Gazi University Journal of Science
GU J Sci
28(2):331-347 (2015)

A Study on Time Series Clustering

Filiz KARDİYEN[1,♠], Hilal GÜNEY[1]

[1]*Gazi University, Department of Statistics, Beşevler, Ankara*

Received: 11/03/2015 Accepted: 28/04/2015

ABSTRACT

In recent years, the topic of classification which is advantageous in terms of time and cost is of great interest in various fields. In particularly, when a large number of series should be analyzed, it is much more practical to classify the series into similar groups, and to make an estimation for each corresponding group rather than to make prediction for every given series individually. For this reason, some studies have been carried out in order to develop classification and clustering methods using characteristics of the time series. In this study, model based approaches: Maharaj's p-value based distance, Piccolo's AR distance, Cepstral based distance and free model based methods: Autocorrelation based distance, Chouakria-Douzal dissimilarity measure, Minkowski distance are compared in terms of clustering performances of time series. The performances of the clustering methods are also investigated for different ordered sets of the processes, and correlation structures among the series. In the result of the study, it is obtained that Maharaj's p-value based distance is the best method regarding to clustering performance, and Piccolo's AR distance based clustering is the least affected method by the different order of the processes.

Keywords: Autocorrelation based distance; AR distance; p-value based distance; Chouakria-Douzal dissimilarity measure; Cepstral based distance; Minkowski distance

1. INTRODUCTION

The purpose of the classification is to divide unlabeled data sets into homogeneous groups in such a way to ensure maximum similarity within groups and maximum dissimilarity between groups [1]. In recent years, time series classification has attracted much attention and has applications in many areas such as economics, business, demography, geology, medicine and climatology. Especially, when a large number of series should be analyzed much more practical to classify the series into similar groups and to make an estimate for each corresponding group rather than to make prediction for every given series individually.This is accompained by advantages in terms of time and cost. For this reason, determining similarities of the time series to constitute classes has been very popular in recent years.

There are many studies in the literature on time series classification and clustering. In general, clustering methods are studied in two main categories: model free-based methods (non-parametric methods), and model-based methods (parametric methods) [21]. Distance-based methods are often used for the classification of time series. Piccolo (1990) defined a metric based on autoregressive expansion for ARIMA models to be used for time series classification and clustering, and applied this parametric measure to determine the similarities between industrial production series. Tong and Dabas (1990) investigated some of the similarities between linear and non-linear models by applying the classical clustering

♠Corresponding author, e-mail: fyuva@gazi.edu.tr

techniques. Shaw and King (1992) used cluster analysis and principal component analysis to classify time series. Maharaj (1996) proposed a hypothesis test to classify ARMA models. Kakizawa et al. (1998) used Kullback-Leibler discrimination information and the Chernoff information measure to discriminate multivariate time series. Maharaj (2000) expanded a hypothesis test and proposed to classify two time series with an algorithm based on the p value and demonstrated the performance of the algorithm with a simulation study and an application. To measure the distance between time series, Galeano and Pena (2000) developed a metric based on autocorrelation function (ACF). Distances based on partial autocorrelation (PACF) and the inverse autocorrelation function (IACF) introduced by Cleveland (1972) has not yet been applied to the problem of clustering. Alonso et al. (2006) proposed a distance measure based on the full probability density of the forecasts. Corduas and Piccolo (2007) classified time series using the AR distance. Chouakria and Nagabhushan (2007) proposed a classification metric that takes into account the behavior of the growth time series. Vilar et.al. (2010) studied a time series clustering method using full forecast densities of time series, and this method showed good performance for nonlinear autoregressive models. Additionally, Liu et. al. (2014) developed an approach to time series classification based on a polarization measure of forecast densities of time series.

Many measurements have been described to measure the distances and similarities of the series in the process of clustering. In this study, first, the clustering methods discussed have been briefly introduced, then the clustering method which uses autocorrelation functions of series as a measure of similarity, Piccolo's clustering method based on AR distance, Maharaj's p-value algorithm, Chouakria-Douzal dissimilarity measure, Minkowski measure and cepstral based distance have been compared in terms of clustering performance.

2. MODEL-BASED APPROACHES

2.1. Maharaj's p value-based distance

Let Z_t be a zero mean stochastic process and a_t be a Gaussian white noise process with zero mean and constant variance. Z_t is assumed a member of the family of stationary and invertible $ARMA$ models. Using the standart notation of Box and Jenkins (1994) ARMA(p,q) model is expressed as

$$\phi(B)Z_t = \theta(B)a_t \tag{1}$$

where B is a backward shift operator and $\phi(B)$ and $\theta(B)$ are polynomials of degree p and q in B.

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \tag{2}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{3}$$

$\phi(B)$ and $\theta(B)$ satisfy the constraints of stationarity and invertibility. Z_t is a weighted sum of past values of Z's plus and added shock a_t, that is,

$$Z_t = \sum_{j=1}^{\infty} \pi_j Z_{t-j} + a_t \tag{4}$$

where the $AR(\infty)$ operator is defined by $\pi(B) = \varphi(B)\theta^{-1}(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$, and π_j's are the coefficients of the $AR(\infty)$ operators. See, for instance [2,20].

Let X_t and Y_t, $t = 1,2,...,T$ be two stationary time series. For modelling autoregressive structures, using a definite criterion such as Akaike's Information Criterion (AIC) or Schwartz's Bayesian Information Criterion (BIC) truncated $AR(\infty)$ models of order k_1 and k_2 can be fitted to X_t and Y_t, respectively. The parameter vectors of X_t and Y_t processes for $AR(k_1)$ and $AR(k_2)$ models are

$$\begin{aligned} \pi_x' &= \left[\pi_{1x} \quad \pi_{2x} \quad \ldots \quad \pi_{k_1 x}\right], \\ \pi_y' &= \left[\pi_{1y} \quad \pi_{2y} \quad \ldots \quad \pi_{k_1 y}\right] \end{aligned} \tag{5}$$

$\hat{\pi}_{jx}, j = 1,2,..k_1$, and $\hat{\pi}_{jy}, j = 1,2,..k_2$ represent the parameter estimates of $AR(k_1)$ and $AR(k_2)$, respectively. Let $k = \max(k_1, k_2)$. In constructing the test statistic, the maximum determined order k will be fitted to both series. The models with $T-k$ observations fitted to X_t and Y_t can be expressed as

$$X = W_x \pi_x + a_x, \quad Y = W_y \pi_y + a_y \tag{6}$$

Where

$$W_X = \begin{bmatrix} x_k & x_{k-1} & \cdots & x_1 \\ \cdots & \cdots & \cdots & \cdots \\ x_{t-2} & x_{t-3} & \cdots & x_{T-k-1} \\ x_{T-1} & x_{T-2} & \cdots & x_{T-k} \end{bmatrix} \tag{7}$$

$$X' = [x_{k+1} \quad \dots \quad x_{T-1} \quad x_T],$$
$$\pi'_X = [\pi_{1x} \quad \pi_{2x} \quad \dots \quad \pi_k],$$
$$a'_X = [a_{k+1,x} \quad \dots \quad a_{T-1,x} \quad a_{T,x}] \quad (8)$$

Y', W_y , π_y and a_y' can also be written in a similar way. Additionally

$$E[a_x] = 0 \, , E[a_x a_x'] = \sigma_x^2 I_{T-k} \quad (9)$$

$$E[a_y] = 0 \, , E[a_y a_y'] = \sigma_y^2 I_{T-k} \quad (10)$$

where I_{T-k} is the identity matrix with the dimension

$(T-k) \times (T-k)$.

