Machine Learning with Applications 10 (2022) 100417

Contents lists available at ScienceDirect

Machine Learning with Applications

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

INGARCH-based fuzzy clustering of count time series with a football application

Roy Cerqueti [a,b,c,1], Pierpaolo D'Urso [a,1], Livia De Giovanni [d,*,1], Raffaele Mattera [a,1], Vincenzina Vitale [a,1]

[a] Department of Social and Economic Sciences, Sapienza University of Rome, Italy
[b] School of Business, London South Bank University, United Kingdom
[c] GRANEM, Université d'Angers, France
[d] Department of Political Sciences, LUISS Guido Carli, Rome, Italy

ARTICLE INFO

Keywords:
Fuzzy C-medoids
INGARCH
Poisson distribution
Sport analytics

ABSTRACT

Although there are many contributions in the time series clustering literature, few studies still deal with count time series data. This paper aims to develop a fuzzy clustering procedure for count time series data. We propose an Integer GARCH-based Fuzzy C-medoids (INGARCH-FCMd) method for clustering count time series based on a Mahalanobis distance between the parameters estimated by an INGARCH model. We show how the proposed clustering method works by clustering football teams according to the number of scored goals.

1. Introduction

Clustering is an unsupervised learning technique used to find similar structures in a given dataset, so it is commonly implemented for pattern recognition. The clustering task becomes more arduous when the dataset includes time series rather than cross-sectional objects. As claimed by Liao (2005), most time series clustering methods try to modify standard – hierarchical and non-hierarchical – clustering methods so that time series data can be appropriately handled. In particular, how the dissimilarity among time series is computed is of great importance.

Following Maharaj et al. (2019), we can distinguish between three approaches: (a) observation-based, (b) feature-based, and (c) model-based. The observation-based approaches define the dissimilarities among time series considering actual data, whereas the future-based ones compute them considering interesting features extracted from the time series, such as the autocorrelation function (ACF) (D'Urso & Maharaj, 2009), periodogram (Caiado et al., 2006, 2020), cepstral coefficients (D'Urso et al., 2020; Savvides et al., 2008) or, more recently, quantile autocovariance (Lafuente-Rego et al., 2020; Vilar et al., 2018) and quantile spectral cross-spectral density (López-Oriona & Vilar, 2021; López-Oriona et al., 2022). The last approach, i.e. model-based clustering, computes the dissimilarities among time series considering the proximity among the fitted statistical models, for example, based on the estimated parameters. Well known examples are given by the ARIMA (Piccolo, 1990), GARCH (D'Urso et al., 2016; Otranto, 2008) and, more recently, the GAS (Cerqueti et al., 2022, 2021). A great advantage of the model-based approaches is that the time series involved in the clustering process is not required to have the same length. Furthermore, if the statistical model is correctly specified, there is evidence showing that model-based approaches provide excellent performances in correctly classifying the time series (Díaz & Vilar, 2010).

Although there are many contributions in the literature on time series clustering, few studies still deal with count time series data. Count time series arise in many real-life problems. Some examples are the emergency call arrivals (Matteson et al., 2011), the number of transactions in the stock market (Rydberg & Shephard, 2000), the number of goals scored by a football team (Angelini & De Angelis, 2017) and so on. These time series are based on counts so that they are integer-valued.

The most popular approaches for modelling count time series belong to the class of observation-driven models (Cox et al., 1981), where the observed counts are modelled considering lagged observations in the conditional mean function and lagged counts. Assuming that counts are conditionally Poisson distributed, a well-known statistical model used for modelling and predicting count time series is the Integer GARCH (INGARCH) (Ferland et al., 2006), also called Autoregressive Conditional Poisson (ACP) model (Fokianos et al., 2009). Since the conditional mean of a Poisson distribution equals the conditional variance, the resulting statistical model mimics a GARCH process.

[*] Corresponding author.
E-mail addresses: roy.cerqueti@uniroma1.it (R. Cerqueti), pierpaolo.durso@uniroma1.it (P. D'Urso), ldegiovanni@luiss.it (L. De Giovanni), raffaele.mattera@uniroma1.it (R. Mattera), vincenzina.vitale@uniroma1.it (V. Vitale).
[1] The authors have provided equal contribution(s) to the paper.

https://doi.org/10.1016/j.mlwa.2022.100417
Received 12 July 2022; Received in revised form 20 September 2022; Accepted 21 September 2022
Available online 27 September 2022

