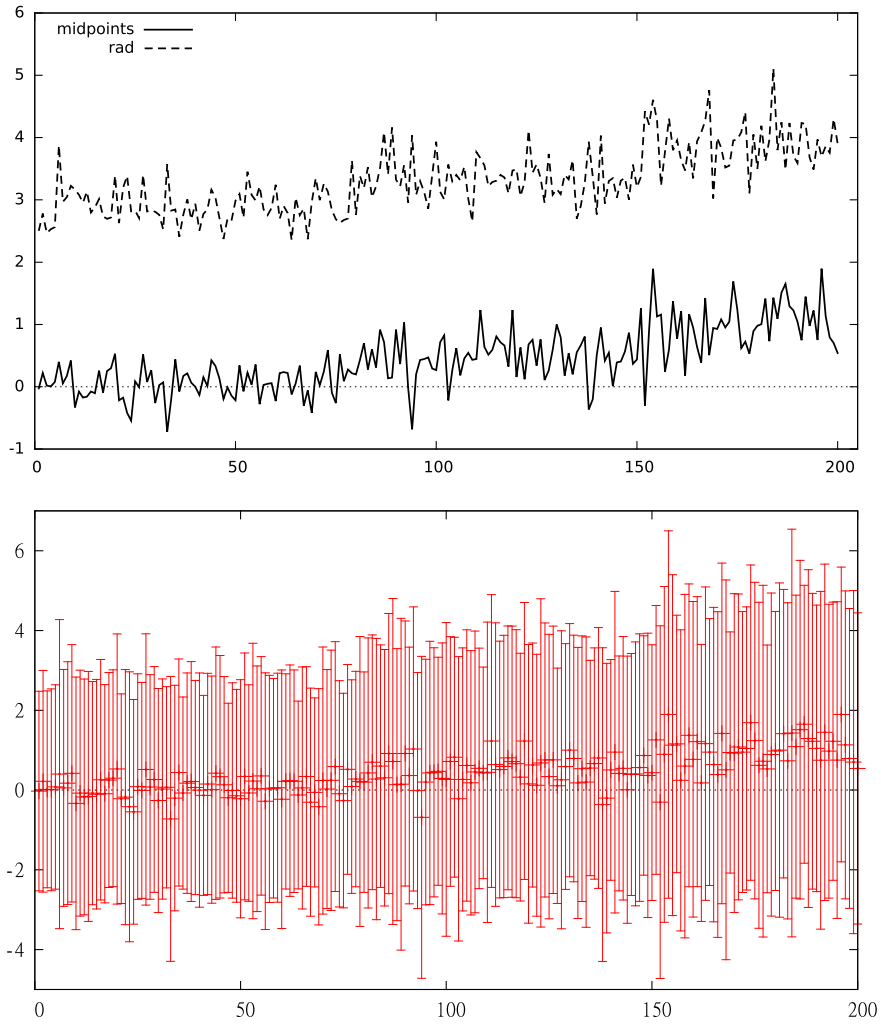

Fig. 3. Changes in both centers and radii, one simulated series.

density, size, shape, and electrical charge, from place to place and from time to time. Smaller particles are likely responsible for adverse health effects because of their ability to reach the lower areas of the respiratory tract. The PM10 standard includes particles, measured as thousands for cubic meter, with a diameter of 10 μm or less. According to health-based EU air quality standards the average daily concentration of PM10 should not exceed 50 $\mu g/m^3$.

We have analyzed the sample period 01 August–30 November and thus $T = 122$. The original time series provides daily values recorded on hourly basis and thus time units are represented by interval-valued data whose conversion into a single value for each day entails loss of information and inaccuracy. In Fig. 5 it is depicted the time series of the centers and radii whereas Fig. 6 reports the error bar graph corresponding to the interval-valued time units.

Table 4
Changes in centers and radii—simulation results averaged over the 1000 MC replications.

cp = 2.03			
ci	$ci \pm 1$	$ci \pm 2$	$ci \pm 4$
Break 1 ($T_1 = 80$)			
.58	.79	.88	.94
Break 2 ($T_2 = 150$)			
.59	.81	.89	.96

The visual inspection of both series in first place shows that the PM10 values often exceed the standards, moreover the plot suggests the presence of a possible change located at the beginning of October as well as of possible outliers (time units characterized by relatively larger midpoint and/or radius) located at the extremes.

