Wiley
Computational and Mathematical Methods
Volume 2024, Article ID 5540143, 13 pages
https://doi.org/10.1155/2024/5540143

WILEY

Research Article

Affinity Coefficient for Clustering Autoregressive Moving Average Models

Ana Paula Nascimento[1,2] **Alexandra Oliveira**[2,3] **Brígida Mónica Faria**[2,3] **Rui Pimenta**[2,4] **Mónica Vieira**[1,5] **Cristina Prudêncio**[1,5] **and Helena Bacelar-Nicolau**[6,7]

[1]*Center for Translational Health and Medical Biotechnology Research (TBIO)/Health Research Network (RISE-Health), ESS, Polytechnic of Porto, R. Dr. António Bernardino de Almeida, 400, 4200-072 Porto, Portugal*
[2]*Biomatemática, Bioestatística e Bioinformática, ESS, Polytechnic of Porto, R. Dr. António Bernardino de Almeida, 400, 4200-072 Porto, Portugal*
[3]*Artificial Intelligence and Computer Science Laboratory (LIACC Member of LASI), University of Porto, Porto, Portugal*
[4]*Centre for Health Studies and Research of the University of Coimbra/Centre for Innovative Biomedicine and Biotechnology (CEISUC/CiBB), ESS, Polytechnic of Porto/REQUIMTE/LAQV, Rua Dr. António Bernardino de Almeida 4200-072 Porto, Portugal*
[5]*Ciências Químicas e das Biomoléculas, ESS, Polytechnic of Porto, R. Dr. António Bernardino de Almeida, 400, 4200-072 Porto, Portugal*
[6]*Faculty of Psychology, University of Lisbon (FPUL), Lisboa, Portugal*
[7]*Institute of Environmental Health, Faculty Medicine, University of Lisbon (ISAMB-FMUL), Lisboa, Portugal*

Correspondence should be addressed to Ana Paula Nascimento; ananascimento@ess.ipp.pt

Received 2 August 2023; Revised 22 March 2024; Accepted 29 April 2024; Published 24 May 2024

Academic Editor: Higinio Ramos

In various fields, such as economics, finance, bioinformatics, geology, and medicine, namely, in the cases of electroencephalogram, electrocardiogram, and biotechnology, cluster analysis of time series is necessary. The first step in cluster applications is to establish a similarity/dissimilarity coefficient between time series. This article introduces an extension of the affinity coefficient for the autoregressive expansions of the invertible autoregressive moving average models to measure their similarity between them. An application of the affinity coefficient between time series was developed and implemented in R. Cluster analysis is performed with the corresponding distance for the estimated simulated autoregressive moving average of order one. The primary findings indicate that processes with similar forecast functions are grouped (in the same cluster) as expected concerning the affinity coefficient. It was also possible to conclude that this affinity coefficient is very sensitive to the behavior changes of the forecast functions: processes with small different forecast functions appear to be well separated in different clusters. Moreover, if the two processes have at least an infinite number of π- weights with a symmetric signal, the affinity value is also symmetric.

Keywords: affinity coefficient; clustering; distance; similarity coefficient; time series

1. Introduction

Cluster analysis is a set of exploratory multivariate data analysis methods for identifying natural groups in data, based on a coefficient of similarity or dissimilarity between variables or between individuals, or more generally between statistical units of data [1]. The objective of this analysis is to identify clusters of statistical data units within the dataset that share similar characteristics, resulting in high similarity among units within a cluster and low similarity among units belonging to different clusters [1, 2]. Hierarchical cluster analysis algorithms, namely, agglomerative hierarchical cluster analysis, provide a hierarchy of partitions and are the most used [3]. Agglomerative strategy algorithms usually start with all

statistical units separated into unit sets (singletons), forming a cluster partition equal to the number of statistical units, and successively grouping the most similar groups (according to some measure of similarity or dissimilarity between statistical units and criterion of aggregation between groups) in the same cluster, until it forms a partition of a single cluster. The hierarchy obtained with the applied agglomerative hierarchical clustering analysis depends on the agglomerative method. The average linkage method produces more balanced hierarchies. This method does not have the chain effect of the single linkage method, and it is an intermediate method between the single linkage and complete linkage methods. Moreover, the average linkage method uses more information than the others [1]. Cluster analysis is also important in the domain of time series. In economics, finance, engineering, bioinformatics, geology, medicine, and biotechnology, the data is stored sometimes in the form of time series data. Automated acquisition systems and the growing storage capacity have made time series data available in these (and other) domains. It includes different applications, such as financial stock prices, electrocardiogram (ECG) measurements, blood pressure in health, satellite images, earthquake in earth observation, and even social media, among others [4]. Time series are prevalent in data mining applications, and clustering time series can be an appropriate alternative for time series classification models because annotations can be hard to get [5]. Time series data is seen as many data points but can be seen as a single object. Clustering these objects is advantageous because it allows the discovery of interesting patterns in time series data [4]. For example, economics may be interesting to look at some time series indicators and cluster them, such as Gross National Product or population growth [6]. In Engineering, Niennattrakul and Ratanamahatana demonstrated the utility of time series representation in the task of clustering multimedia data [7]. In medicine, the study of biological signals requires discrimination between signals caused by particular illnesses, namely, in the cases of ECG [8] and electroencephalogram (EEG) [9], data are stored in time series data. Clustering time series is useful, and a general overview of time series clustering methods can be found for instance in Maharaj, D'Urso, and Caiado and many others [4, 10–12]. A more usual approach in clustering time series is whole time-series clustering. In this approach, each time series is considered as a single object and clustering is performed concerning to their similarities [4, 10]. This approach can be performed considering that each time series is generated by a model (model-based approach), and the similarity between fitted models is evaluated. In the model-based approach, a raw time series is transformed into model parameters and then a suitable model dissimilarity/similarity and clustering algorithm is chosen and applied [4, 10]. Then, a proper dissimilarity/similarity measure for clustering time series based on the nature and specific purpose of the clustering task is needed.

1.1. Measures in Whole Time Series Clustering. In whole time-series clustering, Piccolo [13] introduced a Euclidean distance for Autoregressive Integrated Moving Average (ARIMA) models based on autoregressive (AR) representation of these processes. Corduas and Piccolo and Maharaj [14, 15] extended the model-based clustering idea of Piccolo by developing a testing procedure for differences between underlying models of each pair of independent time series by using equivalent AR expansions and using the p values of each test in an algorithm to cluster the time series. Also, Maharaj [16] extended the testing procedure for significant differences between underlying models of each pair of independent time series to that of related series. In the case of clustering seasonal time series, Piccolo's method can be extended by fitting seasonal ARIMA models. Scotto, Alonso, and Barbosa [17, 18] and D'Urso, Maharaj, and Alonso [19] proposed different aspects of clustering of seasonal time series based on the estimates of fitted generalized extreme value models. In every case, it is used a dissimilarity/similarity measure. Several dissimilarity/similarity measures between time series have been proposed [4, 10, 11]. One of the most used dissimilarity measures is the Euclidean distance between AR expansions of the fitted ARIMA models proposed by Piccolo and Corduas and Piccolo [13, 14]. This distance is independent of the signal of the values of the coefficients of AR expansions of the fitted ARIMA models and is very dependent on the magnitude of the coefficients of AR expansions of the fitted ARIMA models. Thus, cluster analysis based on a coefficient, which measures the similarity between the "profiles," in a domain of time series is needed.

An extension of the affinity coefficient proposed by Bacelar-Nicolau and Nicolau [20] and Nicolau and Bacelar-Nicolau [21] as a similarity measure between the AR expansions of the invertible linear time series for autoregressive moving average (ARMA) models is presented in this work. The affinity coefficient was introduced by Matusita [22] to measure the proximity between two distribution functions. Bacelar-Nicolau introduced the affinity coefficient in cluster analysis, as a basic similarity coefficient between variables or individuals [23, 24]. In the beginning, the affinity coefficient was applied to positive frequencies [23–25], that is, positive profiles. According to Bacelar-Nicolau and Nicolau [20] and Nicolau and Bacelar-Nicolau [21], the affinity coefficient, also called the Matusita–Nicolau coefficient [26], was generalized to real numbers. Later, this coefficient was extended to clustering of statistical data units, mainly in a three-way approach: while in the two-way case, each cell of the data matrix contains a single value, and in a three-way approach, each cell of the data matrix may contain a set of values, describing a probability distribution, a histogram (frequency distribution), or integer frequencies, for instance, instead of one single value [1, 2, 24, 27–30]. There are real advantages in using the affinity coefficient to measure the similarity between profiles because of its properties [2]. In fact, the affinity coefficient presents several relevant properties in its application to the field of exploratory analysis of multivariate data, revealing, for example, to be a more robust coefficient than the well-known correlation coefficient of Pearson, when it is intended to cluster a set of statistical of variables [31]. The associated distance is also a function of the profiles, so it depends less on the magnitude of values than the Euclidean distance. These properties also apply when the statistical units are time series, or multiple time series, or

similarity search models in time series [32, 33]. It is therefore natural to hope for a good quality in the results of time series cluster analysis based on the extension affinity coefficient or on the associated distance. Furthermore, cluster analysis that will be used, based on the affinity coefficient, admits either an empirical or a probabilistic approach, not yet applied to time series sets; therefore, it was proposed below an extension of the affinity coefficient to the domain of time series [24, 29, 30, 34, 35].