It will be assumed that the disturbances of two models are correlated at the same points in time but uncorrelated across observations,

$$E(a_x a_y') = \sigma_{xy} I_{T-k} \quad (11)$$

Then, the combined model of the two equations in (6) becomes

$$Z = W\pi + a \quad (12)$$

Where

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}, W = \begin{bmatrix} W_x & 0 \\ 0 & W_y \end{bmatrix}, \pi = \begin{bmatrix} \pi_x \\ \pi_y \end{bmatrix}, a = \begin{bmatrix} a_x \\ a_y \end{bmatrix} \quad (13)$$

$$E(a) = 0 \, , E(aa') = V = \sum \otimes I_{T-k} \, ,$$

$$\sum = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \quad (14)$$

The parameter π is estimated by generalized least squares given by

$$\hat{\pi} = [W'V^{-1}W]^{-1} W'V^{-1}Z \quad (15)$$

In the p-value algorithm, the null hypothesis is defined as $H_0 : \pi_x = \pi_y$ or $H_0 : R\pi = 0$. where $R = [I_k \quad -I_k]$ and I_k is a identity matrix with the size of kxk.

The recommended statistics $D = (R\hat{\pi})' (R(W'\hat{V}^{-1}W)R')(R\hat{\pi})$ for this test is asymptotically chi-square distributed with k degree of freedom. The clustering procedure begins with performing the test of hypotheses for every pair of series and determining the p-value associated with the test *D*. Then groups of series are determined by comparing p-values. The steps of algorithm are shown in Figure 1 [13,14,15], in detail.

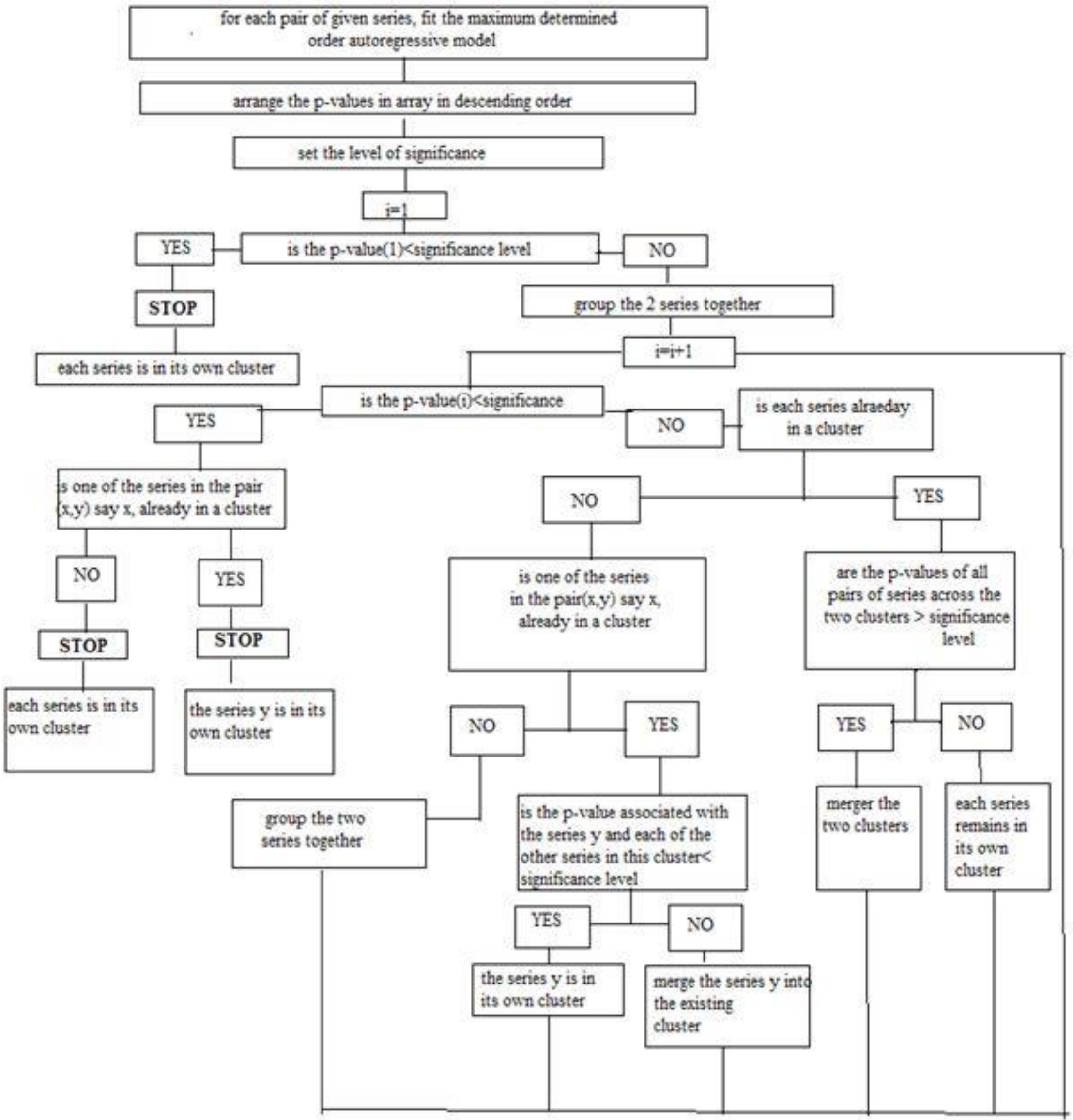

Figure 1. Maharaj's Clustering Algorithm

2.2. Piccolo's AR distance

In the proposition by Piccolo (1990), ARIMA models are fitted to all given time series, and then clustering the fitted series is performed by utilising a distance measure based on the coefficients of the corresponding $AR(\infty)$ operators [13]. The sequence of these coefficients $\pi' = (\pi_1, \pi_2, \ldots)$, which specifies completely the distribution of the process Z_t, fully characterize any stationary and invertible process as in equation (1). A measure of structural diversity between X_t and Y_t can be obtained by comparing the respective π sequences as follows:

$$d(X,Y) = \sqrt{\sum_{j=1}^{\infty} \left(\pi_{xj} - \pi_{yj}\right)^2} \tag{16}$$

$d(X,Y)$, satisfies the classical properties of a distance properties:

i. Non-negativity $d(X,Y) \geq 0$

ii. Symmetry $d(X,Y) = d(Y,X)$

iii. Triangularity $d(X,Y) \leq d(X,h) + d(h,Y)$

Distance matrix is constituted with the elements given in (16). Then, time series are grouped using a hierarchical clustering method [6,13,15,16].

2.3. Cepstral-based distance

In this approach, a distance measure based on the linear predicting coding (LPC) cepstrum is proposed. Cepstrum analysis is a kind of signal processing analysis.

The LPC coefficients for $AR(p)$ autoregressive time series are defined by;

$$c_n = \begin{cases} -\phi_1 & n = 1 \\ -\phi_n - \sum_{m=1}^{n-1}\left(1 - \frac{m}{n}\right)\phi_m c_{n-m} & 1 < n \leq p \\ -\sum_{m=1}^{p}\left(1 - \frac{m}{n}\right)\phi_m c_{n-m} & p < n \end{cases} \quad (17)$$

where ϕ_i's are , $i = 1,2,..., p$ are the autoregression coefficients. Note that for every ARIMA model there exists an equivalent AR model.

Euclidean distance between cepstral coefficients given in (17)

$$d_{CEPS} = \left\{ \sum_{t=1}^{T} \left(c_t^X - c_t^Y\right)^2 \right\}^{1/2} \quad (18)$$

with c^X and c^Y the cepstral coefficients of the series X_t and Y_t, respectively [8,13,16]. Following the determination of the distance matrix by distances defined in (18), the time series are clustered using a hierarchical clustering method.