In this paper, a new fuzzy clustering method is proposed for count time series data. Two main contributions can be highlighted. First, following a k-medoids approach, we propose a model-based clustering method based on the INGARCH process, one of the most popular approaches for modelling this type of data. Second, we introduce fuzziness in the clustering method. The adoption of a fuzzy clustering approach introduces uncertainty in the clustering process. Indeed, fuzzy clustering allows a time series to be allocated to two or more clusters with a given level of uncertainty represented by the so-called membership degree. Identifying a clear boundary between clusters is not easy in many real-world problems, so the membership degree highlights whether a second-best cluster is possible. Traditional clustering methods are not able to highlight such conclusions. As a result, we propose an INGARCH-based Fuzzy C-medoids (INGARCH-FCMd) method for clustering count time series based on a Mahalanobis distance computed considering the parameters estimated by an INGARCH model.

We show how the proposed clustering method works by providing an application to football data. In particular, we study the problem of clustering football teams considering the number of scored goals. Applying quantitative methods to sports data raised the interest of statistics and machine learning research communities. Statistical methods are nowadays commonly used for predicting matches' results (Angelini & De Angelis, 2017; Mattera, 2021), pricing players' value (Behravan & Razavi, 2021) and for evaluating teams performances (Sarlis & Tjortjis, 2020). Clustering gives another significant application of machine learning techniques for sports analytics (D'Urso et al., 2022; Narizuka & Yamazaki, 2019; Ulas, 2021).

Scored goals are integer-valued count data, and there is a long tradition in modelling such a variable through the Poisson process. For example, Maher (1982) proposed using Poisson distributions to model the number of goals scored by teams in a football match. More recently, many authors considered relaxing the hypothesis under which previous outcomes, i.e. the goals scored in previous matches, do not affect the current ones – that is, the hypothesis of independence over time. With this respect, recent studies (e.g. see Angelini & De Angelis, 2017; Koopman & Lit, 2015) modelled the scored goals by means of dynamic Poisson processes. Similarly, following this strand of literature, in this paper we consider the number of scored goals in different matches as count time series data.

The number of scored goals is an important indicator for discriminating against football teams since it is an important performance indicator. Furthermore, the final match outcome depends on the number of goals scored by the teams. Hence, forecasting how many goals will be scored by the teams is crucial for predicting what will be the final match outcome. Scored goals are a count variable, so, using past observations, previous literature (e.g. see Angelini & De Angelis, 2017; Koopman & Lit, 2015) used count time series models for predicting the scored goals that there will occur in the future matches. Since the number of scored goals is a count time series following a Poisson distribution, the developed INGARCH-based clustering approach is well suited for clustering these kinds of time series data.

The paper structure is the following. In Section 2.1, we discuss the INGARCH model, while in Section 2.2, we present the fuzzy clustering procedure adopted in the paper. In Section 3, we show the application of the model to real data, and in the last Section 4, we conclude with some final remarks and future possible research directions.

2. Clustering of count time series

The following section presents the proposed clustering method. First, the INGARCH process is introduced in Section 2.1; then, the clustering method is presented and discussed in detail in Section 2.2.

2.1. The INGARCH process

Let use denote $\{x_t : t = 1, \dots, T\}$ as a count time series. By assuming that x_t follows a Poisson distribution, we have that $\mathrm{E}\left(x_t \mid \mathcal{F}_{t-1}\right) = \lambda_t$ is the conditional mean of the count time series process, with $\mathcal{F}_{t-1}$ defined as a combination of the lagged values of both x_t and λ_t. In the case of Poisson distribution the conditional mean equals the conditional variance, i.e. $\mathrm{E}\left(x_t \mid \mathcal{F}_{t-1}\right) = \mathrm{Var}\left(x_t \mid \mathcal{F}_{t-1}\right) = \lambda_t$, so the following count time series process:

$$\begin{cases} x_t \mid \mathcal{F}_{t-1} : \mathcal{P}\left(\lambda_t\right) \\ \lambda_t = \gamma_0 + \sum_{i=1}^{q} \gamma_i x_{t-i} + \sum_{j=1}^{p} \delta_j \lambda_{t-j} \end{cases} \tag{1}$$

is commonly called INGARCH(p,q) (Ferland et al., 2006) with parameters $\gamma_0 > 0, \gamma_i \geqslant 0, i = 1, \dots, q, \delta_j \geqslant 0, j = 1, \dots, p$, because its structure parallels the one of the GARCH(p,q) model (Bollerslev, 1986). The (unconditional) long-run mean of the process $\{x_t : t = 1, \dots, T\}$ is (see Ferland et al., 2006):