In the present work, it extended the affinity coefficient, related to the generalized affinity coefficient for real numbers, for the AR expansions of the invertible ARMA models (ARMA(p,q)) as a similarity measure between linear time series and particularized for ARMA(1,1). Also, it simulated ARMA(1,1) processes and the likelihood estimates of their coefficients are obtained to perform agglomerative hierarchical cluster analysis with the associated distance with the proposed extension of the affinity coefficient. These simulated results and discussions are shown.

2. Affinity Coefficient and Associated Distance for ARMA Models

The affinity coefficient and its extensions are measures of similarity used in hierarchical and nonhierarchical cluster analysis. Here, it proposed an extension of the affinity coefficient to the domain of time series.

2.1. Definition and Properties of the Affinity Coefficient for ARMA Models. A time series is a set of dependent observations made at successive points in time that have something structurally stable. Here, it is considered that the dependence between the observations verifies the linear equation of the ARMA model [36, 37]. A set of observations $X = (X_t, t \in Z)$ is called a process ARMA model of order p and q (ARMA(p,q)) if it is stationary and satisfies the equation:

$$X_t - \varphi_{x1}X_{t-1} - \cdots - \varphi_{xp}X_{t-p} = \varepsilon_t - \theta_{x1}\varepsilon_{t-1} - \cdots - \theta_{xq}\varepsilon_{t-q} \quad (1)$$

or, equivalently,

$$\varphi(B)X_t = \Theta(B)\varepsilon_t$$

with $\varphi(B) = 1 - \varphi_{x1}B - \cdots - \varphi_{xp}B^p$, $\Theta(B) = 1 - \theta_{x1}B - \cdots - \theta_{xq}B^q$ where $\varphi_{x1}, \cdots, \varphi_{xp}, \theta_{x1}, \cdots, \theta_{xq}$, are real coefficients, B is the backshift operator such that $B^kX_t = X_{t-k}$, $k = 0, 1, \cdots$, and $(\varepsilon_t, t \in Z)$ are white noises with constant variance σ^2 different from zero. The stationary process X satisfies the model ARMA(p,q), as in Equation (1), with the following conditions:

i. If $\varphi_{xp} \neq 0$, $\theta_{xq} \neq 0$,

ii. The solutions of the equations $\varphi(B) = 0$ and $\Theta(B) = 0$ do not have common roots,

iii. The solutions of the equations $\varphi(B) = 0$ and $\Theta(B) = 0$ are in a module strictly larger than one,

Then, there is one and only one stationary solution and the representation (1) is unique [36, 37].

Let $X = (X_t, t \in Z)$ be a zero mean ARMA(p,q) process defined as in Equation (1) satisfying the previous assumptions. Thus, because the solutions of the equation $\Theta(B) = 0$ are in a module strictly larger than one, then X is invertible and so can be represented in terms of its past values according to the AR (∞) formulation [36, 37], that is,

$$\prod(B)Xt = \varepsilon t \quad (2)$$

with

$$\prod(B) = \frac{\varphi(B)}{\Theta(B)} = 1 - \sum_{j=1}^{\infty}\pi_{x,j}B^j \quad (3)$$

$$\sum_{j=1}^{\infty}|\pi_{x,j}| < \infty \quad (4)$$

Note that the π- weights of the AR (∞) formulations, as in Equation (3), are the coefficients of the best linear forecast of X given in the past; that is, they are the coefficients of the orthogonal projection of the process given in the past. Note that also, if $q = 0$, the X process is called the AR process and denoted by AR(p). These processes are not represented in terms of their past values according to the AR (∞) formulation. If $p = 0$, the X process is called the moving average process and is denoted by MA(q). These processes can be represented in terms of their past values according to the AR (∞) formulation if the solutions of the equation $\Theta(B) = 0$ are in a module strictly larger than one [37]. In the following, let the processes X and Y satisfy the model ARMA(p,q), as in Equation (1), and satisfy the conditions i, ii, and iii. Thus, let be X and Y processes that can be represented according to the AR (∞) formulation. Let be also X and Y processes that their white noise is a Gaussian process, and then given the initial values, the operators $\varphi(B)$ and $\Theta(B)$ and the variance of the white noise the probabilistic structure of the processes X and Y are completely characterized [13, 14]. Thus, instead of using the X and Y processes to define the similarity measure, it is possible to use their representation according to the AR (∞) formulation. From this consideration, Piccolo [13] introduced the Euclidean distance between the π- weights of the AR (∞) formulations as a measure of structural dissimilarity between two ARIMA processes, with given orders [13]. In the same way, it may extend the affinity coefficient, according to Bacelar-Nicolau and Nicolau [20] and Nicolau and Bacelar-Nicolau [21], between the π- weights of the AR (∞) formulations as a measure of similarity between two ARMA processes, with given orders. In fact, Matusita [22] defined the proximity of finite or infinite distribution functions (positive valued data) with the affinity coefficient, and Bacelar-Nicolau [20] and Nicolau [21] prove that the affinity coefficient can be extended as a measure of similarity over real (thus, also negative) valued data. Since the π- weights of the AR (∞) formulations characterize the probabilistic structure of the

processes X and Y, under the conditions mentioned above, it may extend the affinity coefficient to the affinity coefficient between the π- weights of the AR (∞) formulations as a measure of similarity between two ARMA processes with given orders.

Let X and Y be a zero mean processes of ARMA(p,q) processes, as in Equation (1), where the white noise is Gaussian with constant variance σ^2 different of zero and satisfy the conditions i, ii, and iii. Then, the affinity coefficient between the processes X and Y is extended by

$$\text{Aff}(X,Y) = \sum_{j=1}^{\infty} \text{sign}\left(\frac{\pi_{x,j}}{\pi_{x.}}\right) \text{sign}\left(\frac{\pi_{y,j}}{\pi_{y.}}\right) \sqrt{\left|\frac{\pi_{x,j}\pi_{y,j}}{\pi_{x.}\pi_{y.}}\right|} \quad (5)$$

where $(\pi_{x,j}, j = 1, 2, \cdots,)$ and $(\pi_{y,j}, j = 1, 2, \cdots,)$, with $\pi_{x,j} \neq 0$ and $\pi_{y,j} \neq 0$, for some j, are the π- weights of the AR (∞) formulation of the X and Y ARMA processes with

$$\pi_{x.} = \sum_{j=1}^{\infty} |\pi_{x,j}| < \infty \quad (6)$$

$$\pi_{y.} = \sum_{j=1}^{\infty} |\pi_{y,j}| < \infty \quad (7)$$

Note that the expression (5) is the inner product between the following successions: $(\text{sign}(\pi_{x,1}/\pi_{x,.})\sqrt{|\pi_{x,1}/\pi_{x,.}|}, \cdots)$ and $(\text{sign}(\pi_{y,1}/\pi_{y,.})\sqrt{|\pi_{y,1}/\pi_{y,.}|}, \cdots)$. In case of $\pi_{x,j} \neq 0$, for some j, $j = 1, \cdots,$ and $\pi_{y,j} = 0, \forall j$, the affinity coefficient between the vectors $(\text{sign}(\pi_{x,1}/\pi_{x,.})\sqrt{|\pi_{x,1}/\pi_{x,.}|}$, $\text{sign}(\pi_{x,2}/\pi_{x,.})\sqrt{|\pi_{x,2}/\pi_{x,.}|}, \cdots)$, (0,0,0….) is 0, by the properties of inner product. On the other hand, if $\pi_{x,j} = 0, \forall j$ and $\pi_{y,j} = 0, \forall j$, both the processes are white noises, so the natural value for the affinity coefficient is one. Thus, the affinity coefficient can be applied also in these cases. The affinity coefficient defined, in Equation (5), satisfies the following property.