3. FREE MODEL APPROACHES

3.1. Autocorrelation based distance

Cluster analysis is a statistical analysis that groups units, combining units with similar characteristics. In the literature, hierarchical and non-hierarchical clustering techniques have been used in the classification of time series and various distance measures have been proposed.

Galeano and Pena (2000) proposed a metric based on the estimated autocorrelation function (ACF) to take into account the dependence structure of the time series.

Let $\hat{\rho}_X = \left(\hat{\rho}_{X1},...,\hat{\rho}_{XR}\right)$ and $\hat{\rho}_Y = \left(\hat{\rho}_{Y1},...,\hat{\rho}_{YR}\right)$ be the estimated autocorrelation vectors of the series X and Y, respectively, for some R such that $\hat{\rho}_{Xi} \approx 0$ and $\hat{\rho}_{Yi} \approx 0$ for $i > R$.

Euclidean distance between estimated autocorrelation functions

$$d_{XY} = \sqrt{\sum_{r=1}^{R} \left(\hat{\rho}_{Xr} - \hat{\rho}_{Yr}\right)^2} \quad (19)$$

where $r = 1,\dots,R$ refers to lags of X and Y series [7,8].

Time series are divided into groups with a hierarchical clustering method after the distance matrix whose elements are given in (19).

3.2. Chouakria-Douzal dissimilarity measure

Chouakria and Nagabhushan (2007), suggested dissimilarity measure that takes into account the dynamic behaviour of the time series. The first order temporal correlation coefficient used as a measure of the first proximity between the dynamic behaviour of the series

$$CORRT(X_t, Y_t) = \frac{\sum_{t=1}^{T-1}\left(x_{t+1} - x_t\right)\left(y_{t+1} - y_t\right)}{\sqrt{\sum_{t=1}^{T-1}\left(x_{t+1} - x_t\right)^2}\sqrt{\sum_{t=1}^{T-1}\left(y_{t+1} - y_t\right)^2}}$$

(20)

Temporal correlation coefficient belongs to the interval $[-1,\ 1]$. If $CORRT(X_t, Y_t) = 1$, then the two series have the same dynamic structure, and change in the same direction whereas if $CORRT(X_t, Y_t) = -1$ two series have the same dynamic structure, and change in the opposite direction and if $CORRT(X_t, Y_t) = 0$ then series are statistically linearly independent.

Chouakria-Douzal defines a dissimilarity index D as follows:

$$D_{CORRT}(X_t, Y_t) = f\left(CORRT\left(X_t, Y_t\right)\right)\delta_{conv}\left(X_t, Y_t\right)$$

(21)

where, $f(.)$ is an adaptive tuning function to modulate a given raw data distance $\delta_{conv}(X_t, Y_t)$ according to the temporal correlation. Exponentially adaptive function is chosen as an adaptive tuning function

$$f(u) = \frac{2}{1+\exp(ku)}, k \geq 0 \tag{22}$$

For $k = 0$, D_{CORRT} equals $\delta\left(X_t, Y_t\right)$ [4,15].

Then, a hierarchical clustering method using the distances given in (21) is employed to cluster the time series.

3.3. Minkowski distance

Minkowsky distance is given by:

$$D_{MIK}(X,Y) = \left(\sum_{t=1}^{T}(x_t - y_t)^q\right)^{1/q} \tag{23}$$

with q a positive integer [15]. Following the determination of the distance matrix consists of the distances defined in (23), time series are clustered using a hierarchical clustering method.

4. SIMULATION STUDY

To compare the clustering performances of the methods, five different time series processes were considered, and the experimental study is performed through the followings.

$AR(1)$	$\phi_1 = 0.5$	
$MA(1)$	$\theta_1 = 0.7$	
$AR(2)$	$\phi_1 = 0.6$	$\phi_2 = 0.2$
$MA(2)$	$\theta_1 = 0.8$	$\theta_1 = -0.4$
$ARMA(1,1)$	$\phi_1 = 0.8$	$\theta_1 = 0.2$

To consider the potential influence of order that the series put into clustering processes, three different ordered sets of time series models were described: [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$]; [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$] and [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$]. For each sets of models, four series of length *200* were generated for each models. Throughout the generating procedure of four series of any considered model within the set, the correlations between the disturbances of each pair of series 0, 0.5, 0.75 ve 0.9 were used. The same correlation values were employed for each model within the set. Implementing Maharaj's p-value based distance, the simulated series were fitted with truncated *AR(k)* models, where the order *k* (up to *10*) was selected by Bayesian Information Criterion, BIC. After this initial step, p-values for each pair of series were computed, and then hypothesis testing was conducted using these values at *alpha=0.05* significance level. For AR distance method, autocorrelation coefficient was calculated from the whole series, instead of one quarter lenght of the series. The *h* values were determined as *1,2* and *5* for Chouakria-Douzal dissimilarity measure.

The clustering procedure was simulated 500 times. For each of the simulations, number of exactly correct clusters was counted for each method. By subtracting this number from the true number of correct clusters, measure of discrepancy was computed, that is,

MD=4-number of exactly correct clusters. Since series were generated from four different processes, then the true number of correct is four. Percentages of MD in 500 simulations are given in Table 1-3. To clarify the ouput of calculations given in tables, it would be good to explain one cell of table. In Table 1, for h=1 ve Correlation=0 , the percentage of only one misclustering is 41.2 %. The distribution of MD for each set of models is given in Figures 2(a)-4(d). Since Choukaria-Douzal dissimilarity measure shows the best performance at this value of *h*, figures are given for *h*=5.

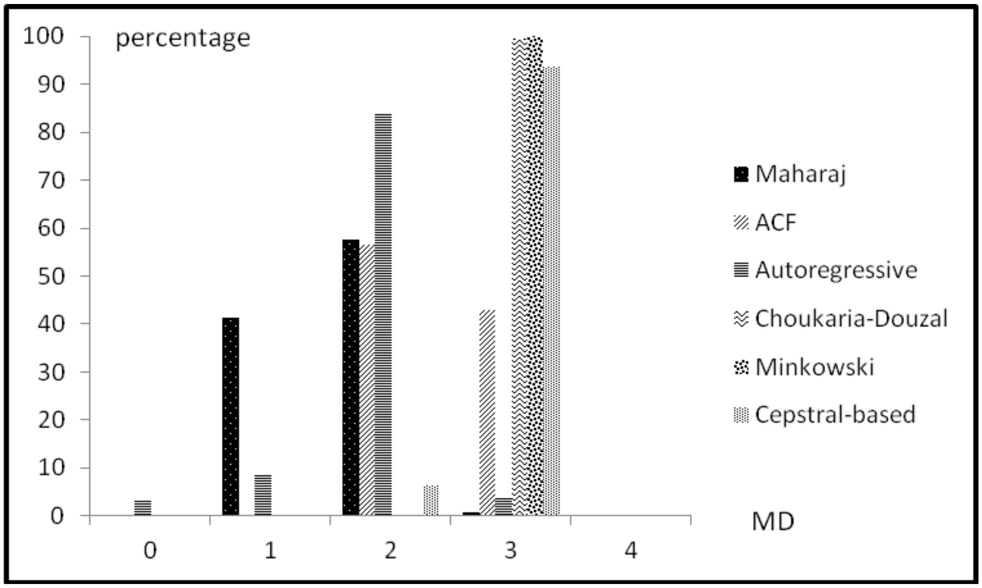

Figure 2(a). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$] when the Correlation is 0