We have applied the regression tree based procedure setting a minimum number of observations per segment of 7 days and we got a large tree with nine terminal nodes corresponding to eight candidate change points that are indicated in Fig. 6. Then we have generated a sequence of nested partitions by pruning back the tree and we have employed the proposed modified BIC to select the optimal number of changes.

In Fig. 7 it is displayed in the form of a tree diagram the optimal partition corresponding to the change points selected according to the modified BIC whereas Table 6 summarizes some stylized facts of the entire series and the five subperiods identified by the change points.

As expected the first and strongest change occurs around the middle of the series (observation 68, 7th of October), when the series of PM10 shifts towards higher levels of both centers and variability of the radii. In particular, in the first segment, running from the 1st of August to the 7th of October, the average daily concentration of PM10 is 47.9, and thus it does not exceed the European air quality standards, opposed to the second period where the mean daily concentration is 75.8. This is likely due to increasing vehicles circulation after the summer break. Two more changes occur at the beginning and the end of the series (observations 11 and 114, 11th of August and 22th of November, respectively) and they are both characterized by a sudden shift that identifies outliers.

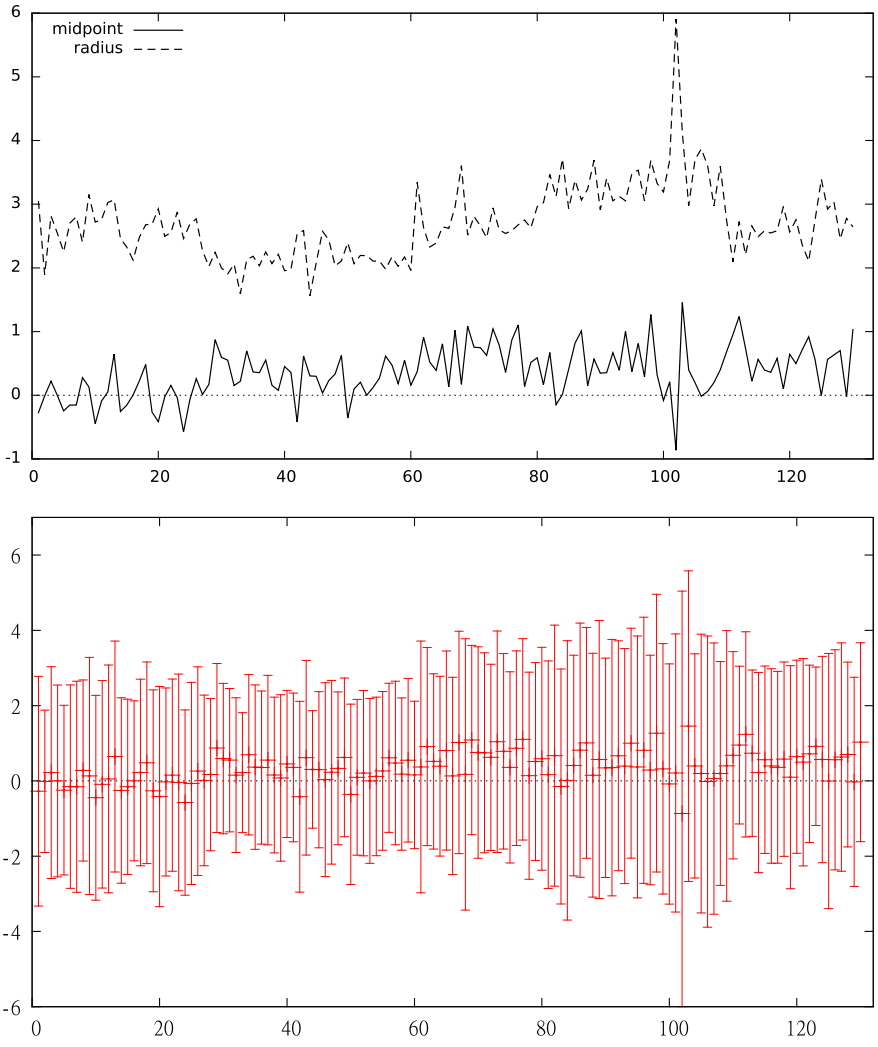

Fig. 4. Four changes not equally spaced in both centers and radii, one simulated series.