Property 1. The given two zero mean ARMA(p,q) processes X and Y defined, as in Equation (1), satisfied the conditions i, ii, and iii with Gaussian white noise with constant variance σ^2 different from zero. Then, the affinity coefficient is defined, as in Equation (5), between the X and Y processes verifying the following properties:

i. $\text{Aff}(X,Y)$ always exists and converges.

ii. $\text{Aff}(X,Y)$ is a symmetric coefficient.

iii. $-1 \leq \text{Aff}(X,Y) \leq \text{Aff}(X,X) = 1$.

iv. $\text{Aff}(X,Y) = 1$ if and only if $\pi'_y n = \alpha \pi'_x n, \forall n \in N$.

Proof 1. To define the similarity measure between X and Y, let us use their representation according to the AR (∞) formulation. Let $\pi_{x,j}, j = 1, \cdots$ and $\pi_{y,j}, j = 1, \cdots$ be the π-weights of the AR (∞) formulation of X and Y. Let be some j, $j = 1, 2, \cdots$, such that, $\pi_{x,j} \neq 0$ and let be some i, $i = 1, 2, \cdots$, such that $\pi_{y,i} \neq 0$.

i. Consider the series $\sum_{j=1}^{\infty} \sqrt{|\pi_{x,j}\pi_{y,j}/\pi_{x,.}\pi_{y,.}|}$. Let be $\forall n \in N$, $\pi'_x n = (\sqrt{|\pi_{x,1}/\pi^n_{x.}|}, \cdots, \sqrt{|\pi_{x,n}/\pi^n_{x.}|})$, where $\pi^n_{x.} = \sum_{j=1}^{n} |\pi_{x,j}|$.

Let us consider the succession of partial sums given by $\forall n \in \text{N}, S_n = \sum_{j=1}^{n} \sqrt{|\pi_{x,j}/\pi^n_{x.}||\pi_{y,j}/\pi^n_{y.}|}$. This succession is the inner product between $\pi'_x n$ and $\pi'_y n$, $\forall n \in \text{N}$. Then, by Cauchy–Schwarz inequality

$$\forall n \in N, \left|\left\langle \pi'_x n, \pi'_y n \right\rangle\right| \leq \left\|\pi'_x n\right\| \left\|\pi'_y n\right\|.$$

Thus, $\forall n \in \text{N}, S_n$ satisfies

$$S_n = \left\langle \pi'_x n, \pi'_y n \right\rangle = \left|\left\langle \pi'_x n, \pi'_y n \right\rangle\right| \leq \left\|\pi'_x n\right\| \left\|\pi'_y n\right\| = \sqrt{\frac{|\pi^n_{x.}|}{|\pi^n_{x.}|}} \sqrt{\frac{|\pi^n_{y.}|}{|\pi^n_{y.}|}} = 1$$

As a series of nonnegative terms is convergent if and only if the succession of partial sums is limited and because all absolutely convergent series are convergent series, then $\text{Aff}(X,Y)$ always exists and converges.

ii. From expression (5), the affinity coefficient, $\text{Aff}(X,Y)$, is a symmetric coefficient.

iii. Let be $\forall n \in N, T_n = \sum_{j=1}^{n} \text{sign}(\pi_{x,j}/\pi^n_{x.}) \text{sign}(\pi_{y,j}/\pi^n_{y.}) \sqrt{|\pi_{x,j}/\pi^n_{x.}||\pi_{y,j}/\pi^n_{y.}|}$. It can be seen that $\forall n \in N, T_n \leq |T_n| \leq S_n$. Then, $\text{Aff}(X,Y) \leq \sum_{j=1}^{\infty} \sqrt{|\pi_{x,j}\pi_{y,j}/\pi_{x,.}\pi_{y,.}|} \leq 1$, because all absolutely convergent series are also convergent and are proved above that $\forall n \in \text{N}, S_n \leq 1$, so $\text{Aff}(X,Y) \leq 1 = \text{Aff}(X,X) = |\pi_{x,.}/\pi_{x,.}|$. On the other hand, $|\text{Aff}(X,Y)| \leq \sum_{j=1}^{\infty} \sqrt{|\pi_{x,j}\pi_{y,j}/\pi_{x,.}\pi_{y,.}|} \leq 1$, then $-1 \leq \text{Aff}(X,Y) \leq 1$.

iv. By Cauchy–Schwarz inequality $|\langle \pi'_x n, \pi'_y n \rangle| = \|\pi'_x n\| \|\pi'_y n\|$ if and only if $\pi'_y \text{n} = \alpha \pi'_x n, \forall n \in \text{N}$.

Then

$$\forall n \in N, S_n = \left\langle \pi'_x n, \pi'_y n \right\rangle = \left|\left\langle \pi'_x n, \pi'_y n \right\rangle\right| = \left\|\pi'_x n\right\| \left\|\pi'_y n\right\| = \sqrt{\frac{|\pi^n_{x.}|}{|\pi^n_{x.}|}} \sqrt{\frac{|\pi^n_{y.}|}{|\pi^n_{y.}|}} = 1$$

if and only if $\pi'_y n = \alpha\pi'_x n, \forall n \in \mathrm{N}$. Thus, if $\forall n \in N, T_n = S_n$ and $\pi'_y n = \alpha\pi'_x n$ follows that $\mathrm{Aff}(X, Y) = 1$. On the other hand, if $\forall n \in N, T_n = -S_n$ and $\pi'_y n = \alpha\pi'_x n$ follows that $\mathrm{Aff}(X, Y) = -1$. □

Thus, in the conditions of the property, the π- weights uniquely determine the forecast function for future values given present and past values [13]. Assuming that ε_t is a Gaussian process and the orders are known, given the same set of initial values, we have that $|\mathrm{Aff}(X, Y)| = 1$ if the models X and Y produce the same forecasts or $\pi'_y n = \alpha\pi'_x n$, $\forall n \in N$.So, it is proved in the property above that the affinity coefficient defined as in Equation (5) is a similarity measure between the ARMA(p,q) processes X and Y satisfying the above conditions. This similarity measure generalizes the classical similarity measures from positive to real values.

Note that the affinity coefficient defined as in Equation (5) is also a similarity measure which satisfies the properties above for the MA(q) processes which can be represented in terms of its past values according to the AR (∞) formulation. When $q = 0$ and $p \neq 1$, X is an AR(p) process, and then, it cannot be represented in terms of its past values according to the AR (∞) formulation since $q = 0$; however, the coefficient defined as in Equation (5) remains a measure of similarity. When $q = 0$ and $p = 1$, that is, when the process is only a AR(1) and X cannot be represented in terms of its past values according to the AR(∞) formulation since $q = 0$, it happens that the affinity is equal to 1 and the models X and Y do not produce the same forecasts. Thus, the affinity coefficient behaves as a similarity coefficient for ARMA(p,q) processes, MA(q) processes, and AR(p) processes with $p \neq 1$, but for AR(1) processes, it does not verify the uniqueness because $\pi'_y n = \alpha\pi'_x n, \forall n \in N$. This restriction is not a problem, because a proper similarity measure must be based on the nature and specific purpose of the clustering task, thus allowing to interpret the clustering solutions in terms of grouping target [38]. For this coefficient, it can also use the associated distance to the domain of time series. In the following section, it presented the distance associated with the proposed $\mathrm{Aff}(X, Y)$ between ARMA(p,q) processes.

2.2. The Affinity Associated Distance for ARMA Models. Given two zero mean ARMA(p,q) processes X and Y, satisfying the conditions of the property above, the affinity associated distance between X and Y will be defined by the usual relationship

$$d(X, Y) = \sqrt{d^2(X, Y)} = \sqrt{2(1 - \mathrm{Aff}(X, Y))} \qquad (8)$$

Note that this is the equation which relates the standard affinity coefficient to the associated distance, in the case of X and Y being elements of $R^{n+} U \{0\}$, now extended to R^{∞}.

Property 2. The measure defined as in Equation (8) satisfies the properties of a distance between ARMA(p,q) processes. Thus, it satisfies the following properties:

i. $\forall X, Y, 0 = d(X, X) \leq d(X, Y) = d(Y, X)$.

ii. $\forall X, Y, d(X, Y) = 0$ if and only if $\pi'_y n = \alpha\pi'_x n, \forall n \in N$.

iii. $\forall X, Y, Z, d(X, Y) \leq d(X, Z) + d(Z, Y)$.