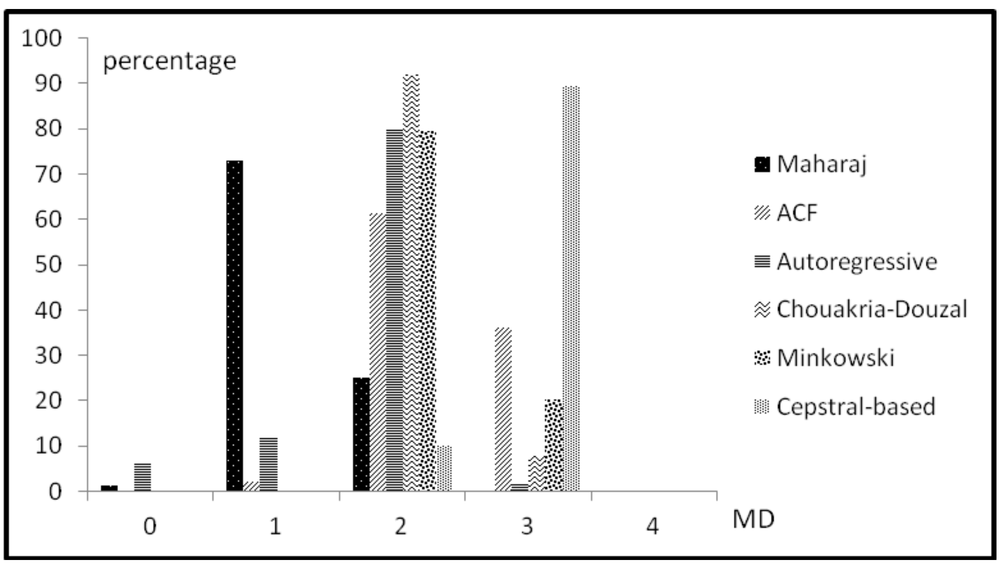

Figure 2(b). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$] when the Correlation is 0.5

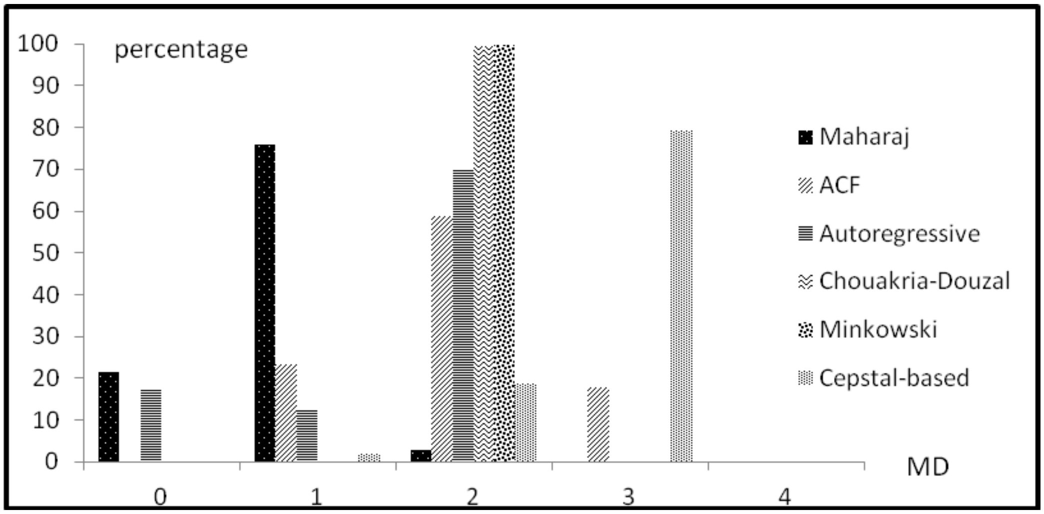

Figure 2(c). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$] when the Correlation is 0.75

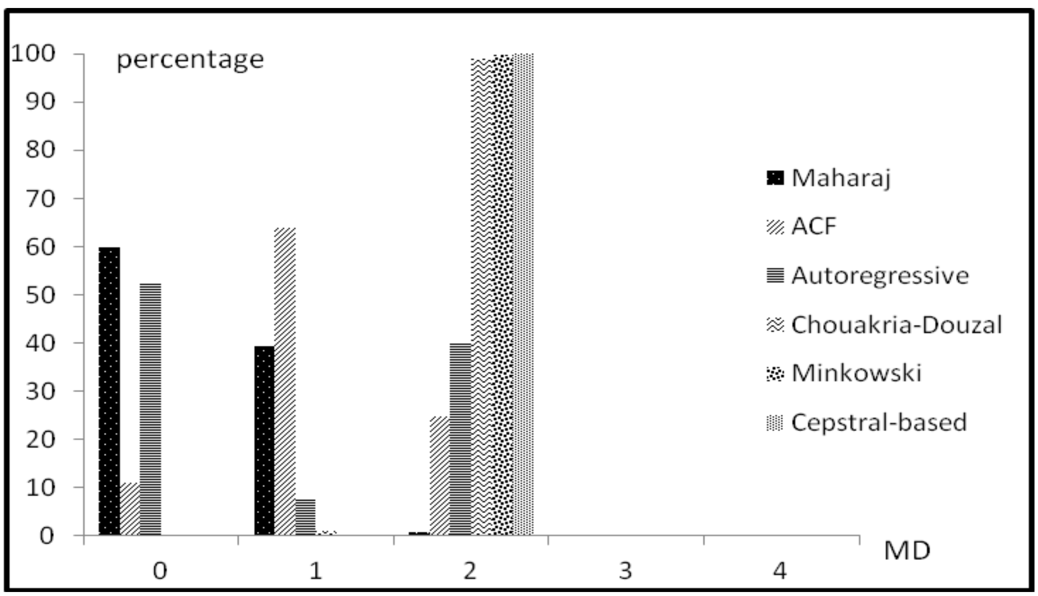

Figure 2(d). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$] when the Correlation is 0.9

Looking at Table1 and Figures 2(a)-2(d), which cover the results for the set of [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$], it is observed that the performances of all methods are better as the correlation between disturbances of series in same process increases. It is more obvious for Maharaj's p value algoritm. Maharaj's p-value algorithm outperforms the other methods with the highest percentage for MD of zero and one. In terms of performance, this is followed by the AR distance and ACF method, while Cepstral based distance method shows low performance for all correlation structures.

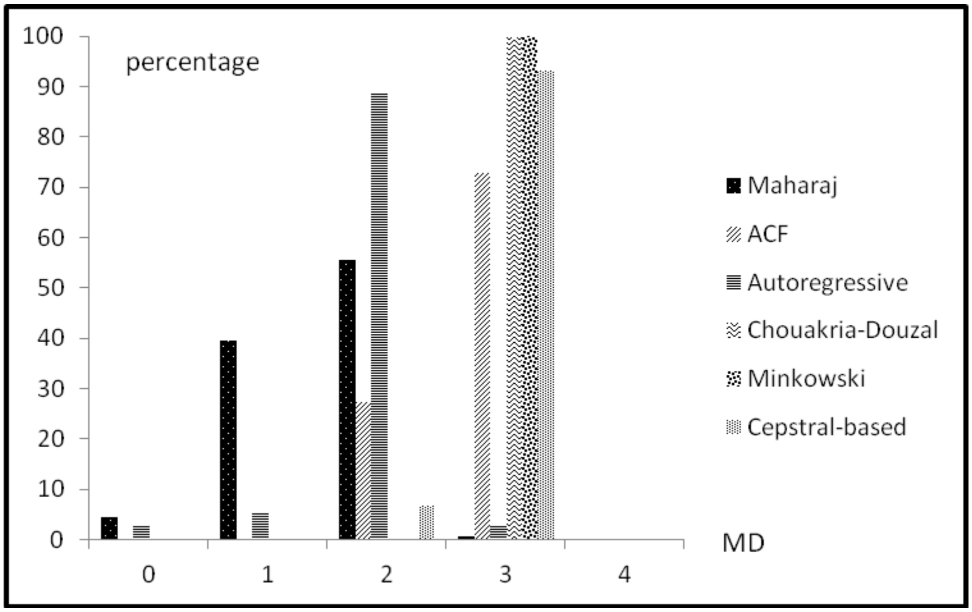

Figure 3(a). Distribution of the measure of discrepancy for the set [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$] when the Correlation is 0