Moreover the change at time 11 separates the period (running from 12th of August to 7th of October) of lowest concentration of PM10 in terms of mean values (centers), variability and width of the radii as we can see from the mean and standard deviations associated to this subperiod (see Table 5). We may see this period as the most respectful one in terms of limit values. The last change occurs at observation 95 (4th of November), and it identifies two subperiods of both high mean levels and daily variability of PM10. A further insight into the partitioning process is provided by the graph of the values of the objective function (3) associated to each possible split of the tree nodes, reported in Fig. 8.

As we can see, all changes correspond to a strong minimum of the objective function; moreover the behavior of the plot for changes occurring at observations 11 and 114 confirms that they identify outliers. On the whole, the application findings confirm how the proposed approach is effective in detecting change points in IVTS separating periods that differ from each other.

Table 5
Four changes not equally spaced in centers and radii—simulation results averaged over the 1000 MC replications.

cp = 3.87			
ci	$ci \pm 1$	$ci \pm 2$	$ci \pm 4$
Break 1 ($T_1 = 25$)			
.70	.91	.95	.96
Break 2 ($T_2 = 60$)			
.83	.93	.96	.99
Break 3 ($T_3 = 80$)			
.47	.75	.84	.95
Break 4 ($T_4 = 110$)			
.76	.92	.97	.99

5. Conclusions

In several real life and research situations data are collected in the form of intervals. To analyze interval-valued data, usually researchers summarize the original data into single values, such as the centers or the medians of the intervals, but by doing so some important information in the original data is lost such as the range of the interval.

In the last years efforts have been done either to extend classical methods or to develop new approaches to deal with interval valued data. This paper has addressed the problem of detecting change points in interval-valued time series proposing, in the framework of atheoretical regression trees the use of a proper distance measure that accounts for the interval structure of the time ordered units. Indeed, change point analysis is a useful tool for monitoring and control and in the last decades it has emerged as a relevant research topic. However various

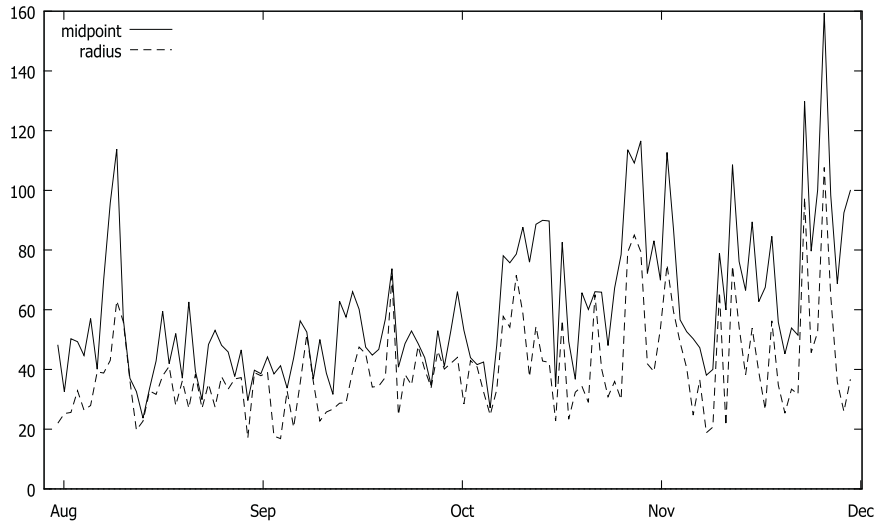

Fig. 5. PM10—time series of midpoints and radii.