Proof 2.

i. $\forall X, Y, 1 = \mathrm{Aff}(X, X) \geq \mathrm{Aff}(X, Y) = \mathrm{Aff}(Y, X)$; then, from Equation (8), $0 = d(X, X) \leq d(X, Y) = d(Y, X)$.

ii. $\forall \mathrm{X}, \mathrm{Y}, \mathrm{Aff}(X, Y) = \pm 1$ if and only if $\pi'_y n = \alpha\pi'_x n, \forall n \in N$; then, from Equation (8), $d(X, Y) = 0$ if and only if $\pi'_y n = \alpha\pi'_x n, \forall n \in N$.

iii. $\forall X, Y, Z, \quad d(X, Y) = \lim \|\pi''^n_x - \pi''^n_y\|$ by definition where $\forall n \in N, \pi''^n_x = (\mathrm{sign}(\pi_{x,1}/\pi^n_{x.})\sqrt{|\pi_{x,1}/\pi^n_{x.}|}, \cdots, \mathrm{sign}(\pi_{x,n}/\pi^{\mathrm{n}}_{\mathrm{x.}})\sqrt{|\pi_{x,n}/\pi^n_{x.}|})$ and $\|.\|$ is the norm in R^n. It can be seen that $\|\pi''_x n - \pi''_y n\| = \|\pi''_x n - \pi''_y n - \pi''_z n + \pi''_z n\|$, and due to properties of the triangular inequality of norm in R^n and the notion of convergence of series, it follows

$$\mathrm{d(X, Y)} = \lim \left\| \pi''_x n - \pi''_y n - \pi''_z n + \pi''_z n \right\| \leq$$

$$\leq \lim \|\pi''_x n - \pi''_z n\| + \lim \|\pi''_{\mathrm{z}} \mathrm{n} - \pi''_{\mathrm{y}} \mathrm{n}\| = \mathrm{d(X, Z)} + \mathrm{d(Z, Y)}$$

because the limits exist. □

Therefore, it is proved above that the measure in Equation (8) is a distance measure between the zero mean ARMA(p,q) processes X and Y defined, as in Equation (1), satisfying the conditions of a distance if and only if $\pi'_y n \neq \alpha\pi'_x n, \forall n \in N$. This is the associated distance to the affinity coefficient between the two zero mean ARMA(p,q) processes X and Y. Note that in the case of AR(1) $\pi'_y n = \alpha\pi'_x n, \forall n \in N$ so in this case, the measure defined in Equation (8) is a semidistance.

The affinity coefficient was defined in Equation (5), and consequently, the distance measure defined in Equation (8) can have different expressions depending on the order of the processes. Thus, even though ARMA(1,1) models are relatively simple time series models, they are a very important case because they fit particularly well in many applications to real phenomena. Consequently, the following section presents the expression of the affinity coefficient in the case of ARMA(1,1) processes.

2.3. Affinity Coefficient for ARMA Model of Order One. In the case of the ARMA(1,1) and in the cases of AR(1) and MA(1), that is, when $q = 0$ and $p = 0$, respectively, the expression of the affinity coefficient is now presented. Consider X a zero-mean invertible ARMA(1,1) process defined, as in the Equation (1), satisfying the conditions i, ii, and iii with Gaussian white noise with constant variance σ^2 different from zero; then, in this case, the π- weights of

the AR (∞) formulation of ARMA(1,1) processes are given by [13]

$$\pi_{x,j} = \theta_x^{j-1}(\varphi x - \theta x),\ j = 1, 2 \tag{9}$$

Let Y be an ARMA(1,1) process in the same conditions of X, and then, it can be proved that

1. Aff $(X, Y) = 0$, if

i. $(\varphi x - \theta x) = 0$ and $(\varphi y - \theta y) \neq 0$,

ii. $(\varphi x - \theta x) \neq 0\ (\varphi y - \theta y) = 0$.

2. Aff $(X, Y) = 1$, if

i. $\theta x(\varphi x - \theta x) = \theta y(\varphi y - \theta y)$, $\forall j$,

ii. $\theta_x = 0$, $\theta_y = 0$, $\varphi_x \neq 0$, $\varphi_y \neq 0$, and sign $(\varphi x/|\varphi x|)$ sign $(\varphi y/|\varphi y|) > 0$,

and

3. Aff $(X, Y) = -1$, if

i. $\theta x = 0$, $\theta y = 0$, $\varphi x \neq 0$, $\varphi y \neq 0$, and sign $(\varphi x/|\varphi x|)$ sign $(\varphi y/|\varphi y|) > 0$.

Moreover

$$\text{Aff}\,(X, Y) = \frac{\sqrt{(1 - |\theta_x|)(1 - |\theta_y|)}}{\left(1 - \sqrt{|\theta_x\theta_y|}\right)} \tag{10}$$

if

i. $(\varphi_x - \theta_x) > 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y > 0$,

ii. $(\varphi_x - \theta_x) < 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y > 0$,

iii. $(\varphi_x - \theta_x) > 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y < 0$,

iv. $(\varphi_x - \theta_x) < 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y < 0$,

v. $\theta_x = 0$, $\varphi_x \neq 0$, $(\varphi_y - \theta_y) \neq 0$, $\theta_y \neq 0$ and sign $(\varphi x/|\varphi x|)$ sign $((\varphi y - \theta y)(1 - |\theta y|)/|\varphi y - \theta y|) > 0$,

vi. $\theta_y = 0$, $\varphi_y \neq 0$, $(\varphi_x - \theta_x) \neq 0$, $\theta_x \neq 0$ and sign $(\varphi y/|\varphi y|)$ sign $((\varphi x - \theta x)(1 - |\theta x|)/|\varphi x - \theta x|) > 0$,

and

$$\text{Aff}\,(X, Y) = -\frac{\sqrt{(1 - |\theta_x|)(1 - |\theta_y|)}}{\left(1 - \sqrt{|\theta_x\theta_y|}\right)} \tag{11}$$

if

i. $(\varphi_x - \theta_x) < 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y > 0$,

ii. $(\varphi_x - \theta_x) > 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y > 0$,

iii. $(\varphi_x - \theta_x) < 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y < 0$,

iv. $(\varphi_x - \theta_x) > 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y < 0$,

v. $\theta_x = 0$, $\varphi_x \neq 0$, $(\varphi_y - \theta_y) \neq 0$, $\theta_y \neq 0$ and sign $(\varphi x/|\varphi x|)$ sign $((\varphi y - \theta y)(1 - |\theta y|)/|\varphi y - \theta y|) < 0$,

vi. $\theta_y = 0$, $\varphi_y \neq 0$, $(\varphi_x - \theta_x) \neq 0$, $\theta_x \neq 0$ and sign $(\varphi y/|\varphi y|)$ sign $((\varphi x - \theta x)(1 - |\theta x|)/|\varphi x - \theta x|) < 0$.

Also

$$\text{Aff}\,(X, Y) = \frac{\sqrt{(1 - |\theta_x|)(1 - |\theta_y|)}}{\left(1 + \sqrt{|\theta_x\theta_y|}\right)} \tag{12}$$

if

i. $(\varphi_x - \theta_x) > 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y > 0$,

ii. $(\varphi_x - \theta_x) > 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y < 0$,

iii. $(\varphi_x - \theta_x) < 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y < 0$,

iv. $(\varphi_x - \theta_x) < 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y > 0$,

and finally,

$$\text{Aff}\,(X, Y) = -\frac{\sqrt{(1 - |\theta_x|)(1 - |\theta_y|)}}{\left(1 + \sqrt{|\theta_x\theta_y|}\right)} \tag{13}$$

if

i. $(\varphi_x - \theta_x) < 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y > 0$,

ii. $(\varphi_x - \theta_x) > 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y < 0$,

iii. $(\varphi_x - \theta_x) > 0$ and $\theta_x < 0$ and $(\varphi_y - \theta_y) < 0$ and $\theta_y > 0$,

iv. $(\varphi_x - \theta_x) < 0$ and $\theta_x > 0$ and $(\varphi_y - \theta_y) > 0$ and $\theta_y < 0$.

In the case of Equations (10) and (11), the affinity coefficient is unbounded if the product $\theta_x\theta_y$ is near one; however, if the product is one, X cannot be represented in terms of its past values according to the AR (∞) formulation so X does not satisfy the hypothesis.

Because the ARMA(1,1) models fit particularly well in many applications to real phenomena, it was applied the proposed extension of the affinity coefficient and its associated distance to the likelihood estimates of the coefficients of ARMA(1,1) processes, simulated with R, and agglomerative hierarchical cluster analysis is made. In the following section, the simulated results are presented and discussed.

3. Simulation Results

The affinity coefficient proposed and the associated distance for AR(1), MA(1), and ARMA(1,1) were implemented using the programming language R [38, 39]. The used packages were haven, ggplot2, dplyr, lubridate, forecast, TSclust, tidyverse, mefa4, data.table, gridExtra, clValid, stats, tibble, ggdendro, and dendextend with version 4.1.3. To apply this coefficient, AR of order one processes, moving average of order one processes, and ARMA of order one processes were generated with R and the likelihood estimates of the coefficients of the AR(1), MA(1), ARMA(1,1) processes, simulated with R, were obtained. After verifying how this coefficient behaves, the affinity coefficient of the likelihood estimates of the coefficients of AR(1), MA(1), and ARMA(1,1) processes was calculated and agglomerative hierarchical clustering analysis with the affinity coefficient approach was applied. To evaluate the results, it was chosen the average linkage as a method of aggregation between groups.