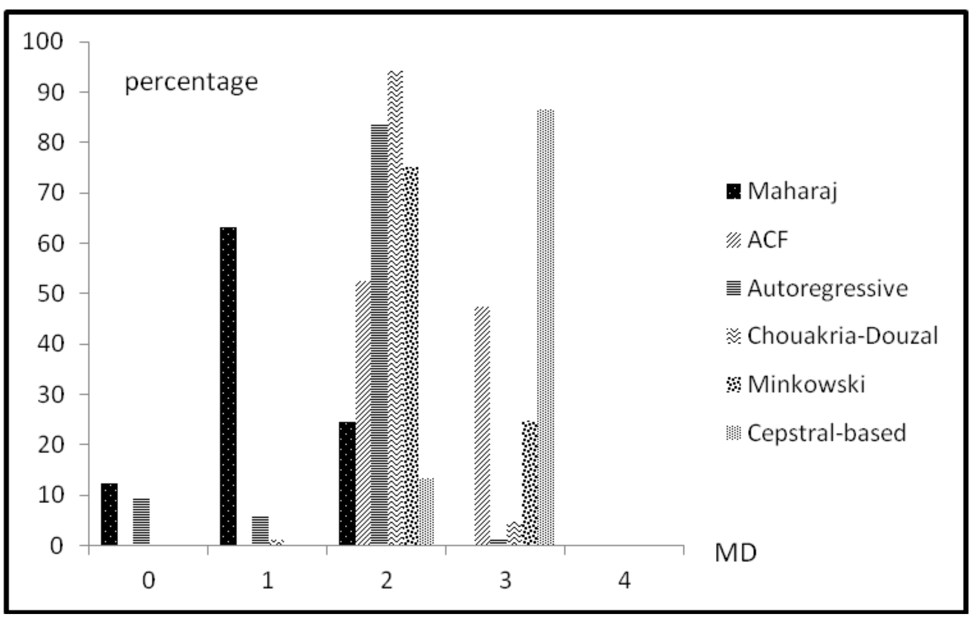

Figure 3(b). Distribution of the measure of discrepancy for the set[$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$] when the Correlation is 0.5

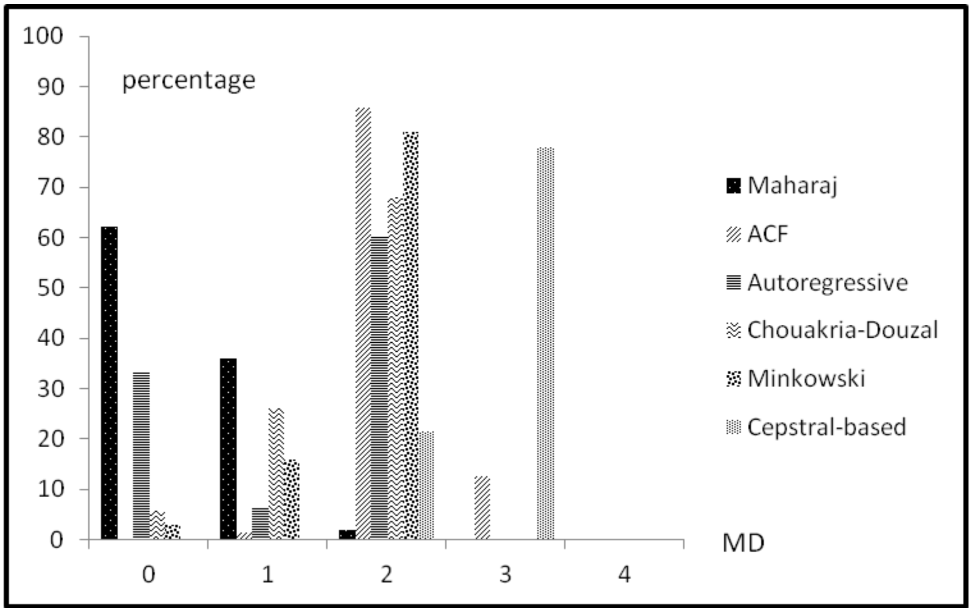

Figure 3(c). Distribution of the measure of discrepancy for the set [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$] when the Correlation is 0.75

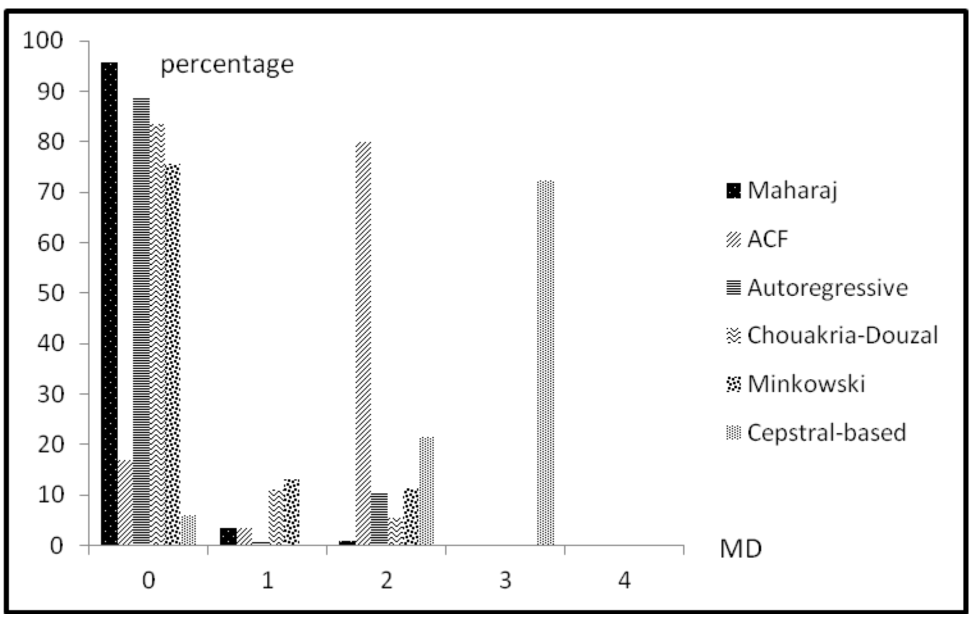

Figure 3(d). Distribution of the measure of discrepancy for the set [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$] when the Correlation is 0.9

Table 2 and Figures 3(a)-3(d) show the MD percentages of methods for the set of [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$]. Comparing this set with the set of [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$], it is seen the performances of the all methods declined. In particularly, the performance of the ACF method dramatically decreases for high correlations. It is observed that the clustering methods have difficulties in distinguishing ARMA(1,1) from AR(1) and MA(1) processes. Maharaj's p- value algorithm is the one having the best performance in clustering.