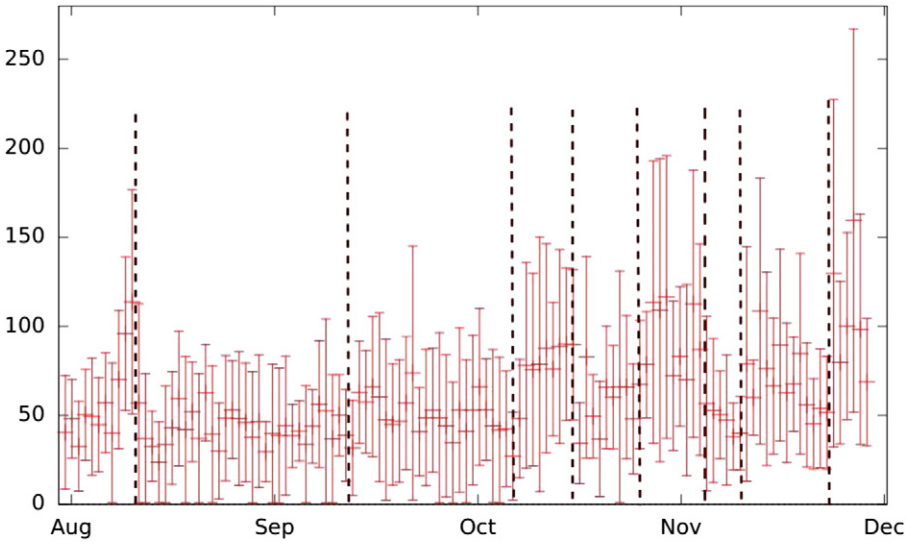

Fig. 6. PM10—error bars time series.

methods proposed in the literature consider the case single-valued time series.

We have presented the results of three simulation studies pertaining to different scenarios and an empirical application to a real interval valued time series that have shown the usefulness of the proposed procedure. Indeed, according to the simulation experiments the method selects the number of changes and their location accurately. In particular, when changes are present only in the centers and thus the data in terms of change point analysis can be treated as single valued, the use of a distance measure that accounts for the width of the intervals favors the correct identification of the change points. In case of changes only in the radii our proposal represents a tool to detect change points that

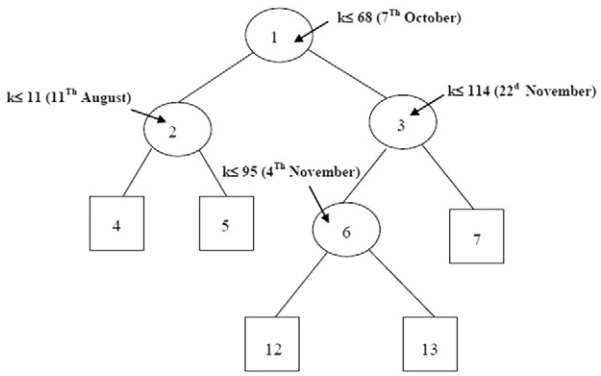

Fig. 7. PM10—regression tree.

Table 6
Stylized facts of the entire series and of the subperiods identified by the change points.

	$\bar{c}$	$\bar{r}$	sd_c	sd_r
Entire series				
1 August–30 November	60.3	40.3	24.2	16.5
Regimes				
01 August–11 August	59.9	36.4	24.6	13.3
12 August–07 October	45.7	34.5	10.5	9.5
08 October–4 November	77.2	49.4	21.33	18.1
05 November–22 November	62.4	39.2	18.2	15.9
22 November–30 November	103.6	58.3	28.7	29.8