3.1. Clustering of ARMA Time Series: An Example. The ARMA processes of order one were generated using the programming language R in the following way: The time series S1 and S2 (Group 1) were created with the AR coefficient given by ar = 0, moving average coefficient given by ma = 0.3, and variance given by 0.0001. In the same way, the time series S3 and S4 (Group 2) were created with the coefficients, ar = 0, ma = −0.2, and variance given by 0.0001. In the third group, the series S5 and S6 (Group 3) were generated with the AR part given by ar = 0 and ma = −0.3 and variance given by 0.0001. The series S7 and S8 (Group 4) were generated with the coefficients, ar = 0, ma = 0.2, and variance given by 0.0001. The series S9 and S10 (Group 5) were generated with the coefficients, ar = 0.3, ma = 0.2, and variance given by 0.0001. The series S11 and S12 (Group 6) were generated with the coefficients, ar = 0.4, ma = 0.3, and variance given by 0.0001. The series S13 and S14 (Group 7) were generated with the coefficients, ar = 0.2, ma = 0.3, and variance given by 0.0001. The series S15 and S16 (Group 8) were generated with the coefficients, ar = 0.3, ma = −0.2, and variance given by 0.0001. The series S17 and S18 (Group 9) were generated with the coefficients, ar = 0.2, ma = −0.3, and variance given by 0.0001. The series S19 and S20 (Group 10) were generated with the coefficients, ar = 0.4, ma = −0.2, and variance given by 0.0001. The series S21 and S22 (Group 11) were generated with the coefficients, ar = 0.2, ma = 0.3, and variance given by 0.0001. All time series were created with a length of 500 samples each [6].

Table 1: Likelihood estimates of the coefficients of the generated autoregressive moving average processes.

Series	Autoregressive coefficients	Moving average coefficients
$\widehat{S}1$	0.00e + 00	2.25e − 01
$\widehat{S}2$	0.00e + 00	2.40e − 01
$\widehat{S}3$	0.00e + 00	−1.62e − 01
$\widehat{S}4$	0.00e + 00	−1.98e − 01
$\widehat{S}5$	0.00e + 00	−2.93e − 01
$\widehat{S}6$	0.00e + 00	−2.95e − 01
$\widehat{S}7$	0.00e + 00	2.17e − 01
$\widehat{S}8$	0.00e + 00	1.60e − 01
$\widehat{S}9$	3.16e − 01	1.61e − 01
$\widehat{S}10$	3.56e − 01	0.00e + 00
$\widehat{S}11$	4.59e − 01	3.08e − 01
$\widehat{S}12$	4.01e − 01	3.07e − 01
$\widehat{S}13$	1.89e − 01	3.13e − 01
$\widehat{S}14$	2.04e − 01	3.24e − 01
$\widehat{S}15$	0.00e + 00	1.23e − 01
$\widehat{S}16$	1.46e − 01	0.00e + 00
$\widehat{S}17$	0.00e + 00	−9.80e − 02
$\widehat{S}18$	0.00e + 00	−1.30e − 01
$\widehat{S}19$	1.86e − 01	0.00e + 00
$\widehat{S}20$	0.00e + 00	1.63e − 01
$\widehat{S}21$	2.88e − 01	2.32e − 01
$\widehat{S}22$	2.14e − 01	3.48e − 01

The likelihood estimates of all parameters of the generated time series were estimated using the programming language R. The likelihood estimates of the coefficients of the generated ARMA processes can be seen in Table 1. It applied agglomerative hierarchical clustering analysis with the distance associated to the affinity coefficient for the processes, given in Table 1, that is, for the estimated time series X and Y, which the autoregressive and moving average coefficients are the likelihood estimates of the coefficients of the generated AR(1), MA(1), ARMA(1,1) processes, and the results are in the following section [14].

3.2. Hierarchical Clustering Results With Affinity Associated Distance. For applying the agglomerative hierarchical clustering analysis, it calculated the affinity associated distance between the estimated time series X and Y. Since the distance depends on the π- weights of the AR (∞) formulations, that is, depending on the coefficients of the best linear forecasts of X and Y given your past, it is expected that processes with similar π- weights have high similarity and processes with distinct π- weights have low similarity. Because the π- weights depend on the autoregressive and moving average coefficients and the proposed affinity coefficient assigns the same weight to all π- weights, then it is expected that the MA(1) and ARMA(1,1) processes have

Table 2: Distances between the estimated autoregressive moving average processes.

	$\hat{S}$ 1	$\hat{S}$ 2	$\hat{S}$ 3	$\hat{S}$ 4	$\hat{S}$ 5	$\hat{S}$ 6	$\hat{S}$ 7	$\hat{S}$ 8	$\hat{S}$ 9	$\hat{S}$ 10	$\hat{S}$ 11	$\hat{S}$ 12	$\hat{S}$ 13	$\hat{S}$ 14	$\hat{S}$ 15	$\hat{S}$ 16	$\hat{S}$ 17	$\hat{S}$ 18	$\hat{S}$ 19	$\hat{S}$ 20	$\hat{S}$ 21	$\hat{S}$ 22
$\hat{S}$1	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$2	1.95e − 02	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$3	1.83e + 00	1.83e + 00	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$4	1.82e + 00	1.81e + 00	5.20e − 02	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$5	1.78e + 00	1.78e + 00	1.78e − 01	1.27e − 01	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$6	1.78e + 00	1.78e + 00	1.81e − 01	1.30e − 01	2.81e − 03	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$7	1.06e − 02	3.01e − 02	1.83e + 00	1.82e + 00	1.79e + 00	1.78e + 00	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$8	9.26e − 02	1.12e − 01	1.86e + 00	1.84e + 00	1.81e + 00	1.81e + 00	8.21e − 02	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$9	2.00e + 00	2.00e + 00	7.45e − 01	7.79e − 01	8.57e − 01	8.58e − 01	2.00e + 00	2.00e + 00	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$10	1.94e + 00	1.93e + 00	4.11e − 01	4.57e − 01	5.64e − 01	5.66e − 01	1.94e + 00	1.96e + 00	4.10e − 01	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$11	2.00e + 00	2.00e + 00	8.68e − 01	8.97e − 01	9.61e − 01	9.62e − 01	2.00e + 00	1.99e + 00	1.98e − 01	5.79e − 01	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$12	2.00e + 00	2.00e + 00	8.68e − 01	8.96e − 01	9.61e − 01	9.62e − 01	2.00e + 00	1.99e + 00	1.98e − 01	5.79e − 01	8.46e − 04	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$13	1.16e − 01	9.68e − 02	1.80e + 00	1.79e + 00	1.75e + 00	1.75e + 00	1.27e − 01	2.07e − 01	1.99e + 00	1.91e + 00	2.00e + 00	2.00e + 00	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$14	1.30e − 01	1.11e − 01	1.80e + 00	1.78e + 00	1.75e + 00	1.75e + 00	1.41e − 01	2.21e − 01	1.99e + 00	1.91e + 00	2.00e + 00	2.00e + 00	1.44e − 02	0.00e + 00	NA	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$15	1.49e − 01	1.68e − 01	1.87e + 00	1.86e + 00	1.82e + 00	1.82e + 00	1.38e − 01	5.65e − 02	2.00e + 00	1.97e + 00	1.98e + 00	1.98e + 00	2.62e − 01	2.76e − 01	0.00e + 00	NA	NA	NA	NA	NA	NA	NA
$\hat{S}$16	1.94e + 00	1.93e + 00	4.11e − 01	4.57e − 01	5.64e − 01	5.66e − 01	1.94e + 00	1.96e + 00	4.10e − 01	0.00e + 00	5.79e − 01	5.79e − 01	1.91e + 00	1.91e + 00	1.97e + 00	0.00e + 00	NA	NA	NA	NA	NA	NA
$\hat{S}$17	1.86e + 00	1.85e + 00	1.02e − 01	1.53e − 01	2.77e − 01	2.79e − 01	1.86e + 00	1.88e + 00	6.74e − 01	3.17e − 01	8.08e − 01	8.08e − 01	1.83e + 00	1.82e + 00	1.90e + 00	3.17e − 01	0.00e + 00	NA	NA	NA	NA	NA
$\hat{S}$18	1.84e + 00	1.84e + 00	4.91e − 02	1.01e − 01	2.26e − 01	2.29e − 01	1.85e + 00	1.87e + 00	7.12e − 01	3.66e − 01	8.40e − 01	8.40e − 01	1.81e + 00	1.81e + 00	1.88e + 00	3.66e − 01	5.27e − 02	0.00e + 00	NA	NA	NA	NA
$\hat{S}$19	1.94e + 00	1.93e + 00	4.11e − 01	4.57e − 01	5.64e − 01	5.66e − 01	1.94e + 00	1.96e + 00	4.10e − 01	0.00e + 00	5.79e − 01	5.79e − 01	1.91e + 00	1.91e + 00	1.97e + 00	0.00e + 00	3.17e − 01	3.66e − 01	0.00e + 00	NA	NA	NA
$\hat{S}$20	8.82e − 02	1.08e − 01	1.86e + 00	1.84e + 00	1.81e + 00	1.81e + 00	7.76e − 02	4.51e − 03	2.00e + 00	1.96e + 00	1.99e + 00	1.99e + 00	2.03e − 01	2.17e − 01	6.10e − 02	1.96e + 00	1.88e + 00	1.87e + 00	1.96e + 00	0.00e + 00	NA	NA
$\hat{S}$21	2.00e + 00	2.00e + 00	8.09e − 01	8.41e − 01	9.11e − 01	9.13e − 01	2.00e + 00	2.00e + 00	9.94e − 02	4.97e − 01	1.00e − 01	9.94e − 02	2.00e + 00	2.00e + 00	1.99e + 00	4.97e − 01	7.44e − 01	7.78e − 01	4.97e − 01	2.00e + 00	0.00e + 00	NA
$\hat{S}$22	1.60e − 01	1.41e − 01	1.79e + 00	1.77e + 00	1.74e + 00	1.74e + 00	1.71e − 01	2.51e − 01	1.98e + 00	1.90e + 00	2.00e + 00	2.00e + 00	4.49e − 02	3.05e − 02	3.05e − 01	1.90e + 00	1.82e + 00	1.80e + 00	1.90e + 00	2.46e − 01	1.99e + 00	0.00e + 00