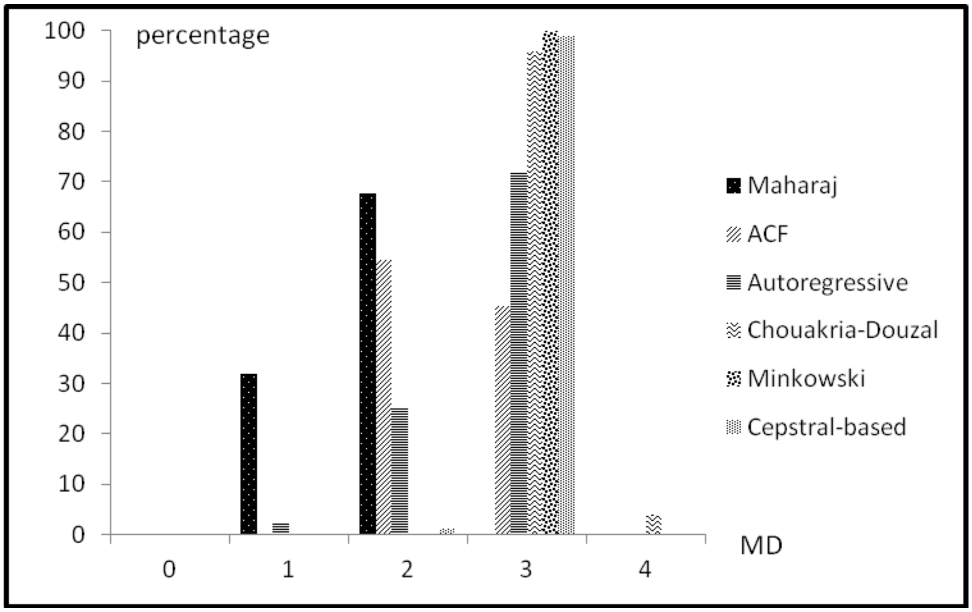

Figure 4(a). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$] when the Correlation is 0

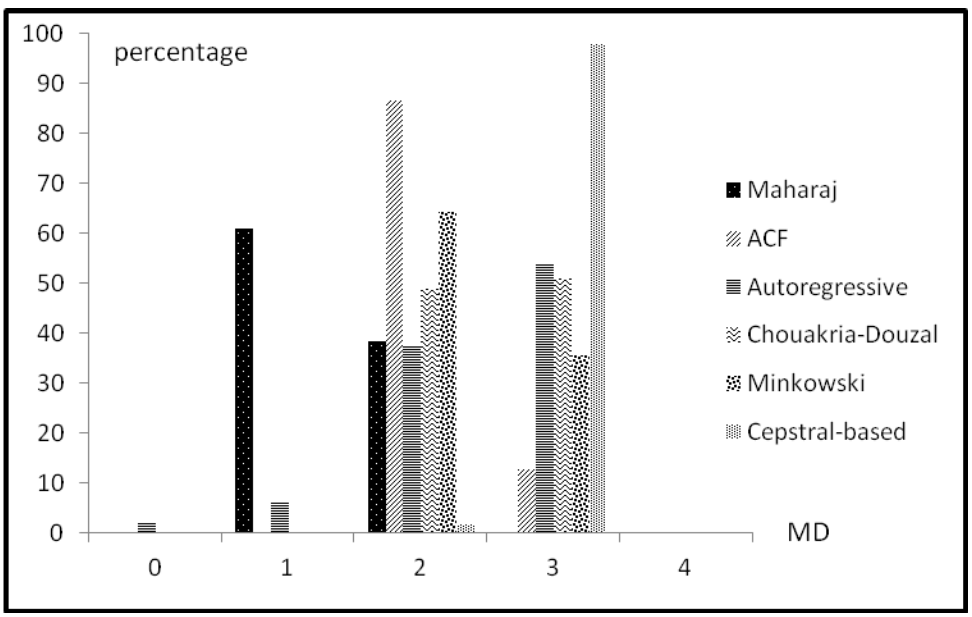

Figure 4(b). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$] when the Correlation is 0.5

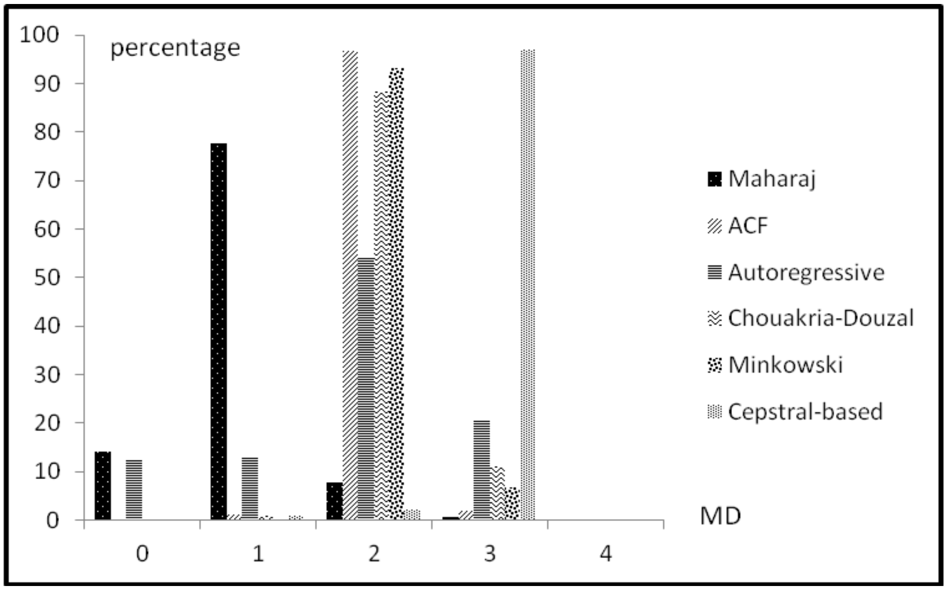

Figure 4(c).Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$] when the Correlation is 0.75

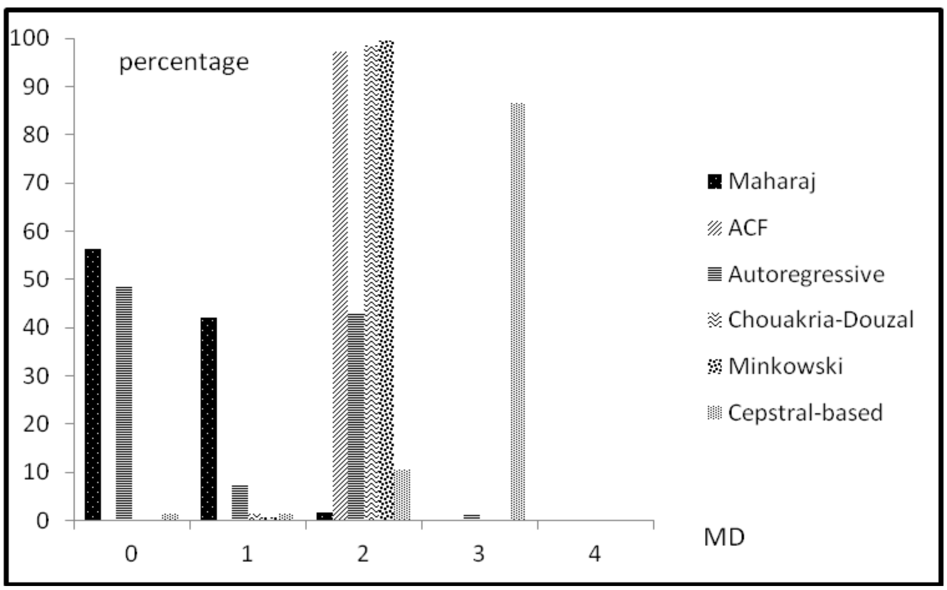

Figure 4(d). Distribution of the measure of discrepancy for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$] when the Correlation is 0.9

Table 3 and Figures 4(a)-4(d) are about the percentages of MD for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$]. The most striking result here is that Maharaj's p-value algorithm has 95% zero misclustering. The performances of AR distance and Choukria-Douzal dissimilarity methods follow Maharaj's p-value algorithm.