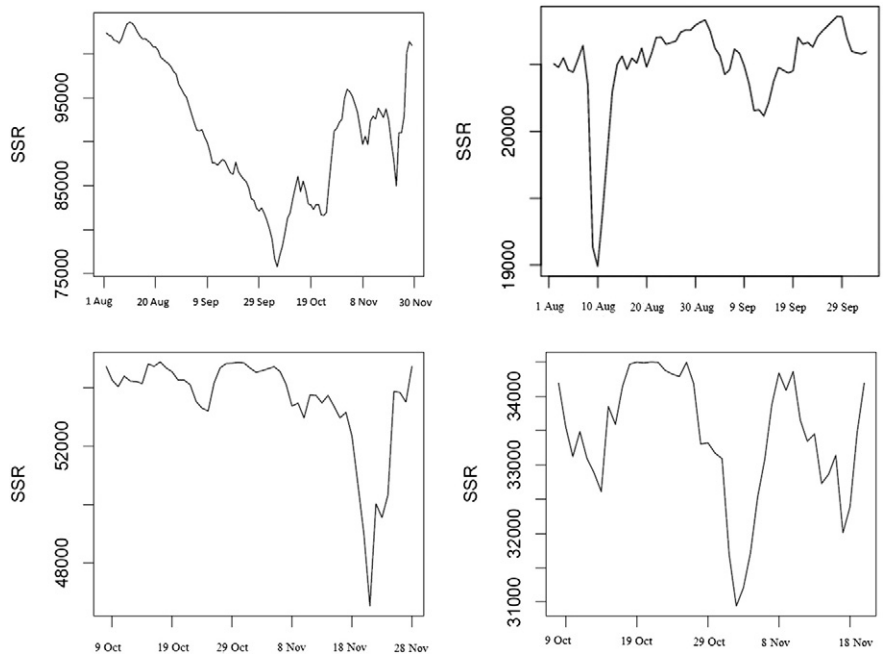

Fig. 8. PM10 behavior of the objective function.

would otherwise be missed by classical methods for single valued data typically applied to the series of the centers. Eventually the method provides the best performances when both centers and radii are subject to change. Overall the simulations confirm that the method is effective and it can be confidently employed for change point analysis.

In the application we have considered a time series of an air pollutant, the particulate matter that is responsible for harmful effects on health. The analysis has shown that the concentration of PM10 does not remain constant over the period of interest and that the method also helps identify outliers.

Eventually the procedure can be easily implemented in any software that provides the classification and regression tree methodology and it provides a quick flexible tool that, due to its simplicity, is particularly useful for applied time series analysis.

Conflict of interest

The authors declare no conflict of interest.