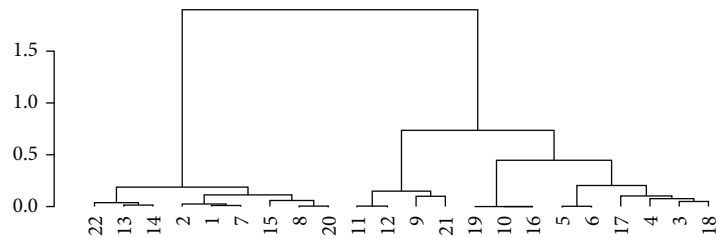

Figure 1: Dendrogram resulting from the hierarchical cluster analysis with average linkage method and affinity-associated distance.

similar behavior (grows and decreases at the same time); that is, similar profiles will be in the same cluster even the magnitude of the coefficients is different. Thus, it is expected that the associated distance to the affinity coefficient proposed depends less on the magnitude and more on the signal of the values of the π- weights of the AR (∞) formulations of the processes than the Euclidean distance. The associated distance between the estimated time series was calculated using Equation (8), that is, using the relationship

$$d(X, Y) = \sqrt{2(1 - \mathrm{Aff}(X, Y))}$$

and the matrix of distances between the estimated time series is in Table 2.

The results obtained after applying agglomerative hierarchical analysis with an average linkage method between groups are shown in dendrogram Figure 1.

From Figure 1, it may be observed the dendrogram has 21 levels. Reading the tree from left to right, we find six well-separated clusters of estimated time series given in Table 3.

With these six clusters and reading the tree from left to right, we find that the distance puts in the first cluster the estimated time series $\widehat{S}13$, $\widehat{S}14$, and $\widehat{S}22$ that are ARMA(1,1) processes with similar magnitude and positive signal of the moving average part and similar magnitude and negative signal of the difference between the autoregressive coefficient and the moving average coefficient. In this case, the π-weights have all negative signals (see Formula (9)). The second cluster comprises the estimated time series $\widehat{S}1$, $\widehat{S}2$, and $\widehat{S}7$ that are MA(1) processes with similar magnitude of the positive moving average part, and the difference between the autoregressive coefficient and the moving average coefficient has a negative signal because the AR coefficient is zero. Thus, the π- weights of these time series have all negative signals (see Formula (9)). The third group contains the estimated time series $\widehat{S}8$, $\widehat{S}15$, and $\widehat{S}20$ that are MA(1) processes with similar magnitude of positive moving average coefficients and which difference between the autoregressive coefficient and the moving average coefficient has a negative signal so the π- weights of these time series have all negative signals (see Formula (9)). Note that the coefficient magnitudes in this group are slightly different from those of the previous group. Thus, the distance puts at the same cluster MA(1) processes with similar forecasts and in different clusters MA(1) processes with small differences in the magnitude of the coefficients, that is, with small differences in their forecasts which is what is intended from the distance. Note that in the superior level, the time series of the previous two groups are at the same cluster, so the cluster analysis includes in the same cluster MA(1) processes with similar forecasts. The fourth cluster contains the estimated time series $\widehat{S}9$, $\widehat{S}11$, $\widehat{S}12$, and $\widehat{S}21$ that are ARMA(1,1) processes with similar magnitude and positive signal of the moving average part and similar magnitude and positive signal of the difference between the autoregressive coefficient and the moving average coefficient. Note that the distance first puts together the time series $\widehat{S}11$ and $\widehat{S}12$ that have very similar autoregressive and moving average coefficients and, after, includes together the time series $\widehat{S}9$ and $\widehat{S}21$ that are also ARMA(1,1) processes but with coefficient magnitudes slightly different from the time series $\widehat{S}11$ and $\widehat{S}12$. In this case, the π- weights have all positive signals (see Formula (9)); then, the forecasts are very different from the forecasts of the time series in the first cluster. The distance puts in different groups in the ARMA(1,1) processes with different π-weights, that is, with different forecasts. Thus, the distance is put at the same cluster ARMA(1,1) processes with similar forecasts and in different cluster ARMA(1,1) processes with different forecasts like expected. The fifth cluster comprises the estimated time series $\widehat{S}10$, $\widehat{S}16$, and $\widehat{S}19$ that are AR(1) processes for these time series; the π- weights have all positive signals (see Formula (9)). Even in this case, when the processes cannot be represented in terms of their past values according to the AR (∞) formulation, the measure puts together time series with similar forecasts. In the last cluster, the distance puts together the time series $\widehat{S}3$, $\widehat{S}4$, $\widehat{S}5$, $\widehat{S}6$, $\widehat{S}17$, and $\widehat{S}18$ that are MA(1) processes with negative coefficients, so the positive signal of the difference between the autoregressive coefficient and the moving average coefficient and

Table 3: Clusters obtained of the estimated autoregressive moving average processes with the affinity-associated distance.

Clusters	Time series in the clusters
Cluster 1	$\{\widehat{S}13, \widehat{S}14, \widehat{S}22\}$
Cluster 2	$\{\widehat{S}1, \widehat{S}2, \widehat{S}7\}$
Cluster 3	$\{\widehat{S}8, \widehat{S}15, \widehat{S}20\}$
Cluster 4	$\{\widehat{S}9, \widehat{S}11, \widehat{S}12, \widehat{S}21\}$
Cluster 5	$\{\widehat{S}10, \widehat{S}16, \widehat{S}19\}$
Cluster 6	$\{\widehat{S}3, \widehat{S}4, \widehat{S}5, \widehat{S}6, \widehat{S}17, \widehat{S}18\}$

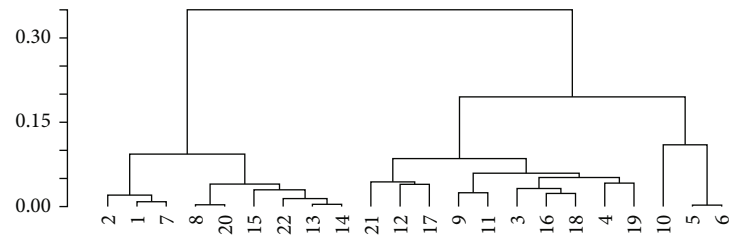

Figure 2: Dendrogram resulting from the hierarchical cluster analysis with average linkage method and Euclidean distance.

then the π- weights has alternate negative and positive signals; therefore, these MA(1) processes have very different forecasts when compared to the other MA(1) processes.

The cut-off occurs at level 20, where the hierarchy separates two well-defined clusters of the time series given by the groups {$\widehat{S}1$, $\widehat{S}2$, $\widehat{S}7$, $\widehat{S}8$, $\widehat{S}13$, $\widehat{S}14$, $\widehat{S}15$, $\widehat{S}20$, $\widehat{S}22$} and {$\widehat{S}3$, $\widehat{S}4$, $\widehat{S}5$, $\widehat{S}6$, $\widehat{S}9$, $\widehat{S}10$, $\widehat{S}11$, $\widehat{S}12$, $\widehat{S}16$, $\widehat{S}17$, $\widehat{S}18$, $\widehat{S}19$, $\widehat{S}21$}. Observing these two groups, it can be seen that the distance puts in the first group MA(1) processes and ARMA(1,1) processes which have negative signals of all π- weights, and in the second group, the distance puts AR(1) processes, ARMA(1,1) processes which have positive signals of all π- weights, and MA(1) processes which have alternate positive and negative signals of all π- weights.