All methods except for ACF are resulted in having the highest performance in this set among all sets of the processes. This can be interpreted as ACF has difficulties in distinguishing AR parameters. In particularly, the results for Chouakria-Douzal method for the set of [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$] is atractive since the percentage for MD of zero and one increased dramatically. On the other hand, the ACF and Cepstral based distance method have the worst clustering performance for this set of processes.

Table1. Measure of discrepancy for the set [$AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$]

		Correlation=0			Correlation =0.5			Correlation =0.75			Correlation =0.9		
	h	1	2	5	1	2	5	1	2	5	1	2	5
Maharaj	0	0.4	0.2	0.2	1.4	2	1.2	21.4	19.2	25.8	50	59.8	56.4
	1	41.2	42	45.8	73	69.4	74.4	75.8	78.6	71.8	48.2	39.4	42.4
	2	57.6	56.8	53.6	25.2	28.4	24.2	2.8	2.2	2.4	1.8	0.8	1.2
	3	0.8	1	0.4	0.4	0.2	0.2	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
ACF	0	-	-	-	-	-	-	0.4	0.8	0.4	11.8	11	9
	1	0.4	-	0.4	2.4	3.8	3.4	23.2	24.4	24.4	62.4	63.8	63.6
	2	56.6	54.6	58.2	61.4	66.2	62.8	58.6	56.6	57.4	25.2	24.8	26.8
	3	43	45.4	41.4	36.2	30	33.8	17.8	18.2	17.8	0.6	0.4	0.6
	4	-	-	-	-	-	-	-	-	-	-	-	-
Autoregressive	0	3.4	2.6	2.8	6.6	5.8	6.4	17.6	19.6	21.4	46.6	52.2	50
	1	8.8	8.2	10.2	12	12	9.8	12.6	9.2	11.4	8.8	7.8	8.4
	2	83.8	84.2	83.4	79.8	81.4	82.4	69.6	71.2	67.2	44.6	40	41.6
	3	4	5	3.6	1.6	0.8	1.4	0.2	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
Chouakria -Douzal	0	-	-	-	-	-	-	-	-	-	-	-	0.8
	1	-	-	-	-	1.2	1.4	0.4	1.8	11.4	0.2	1	7
	2	0.2	0.2	0.4	92	84.6	73.8	99.4	98	87.8	99.8	99	92.2
	3	99.4	86	85.4	8	14.2	24.8	0.2	0.2	-	-	-	-
	4	0.4	13.8	14.2	-	-	-	-	-	-	-	-	-
Minkowsky	0	-	-	-	-	-	-	-	-	-	-	-	-
	1	-	-	-	-	-	-	-	-	-	-	0.2	-
	2	-	-	-	79.6	79.8	79	99.6	99.6	99.6	100	99.8	100
	3	100	100	100	20.4	20.2	21	0.4	0.4	0.4	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
Cepstral	0	-	-	-	-	-	0.2	0.2	0.4	0.4	2.6	1.4	1.6
	1	-	0.2	-	0.4	0.8	10	1.8	0.6	0.6	2.6	1.8	2.4
	2	6.4	6.2	7.2	10.2	11.2	89.8	18.8	16.2	14.4	17.6	18.2	16.4
	3	93.6	93.6	92.8	89.4	88	-	79.2	82.8	84.6	77.2	78.6	79.6
	4	-	-	-	-	-	-	-	-	-	-	-	-

Table 2. Measure of discrepancy for the set [$AR(1)$, $ARMA(1,1)$, $MA(1)$, $MA(2)$]

		Correlation =0			Correlation =0.5			Correlation =0.75			Correlation =0.9		
	h	1	2	5	1	2	5	1	2	5	1	2	5
Maharaj	0	-	0.2	0.2	0.2	0.6	0.6	14	16	14.6	56.4	50.2	59.6
	1	31.8	39.8	36.8	61	60.2	61.2	77.6	76.6	77	42	48	39.4
	2	67.6	59.2	62.6	38.4	38.4	37.6	7.8	7.4	8.2	1.6	1.6	1
	3	0.6	0.8	0.4	0.4	0.8	0.6	0.6	-	0.2	-	0.2	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
ACF	0	-	-	-	-	-	-	-	-	-	-	0.2	-
	1	0.2	0.2	0.2	0.2	0.8	0.6	1.2	1.2	1	2.8	1.4	2.2
	2	54.4	52	58.4	86.8	85.8	84.2	98.6	96.6	97.6	97.2	98.4	97.8
	3	45.4	47.8	41.4	13	13.4	15.2	2	2.2	1.4	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
Autoregressive	0	0.4	0.6	-	2.4	2.2	2	12.2	10.6	12.8	48.4	42.8	48.2
	1	2.4	3.6	3.6	6.2	5	5.2	12.8	12.2	12.2	7.6	14.2	9.4
	2	25.2	25.4	25	37.6	37	35.4	54.2	53.8	52	43	41.8	41.6
	3	72	70.4	71.4	53.8	55.8	57.4	20.8	23.4	23	1	1.2	0.8
	4	-	-	-	-	-	-	-	-	-	-	-	-
Chouakria Douzal	0	-	-	-	-	-	-	-	-	0.2	-	0.4	1.6
	1	-	-	-	-	0.6	-	0.8	3.2	7.8	1.4	2.4	11
	2	-	0.2	0.4	49	41.4	32.6	88.2	78.2	63.2	98.6	97.2	87
	3	96	92.4	87.6	51	58	67.2	11	18.6	28.8	-	-	0.4
	4	4	7.4	12	-	-	0.2	-	-	-	-	-	-
Minkowsky	0	-	-	-	-	-	-	-	-	-	-	-	-
	1	-	-	-	-	-	0.2	-	-	0.4	0.6	0.4	1
	2	-	99.8	-	64.4	66	64.6	93.2	92.8	93	99.4	99.4	98.8
	3	100	0.2	100	35.6	34	35.2	6.8	7.2	6.6	-	0.2	0.2
	4	-	-	-	-	-	-	-	-	-	-	-	-
Cepstral	0	-	-	-	-	-	-	-	-	-	1.4	3	1.4
	1	-	0.2	-	0.2	-	-	0.8	0.4	1.2	1.4	2	1.4
	2	1	0.4	0.4	1.8	99	2	2.2	5.6	4.2	10.6	12.6	11.6
	3	99	99.4	99.6	98	-	98	97	94	94.6	86.6	82.4	85.6
	4	-	-	-	-	-	-	-	-	-	-	-	-

Table 3. Measure of discrepancy for the set [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$]