References

[1] S. Alizadeh, M.W. Brandt, F.X. Diebold, Range-based estimation of stochastic volatility models, J. Financ. 57 (2002) 1047–1091.
[2] C.W.D. de Almeida, R.M.C.R. de Souza, A.L.B. Candeias, Fuzzy Kohonen clustering networks for interval data, Neurocomputing 99 (1) (2013) 65–75.
[3] J. Arroyo, R. Espínola, C. Maté, Different approaches to forecast interval time series: a comparison in finance, Comput. Econ. 37 (2011) 169–191.
[4] J. Bai, P. Perron, Computation and analysis of multiple structural change models, J. Appl. Econ. 18 (2003) 1–22.
[5] J. Bai, P. Perron, Multiple structural change models: a simulation analysis, in: M. Broy, E. Dener (Eds.), Econometric Theory and Practice: Frontiers of Analysis and Applied Research, Cambridge University Press 2006, pp. 212–237.
[6] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey (CA), 1984.
[7] C. Cappelli, P. D'Urso, F. Di Iorio, Change point analysis of imprecise time series, Fuzzy Sets Syst. 225 (2013) 23–38.
[8] C. Cappelli, R. Penny, W. Rea, M. Reale, Detecting multiple mean breaks at unknown points in official statistic, Math. Comput. Simul. 78 (2008) 351–356.
[9] M.G.C.A. Cimino, B. Lazzerini, F. Marcelloni, W. Pedrycz, Genetic interval neural networks for granular data regression, Inf. Sci. 257 (2014) 313–330.
[10] G. Dierckx, J.L. Teugels, Change point analysis of extreme values, Environmetrics 21 (2010) 661–686.
[11] P. D'Urso, Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data, Comput. Stat. Data Anal. 42 (1-2) (2003) 47–72.
[12] P. D'Urso, P. Giordani, A least-squares approach to principal component analysis for interval-valued data, Chemom. Intell. Lab. Syst. 70 (2004) 179–192.
[13] P. D'Urso, P. Giordani, A robust fuzzy k-means clustering model for interval valued data, Comput. Stat. 21 (2006) 251–269.
[14] P. D'Urso, P. Giordani, A weighted fuzzy c-means clustering model for fuzzy data, Comput. Stat. Data Anal. 50 (2006) 1496–1523.
[15] P. D'Urso, L. De Giovanni, Midpoint radius self-organizing maps for interval-valued data with telecommunications application, Appl. Soft Comput. 11 (2011) 3877–3886.
[16] P. D'Urso, L. De Giovanni, R. Massari, Trimmed fuzzy clustering for interval-valued data, ADAC 9 (2015) 21–40.
[17] T. Denoeux, M. Masson, Multidimensional scaling of interval-valued dissimilarity data, Pattern Recogn. Lett. 21 (2000) 83–92.
[18] A. Douzal-Chouakria, L. Billard, E. Diday, Principal component analysis for interval-valued observations, Stat. Anal. Data Min. 4 (2011) 229–246.
[19] C. Gallagher, R. Lund, E. Diday, Changepoint detection in daily precipitation data, Environmentrics 23 (2012) 407–419.
[20] G. Gonzalez-Rivera, W. Lin, Constrained regression for interval-valued data, J. Bus. Econ. Stat. 31 (4) (2013) 473–490.
[21] K.C. Gowda, T.R. Ravi, Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity, Pattern Recogn. Lett. 16 (1995) 647–652.
[22] P.J.F. Groenen, S. Winsberg, O. Rodrguez, E. Diday, I-Scal: multidimensional scaling of interval dissimilarities, Comput. Stat. Data Anal. 51 (2006) 360–378.
[23] D.S. Guru, B.B. Kiranagi, P. Nagabhushan, Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns, Pattern Recogn. Lett. 25 (2004) 1203–1213.
[24] C. Hajjar, H. Hamdan, Interval data clustering using self-organizing maps based on adaptive Mahalanobis distances, Neural Netw. 46 (2013) 124–132.
[25] M. Ichino, H. Yaguchi, Generalized Minkowsky metrics for mixed feature-type data analysis, IEEE Trans. Syst. Man Cybern. 24 (1994) 698–708.
[26] V.K. Jandhyala, S.B. Fotopoulos, N. Evaggelopoulos, Change-point methods for Weibull models with applications to detection of trends in extreme temperature, Environmetrics 10 (1999) 547–564.
[27] V.K. Jandhyala, S.B. Fotopoulos, J. You, Change-point analysis of mean annual rainfall data from Tucumán, Argentina, Environmetrics 21 (2010) 687–697.
[28] M. Jansen, Multiscale change point analysis in Poisson count data, Chemom. Intell. Lab. Syst. 85 (2007) 159–169.
[29] D. Jaurskova, Some problems with application of change-point detection methods to environmental dat, Environmetrics 8 (1997) 469–483.
[30] T. Juster, R. Suzman, An overview of the health and retirement study, J. Hum. Resour. 30 (1995) S7–S56.
[31] J. Le-Rademacher, L. Billard, Symbolic covariance principal component analysis and visualization for interval-valued data, J. Comput. Graph. Stat. 21 (2012) 413–432.
[32] Y. Lertworaprachaya, Y. Yang, R. John, Interval-valued fuzzy decision trees with optimal neighbourhood perimeter, Appl. Soft Comput. 24 (2014) 851–866.
[33] R.B. Lund, X.L. Wang, J. Reeves, Q. Lu, C. Gallagher, Y. Feng, Change point detection in periodic and autocorrelated time series, J. Clim. 20 (2007) 5178–5190.
[34] C.F. Manski, E. Tamer, Inference on regressions with interval data on a regressor or outcome, Econometrica 70 (2002) 519–546.

[35] S.R.K. Raju, Symbolic Data Analysis in Cardiology, in: E. Diday, K.C. Gowda (Eds.), Symbolic Data Analysis and Its Applications, CEREMADE, Universit Paris-Dauphine 1997, pp. 245–249.
[36] W. Rea, M. Reale, C. Cappelli, J.A. Brown, Identification of changes in mean with regression trees: an application to market research, Econ. Rev. 29 (2010) 754–777.
[37] P.M.M. Rodriguez, S. Nazarii, Modeling and forecasting interval time series with threshold models, ADAC (2014), http://dx.doi.org/10.1007/2Fs11634-014-0170-x.
[38] P. Teles, P. Brito, Modelling Interval Time Series with Space-Time Processes, Communications in Statistics - Theory and Methods 2013, http://dx.doi.org/10.1080/03610926.2013.782200.
[39] W.B. Wu, M. Woodroofe, M. Mentz, Isotonic regression: another look at the change-point problem, Biometrika 88 (2001) 793–804.