The distance depends on the π- weights, and the π-weights depend on the signal and magnitude of the moving average's part and the signal and magnitude of the difference between the AR part and the moving average part (see Formula (9)). This distance is very sensitive to changes of the magnitude and signal on the π- weights and is very sensitive to different forecasts like expected. The results show that two processes with the same behavior of the forecast functions are similar with respect to this coefficient, and two processes with slightly different behavior of the forecast functions are different with respect to this coefficient.

3.3. Comparing Results With Euclidean Distance. For comparing the results obtained with the affinity coefficient/associated distance and those obtained with the Euclidean distance as proposed by Piccolo [13] in the agglomerative hierarchical clustering analysis, it calculated the Euclidean distance between the same estimated time series. The results obtained after applying agglomerative hierarchical analysis with an average linkage method between groups are shown in dendrogram (Figure 2). Looking at the dendrogram, it is possible to see in the hierarchy five clusters, given in Table 4.

With these five clusters, the cluster analysis includes the first cluster in the estimated time series $\widehat{S}1$, $\widehat{S}2$, and $\widehat{S}7$ that are MA(1) processes in which the π- weights of these time series have all negative signals. The second cluster comprises the estimated time series, $\widehat{S}8$, $\widehat{S}13$, $\widehat{S}14$, $\widehat{S}15$, $\widehat{S}20$, and $\widehat{S}22$. This second cluster contains ARMA(1,1) processes in which the π- weights have all negative signals and MA(1) processes in which π- weights have all negative signals. In the associated distance, these ARMA(1,1) and MA(1) processes were separated. Note that in the superior level, the time series of the previous two groups are in the same cluster obtained with the associated distance.

Table 4: Clusters obtained of the estimated autoregressive moving average processes with Euclidean distance.

Clusters	Time series in the clusters
Cluster 1	{$\widehat{S}1$, $\widehat{S}2$, $\widehat{S}7$}
Cluster 2	{$\widehat{S}8$, $\widehat{S}13$, $\widehat{S}14$, $\widehat{S}15$, $\widehat{S}20$, $\widehat{S}22$}
Cluster 3	{$\widehat{S}12$, $\widehat{S}17$, $\widehat{S}21$}
Cluster 4	{$\widehat{S}3$, $\widehat{S}4$, $\widehat{S}9$, $\widehat{S}11$, $\widehat{S}16$, $\widehat{S}18$, $\widehat{S}19$}
Cluster 5	{$\widehat{S}5$, $\widehat{S}6$, $\widehat{S}10$}

The third cluster puts together the estimated time series $\widehat{S}12$, $\widehat{S}17$, and $\widehat{S}21$ that are ARMA (1,1) with positive π-weights and MA (1) with negative and positive π- weights. Note that these ARMA(1,1) processes and the MA(1) process were separated in the associated distance. The associated distance puts together these time series at a higher level in the tree. In the fourth cluster, the Euclidean distance puts together the estimated time series $\widehat{S}3$, $\widehat{S}4$, $\widehat{S}9$, $\widehat{S}11$, $\widehat{S}16$, $\widehat{S}18$, and $\widehat{S}19$ that are AR(1) and ARMA(1,1) processes with positive π- weights and MA (1) processes with negative and positive π- weights. The associated distance puts AR(1), ARMA(1,1), and MA(1) processes together at a higher level in the tree. The fifth cluster contains MA(1) processes where the π- weights have alternate negative and positive signals and the AR(1) process.

Looking at the clusters provided by the affinity-associated distance and the Euclidean distance (Tables 2 and 4) it appears that the affinity approach better separates the different types of AR(1), MA(1), and ARMA(1,1) processes depending on the signal and size of the π- weights. Moreover, the affinity coefficient approach appears to be more responsive than the Euclidean distance to small differences in forecasting functions.

For better understanding, the dendrograms were compared using a tanglegram and the entanglement coefficient with the dendextend package of R language.

The entanglement coefficient was calculated and obtained the value 0.19 which means that dendrograms have some similar features but also have significant differences in their structures, as shown in the figure. Note that this coefficient can take values between zero for no entanglement, and one for full entanglement.

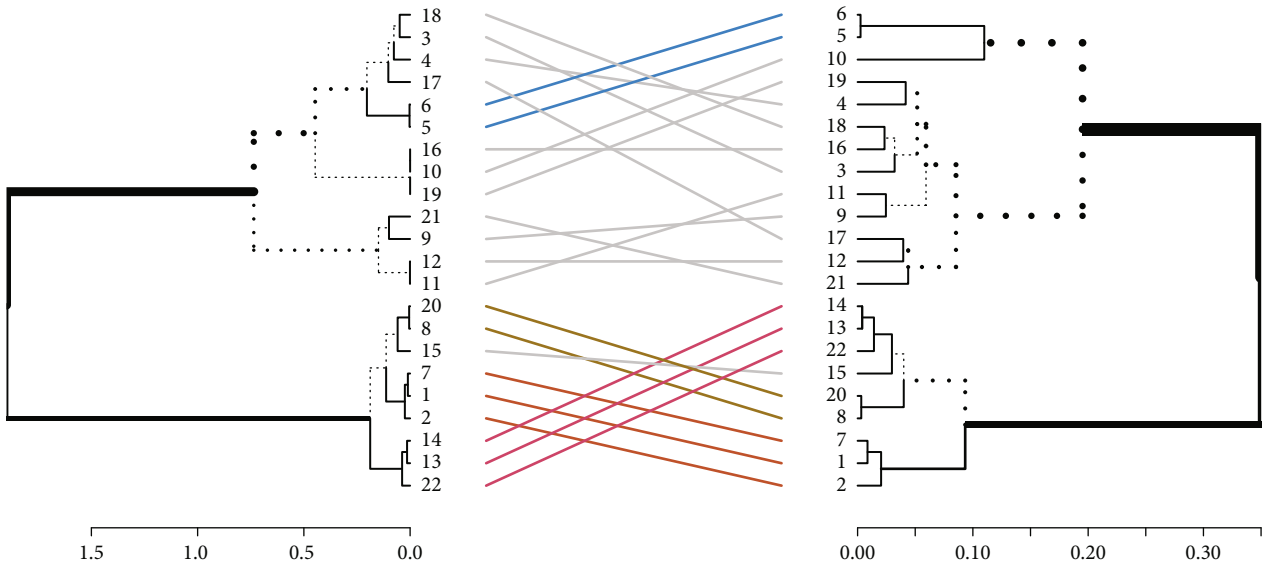

Figure 3: Comparison of dendrograms with the tanglegram function of R.

Figure 3 illustrates the comparison of the dendrograms.

4. Conclusions and Future Work

In this work, it proposed the extended affinity coefficient and associated distance, based on the π- weights of the AR (∞) formulations of the ARMA(p,q) processes, in the domain of time series cluster analysis. The affinity coefficient and associated distance can have different expressions depending on the order of the processes. Thus, in the case of the ARMA (1,1) and in the cases of AR (1) and MA (1), the expression of the affinity coefficient was presented because these models are very important cases, since they appear in many applications to real data. For these particular models and to verify how this coefficient behaves, the affinity coefficient proposed between the AR (1), MA (1), and ARMA(1,1) models was calculated and agglomerative hierarchical clustering analysis with the associated distance was applied. The results showed that this similarity coefficient considers the structure of the time series because it depends on the π- weights. In fact, the π- weights completely characterize the distribution of the process if the initial values, the orders, and the variance of the Gaussian white noise are given. The results show, as expected, that two processes with the same behavior of the forecast functions (same π- weights) are similar with respect to this coefficient. Moreover, if the two processes have at least an infinity number of π- weights with a symmetric signal, the affinity is also symmetric. It is also possible to conclude that this affinity coefficient is very sensitive to the behavior changes of the forecasting functions (same π-weights), because the associated distance is more sensitive to differences of the forecast functions, since it better separates the different types of AR(1), MA(1), and ARMA(1,1) of processes depending more on the sign and size of the π-weights than the Euclidean distance for the estimated time series. The affinity coefficient proposed, and the associated distance can also be implemented for ARMA(p,q) models and ARIMA models. An identified limitation of this coefficient is that for AR(1) processes, the associated distance is only a semidistance, so the measure in this case does not satisfy the uniqueness; however, in this case, the Euclidean distance is only bounded [13].