		Correlation =0			Correlation =0.5			Correlation =0.75			Correlation =0.9		
	h	1	2	5	1	2	5	1	2	5	1	2	5
Maharaj	0	4.4	3.8	3.4	12.2	13.8	13	62.2	63.6	62.2	95.8	96.2	95.4
	1	39.4	40	41	63.2	66.4	63.4	36	34	36	3.4	3.8	4
	2	55.6	55.6	54.8	24.4	19.2	23	1.8	2.4	1.8	0.8	-	0.6
	3	0.6	0.6	0.8	0.2	0.6	0.6	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
ACF	0	-	-	-	-	-	-	0.2	0.8	0.2	16.8	17	15.4
	1	-	-	-	0.2	-	-	1.4	-	1.4	3.4	5.4	5
	2	27.2	22.4	23.8	52.4	52.4	55.4	85.8	88.6	86	79.8	77.2	79.8
	3	72.8	77.6	76.2	47.4	47.6	44.6	12.6	10.6	12.4	-	0.4	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
Autoregressive	0	3	4	2.2	9.4	8	10.2	33.4	36	33.4	88.8	86.4	87.4
	1	5.2	4.8	3.8	6	7.6	4.4	6.2	4.4	6.4	0.6	1.6	0.4
	2	88.8	86.2	90.2	83.6	83.2	83.4	60.2	59.6	60	10.6	12	12.2
	3	3	5	3.8	1	1.2	1.6	0.2	-	0.2	-	-	-
	4	-	-	-	-	-	-	-	-	-	-	-	-
Chouakria -Douzal	0	-	-	-	-	-	-	5.8	7	9.4	83.6	86.6	89.6
	1	-	-	-	1	1.4	2.4	26	27.4	27.8	11	8.4	8.8
	2	-	-	1.6	94.2	84.6	73.8	68	65.6	60	5.4	5	1.6
	3	99.8	97	83.2	4.8	14	23.6	0.2	-	2.8	-	-	-
	4	0.2	3	15.2	-	-	0.2	-	-	-	-	-	-
Minkowsky	0	-	-	-	-	-	-	2.8	1.4	2.8	75.6	74.4	76.2
	1	-	-	-	-	0.2	0.4	15.8	15.4	15.8	13	12.2	13.6
	2	-	-	-	75.2	72.6	71.4	81	83.2	81	11.4	13.4	10.2
	3	100	100	99.8	24.8	27.2	28.2	0.4	-	0.4	-	-	-
	4	-	-	0.2	-	-	-	-	-	-	-	-	-
Cepstral	0	-	-	-	-	-	-	0.4	0.4	0.4	6	3.2	4.6
	1	-	-	-	0.2	-	0.2	0.2	0.4	0.2	0.2	0.8	0.4
	2	6.8	4.2	6.4	13.2	11.2	12.2	21.4	15.2	2.2	21.4	17.4	16.8
	3	93.2	95.8	93.6	86.6	88.8	87.6	78	84	78.2	72.4	78.6	78.2
	4	-	-	-	-	-	-	-	-	-	-	-	-

8. CONCLUSION

In time series clustering, the idea is to investigate similarities of the time series in the identified clusters. It is important to note that clusters that are identified when carrying out clustering methods are not unique since they depend on the distance measure used, the clustering technique and the analyst. A wide variety of algorithms have been developed to classify different types of time-series data. In this study, some of the model-based and model free methods were compared in terms of their classification performance. For this purpose, several well-known model-based methods, Maharaj's p-value, Piccolo's AR distance-based hierarchical clustering method and Cepstral-based hierarchical clustering method; and also, some model-free methods, autocorrelation-based hierarchical clustering method, Choukria-Dousal dissimilarity method and Minkowski distance were evaluated. The simulation study showed that Maharaj's p-value algorithm was found as a method having the highest clustering performance for all correlation structures and all set of processes. Choukaria-Douzal dissimilarity method has displayed high performance when setting the experiment as high value of correlation coefficients and big value of *h*. Cepstral-based and Minkowski methods are not affected much by the increase of correlation between series. Regarding the performance of these methods, they may not preferable to other methods.

In order to examine the potential influence of order that the series put into clustering processes, three different ordered sets of the processes were taken into consideration in the comparison of the methods. The results of the simulation study showed that all of the methods were affected by the change of the order of the processes. It is observed that all clustering methods have difficulties in distinguishing ARMA(1,1) from AR(1) and MA(1) processes. All methods except for ACF are resulted in the highest performance for the set of [$AR(1)$, $AR(2)$, $ARMA(1,1)$, $MA(1)$]. It is also concluded that ACF method has difficulty in distinguishing AR parameters. Cepstral-based method was the least affected one by the order of the processes. Finally, we obtained more homogeneous clusters with the p-value algorithm.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Alonso, A., M., Berrendro, J., R., Hernandez, A., Justel, A., "Time Series Clustering Based on Forecast Densities", Computational Statistics & Data Analysis, Volume 51(2), 762-776, 2006.

[2] Box, G., E., P. and Jenkins, G., M.., Reinsel G. C., Time Series Analysis: Forcesting and Control 3.edn., Prentice Hall, New Jersey, 1994.

[3] Caiado, J., Crato, N., and Pena, D., "A Periodogram-Based Metric for Time Series Classification", Computational Statistics and Data Analysis, Volume 50(10), 2668-2684, 2006.

[4] Chouakria, A., D., Nagabhushan, P., N., "Adaptive Dissimilarity Index for Measuring Time Series Proximity", Advanced in Data Analysis and Classification, Volume 1(1), 5-21, 2007.

[5] Cleveland, W., S., "The Inverse Autocorrelation of Time Series and Their Applications", Tecnometrics, Volume 14(2), 277-293, 1972.

[6] Corduas, M., and Piccolo, D., "Time Series Clustering and Classification by the Autoregressive Metric", Computational Statistics & Data Analysis, Volume 52(4),1860-1872, 2008.

[7] D'Urso, P., Maharaj, E., A., "Autocorrelation-based Fuzzy Clustering of Time Series", Fuzzy Sets and Systems, Volume 160(24), 3565-3589, 2009.

[8] Galeano, P., and Pena, D., "Multivariate Analysis in Vector Time Series", Resenhas the Journal of The Institute of Mathematics and Statistics of The University of Sao Paolo, Volume 4, 383-404, 2000.

[9] Kalpakis, K. , Gada, D., and Vasundhara Puttagunta, V., "Distance Measures for Efective Clustering of ARIMA Time-Series", Proceedings 2001 IEEE International Conference on Data Mining, 273-280, 2001.

[10] Kakizawa, Y., Shumway, R., H., and Taniguchi, M., "Discrimination and Clustering for Multivariate Time Series", Jornal of American Statistical Association, Volume 93(441), 328-340, 1998.

[11] Liao, T., W., "Clustering of Time Series Data-a survey", Pattern Recognition, Volume 38(11), 1857-1874, 2005.

[12] Liu, S., Maharaj, E., A., Inder, B., "Polarization of Forecast densities:A new Approach to Time Series Classification", Computational Statistics and Data Analysis, Volume 70 (February), 345-361, 2014.

[13] Maharaj, E. A., "A Significance Test for Classifying ARMA Models", Journal of Statistical Computation and Simulation, Volume 54(4), 305-331, 1996.

[14] Maharaj, E., A., "Clusters of Time Series", Journal of Classification, Volume 17(2), 297-314, 2000.

[15] Manso, P., M., A Package for Stationary Time Series Clustering, Master thesis, Universidade da Coruna, 2013.

[16] Piccolo, D., "A Distance Measure for Classifying ARIMA Models", Journal of Time Series Analysis, Volume 11(2), 153-164, 1990.

[17] Shaw, C., T., and King, G., P., "Using Cluster Analysis to Classify Time Series", PhysicaD, Volume 58(1-4), 288-298, 1992.

[18] Tong, H., and Dabas, P., "Cluster of Time Series Models: An Example", Journal of Applied Statistics, Volume 17(2), 187-198, 1990.

[19] Vilar, J., A., Alonso, A., M., Vilar, J., M., "Non-Linear Time Series Clusterig Based on Non-Parametric Forecast Densities", Volume 54(11), 2850-2865, 2010.

[20] Wei, W.,W., S., Time Series Analysis Univariate and Multivariate Methods, Addison and Wesley Publishing Company, Canada, 1989.

[21] Xiong, Y., and Yeung, D., Y., "Time Series Clustering with ARMA Mixtures", Pattern Recognition, Volume 37,1675-1689, 2004.