The results obtained when clustering time series with the affinity coefficient approach are important because they can be used in many applications to real phenomena in the areas of medicine and biotechnology. In fact, this approach can be applied in the analysis of data regarding biological signals such as ECG and EEG, noncommunicable diseases [40] or other data that can be analyzed in these terms. From this analysis, it is intended to contribute to the development of biotechnological tools to support the diagnosis and monitoring of various pathologies.

Finally—last but not least—from a methodological point of view, it should be noted that clustering models using affinity coefficient over the pi coefficients of the AR(∞) models, rather than time series themselves, allows us to compare time series of different lengths with different dates. We intend to use and develop this approach over the next works.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

The authors would like to acknowledge Escola Superior de Saúde, ESS–P.PORTO and Instituto de Saúde Ambiental, ISAMB-FMUL-U. Lisboa, for the support throughout this research.

Acknowledgments

The authors would like to thank Gilbert Saporta for his valuable remarks, which improved the paper. The authors would also like to acknowledge Escola Superior de Saúde, ESS–P.PORTO and Instituto de Saúde Ambiental, ISAMB-FMUL-U. Lisboa, for the support throughout this research.

References

[1] H. Bacelar-Nicolau, F. C. Nicolau, Á. Sousa, and L. Bacelar-Nicolau, "Measuring similarity of complex and heterogeneous data in clustering of large data sets," *Biocybernetics and Biomedical Engineering*, vol. 29, no. 2, pp. 9–18, 2009, http://hdl.handle.net/10400.3/2664.

[2] H. Bacelar-Nicolau, F. Costa Nicolau, Á. Sousa, and L. Bacelar-Nicolau, "Clustering of variables with a three-way approach for health sciences," *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, vol. 21, no. 4, pp. 435–447, 2014.

[3] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview, II," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 6, pp. 1942–4787, 2017.

[4] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[5] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier, "End-to-end deep representation learning for time series clustering: a comparative study," *Data Mining and Knowledge Discovery*, vol. 36, no. 1, pp. 29–81, 2022.

[6] J. Caiado, N. Crato, and D. Peña, "A periodogram-based metric for time series classification," *Computational Statistics & Data Analysis*, vol. 50, no. 10, pp. 2668–2684, 2006.

[7] V. Niennattrakul and C. A. Ratanamahatana, "On clustering multimedia time series data using K-means and dynamic time warping," in *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pp. 733–738, Seoul, Korea (South), 2007.

[8] D. Ge, N. Srinivasan, and S. M. Krishnan, "Cardiac arrhythmia classification using autoregressive modeling," *BioMedical Engineering OnLine*, vol. 1, no. 1, p. 5, 2002.

[9] J. Alagón, "Spectral discrimination for two groups of time series," *Journal of Time Series Analysis*, vol. 10, no. 3, pp. 203–214, 1989.

[10] E. A. Maharaj, P. D'Urso, and J. Caiado, *Time Series Clustering and Classification*, Chapman and Hall/CRC, 1st edition, 2019.

[11] S. Rani and G. Sikka, "Recent techniques of clustering of time series data: a survey," *International Journal of Computer Applications*, vol. 52, no. 15, pp. 1–9, 2012.

[12] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[13] D. Piccolo, "A distance measure for classifying Arima models," *Journal of Time Series Analysis*, vol. 11, no. 2, pp. 153–164, 1990.

[14] M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric," *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 1860–1872, 2008.

[15] E. A. Maharaj, "A significance test for classifying Arma models," *Journal of Statistical Computation and Simulation*, vol. 54, no. 4, pp. 305–331, 1996.

[16] E. A. Maharaj, "Cluster of time series," *Journal of Classification*, vol. 17, no. 2, pp. 297–314, 2000.

[17] E. A. Maharaj, A. M. Alonso, and P. D'Urso, "Clustering seasonal time series using extreme value analysis: an application to Spanish temperature time series," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 1, no. 4, pp. 175–191, 2015.

[18] M. G. Scotto, A. M. Alonso, and S. M. Barbosa, "Clustering time series of sea levels: extreme value approach," *Journal of Waterway, Port, Coastal, and Ocean Engineering*, vol. 136, no. 4, pp. 215–225, 2010.

[19] P. D'Urso, E. A. Maharaj, and A. M. Alonso, "Fuzzy clustering of time series using extremes," *Fuzzy Sets and Systems*, vol. 318, pp. 56–79, 2017.

[20] H. Bacelar-Nicolau and F. C. Nicolau, *Estatística e Análise de Dados Multivariados: Passado e Futuro*, I Congresso Anual Da Sociedade Portuguesa de Estatística, 1993.

[21] F. C. Nicolau and H. Bacelar-Nicolau, "Teaching and learning hierarchical clustering probabilistic models for categorical data," in *Proceedings of the 54nd Session of the International Statistical Institute*, pp. 346–349, Berlin, Germany, 2003.

[22] K. Matusita, "On the theory of statistical decision functions," *Annals of the Institute of Statistical Mathematics*, vol. 3, no. 1, pp. 17–35, 1951.

[23] H. Bacelar-Nicolau, "The affinity coefficient in cluster analysis," in *Methods of Operations Research*, M. J. K. Beckmann, K. W. Gaede, K. Ritter, and H. Schneeweiss, Eds., vol. 53, pp. 507–512, Verlag Anton Hain, 1985.

[24] H. Bacelar-Nicolau, "Two probabilistic models for classification of variables in frequency tables," in *Classification and Related Methods of Data Analysis*, H. H. Bock, Ed., pp. 181–186, Elsevier Sciences Publishers B.V, 1988.

[25] H. Bacelar-Nicolau, *Contribuição ao estudo dos coeficientes de comparação em Análise Classificatória*, Universidade de Lisboa, 1980.

[26] I. C. Lerman, *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering. Series*, Springer-Verlag, 1st edition, 2016.

[27] H. Bacelar-Nicolau, "The affinity coefficient," in *Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data*, E. Bock and H. H. Diday, Eds., pp. 160–165, Springer-Verlag, 2000.

[28] H. Bacelar-Nicolau, "On the generalised affinity coefficient for complex data," *Biocybernetics and Biomedical Engineering*, vol. 22, no. 1, pp. 31–42, 2002.

[29] H. Bacelar-Nicolau, F. Nicolau, A. Sousa, and L. Bacelar-Nicolau, "Clustering complex heterogeneous data using a probabilistic approach," in *Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010)*, pp. 85–93, Greece, 2010.

[30] Á. Sousa, H. Bacelar-Nicolau, F. C. Nicolau, and O. Silva, "Clustering of symbolic data based on affinity coefficient: application to a real data set," *Biometrical Letters*, vol. 50, no. 1, pp. 27–38, 2013.

[31] A. L. da Silva, G. Saporta, and H. Bacelar-Nicolau, "Missing data and imputation methods in partition of variables," in *Classification, Clustering, and Data Mining Applications*, D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, Eds., pp. 631–637, Springer, Berlin Heidelberg, 2004.

[32] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Computational Statistics & Data Analysis*, vol. 55, no. 9, pp. 2579–2589, 2011.

[33] T. Silveira Gontijo and M. Azevedo Costa, "Forecasting hierarchical time series in power generation," *Energies*, vol. 13, no. 14, p. 3722, 2020.

[34] H. Bacelar-Nicolau, "On the distribution equivalence in cluster analysis," in *Pattern Recognition Theory and Applications*, P. A. Devijver and J. Kittler, Eds., pp. 73–79, Springer, Berlin Heidelberg, 1987.

[35] F. C. Nicolau and H. Bacelar-Nicolau, *Some Trends in the Classification of Variables BT-Data Science, Classification, and Related Methods*, C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, and Y. Baba, Eds., Springer, Japan, 1998.

[36] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, John Wiley and Sons Inc., 5th edition, 2015.

[37] E. Gonçalves and N. M. Lopes, *Séries temporais: Modelações lineares e não lineares*, S. P. de Estatística, 2 edition, 2008.

[38] A. M. Brandmaier, "pdc: AnRPackage for complexity-based clustering of time series," *Journal of Statistical Software*, vol. 67, no. 5, pp. 1–23, 2015.

[39] P. Montero and J. A. Vilar, "TSclust: an R package for time series clustering," *Journal of Statistical Software*, vol. 62, no. 1, pp. 1–43, 2014.

[40] A. P. Nascimento, C. Prudêncio, M. Vieira, R. Pimenta, and H. Bacelar-Nicolau, "A typological study of Portuguese mortality from non-communicable diseases," *Advances in Science, Technology and Engineering Systems*, vol. 5, no. 5, pp. 613–619, 2020.