$$\lambda = \frac{\gamma_0}{1 - \sum_{i=1}^{q} \gamma_i - \sum_{j=1}^{p} \delta_j} \tag{2}$$

the parameters γ_i, δ_j and γ_0 have to be non-negative, while the model is stationary if the sum of the γ_i and δ_j parameters is less than 1. The INGARCH model is structurally equivalent to the Autoregressive Conditional Poisson (ACP) model (Fokianos et al., 2009). The INGARCH(p,q) process can also be estimated by a procedure that resembles the one used for the traditional GARCH models, i.e. based on Maximum Likelihood Estimation (MLE). The likelihood function of the T observations $x_1, \dots, x_T$ is given by:

$$L(\Theta) = \prod_{t=1}^{T} \frac{\mathrm{e}^{-\lambda_t} \lambda_t^{x_t}}{x_t!} \tag{3}$$

with:

$$\Theta = \left(\gamma_0, \gamma_1, \dots, \gamma_p, \delta_1, \dots, \delta_q\right)' \tag{4}$$

In other words, Θ is the vector containing the static parameters in (1). According to (1), the associated log-likelihood function equals to:

$$\mathcal{L}(\Theta) = \sum_{t=1}^{T} \ell_t(\Theta) = \sum_{t=1}^{T} \left[x_t \log \lambda_t - \lambda_t\right] \tag{5}$$

Additional details about solutions can be found in Ferland et al. (2006) and Fokianos et al. (2009). Given a large enough T, the maximum likelihood estimator $\hat{\Theta}$ follows the following Normal distribution:

$$\hat{\Theta} \sim \mathcal{N}\left(\Theta_0, T^{-1} \mathfrak{I}\left(\Theta_0\right)^{-1}\right) \tag{6}$$

where $E[\Theta] = \Theta_0$ and $\mathfrak{I}\left(\Theta_0\right)$ is the information matrix evaluated at Θ_0. The inverse of information matrix equals the parameters' covariance matrix. Therefore, the standard errors can be calculated and parameters' inference is possible (see Ferland et al., 2006; Fokianos et al., 2009). Despite the generality of the INGARCH(p,q) model, we have to note that important empirical evidence shows that very parsimonious models can achieve an adequate modelling of many real count time series. Among them, a relevant role is played by the case $p = q = 1$ (e.g. see Agosto et al., 2016; Agosto & Giudici, 2020; Aknouche et al., 2021; Chen & Lee, 2017; Lee & Lee, 2019; Xiong & Zhu, 2019). The INGARCH$(1,1)$ model has a simple structure, a parsimonious parametrization and is faster to be estimated. Moreover, its statistical properties have a clear interpretation. For all these reasons, it is widely studied in literature. The INGARCH$(1,1)$ can be written as follows:

$$\begin{cases} x_t \mid \mathcal{F}_{t-1} : \mathcal{P}\left(\lambda_t\right) \\ \lambda_t = \gamma_0 + \gamma_1 x_{t-1} + \delta_1 \lambda_{t-1} \end{cases} \tag{7}$$

for which many proprieties are well understood (Ferland et al., 2006). For example, the autocovariance function is given by:

$$\gamma(r) = \frac{\gamma_1\left(1 - \delta_1\left(\gamma_1 + \delta_1\right)\right)\left(\gamma_1 + \delta_1\right)^{r-1} \mu}{1 - \left(\gamma_1 + \delta_1\right)^2}, \quad \forall r \geqslant 1 \tag{8}$$

The autocovariance function (8) indicates that INGARCH$(1,1)$ is also an ARMA$(1,1)$ process. In particular, a generic INGARCH$(1,1)$ process can be written as the following ARMA$(1,1)$ (Ferland et al., 2006):

$$\left(x_t - \lambda\right) - \left(\gamma_1 + \delta_1\right)\left(x_{t-1} - \lambda\right) = e_t - \delta_1 e_{t-1}, \tag{9}$$

with e_t a white noise process of variance $\sigma^2 = \lambda = \gamma_0 / \left(1 - \gamma_1 - \delta_1\right)$.

This feature is important for two reasons. First of all, it suggests that the INGARCH model admits an $AR(\infty)$ representation. Second, it means that the parameters can be easily estimated with the usual Conditional Least Square (CLS) approach (Ferland et al., 2006; Fokianos et al., 2009). However, Fokianos et al. (2009) demonstrated with simulations that CLS provides larger MSE than MLE, thus representing a less attractive option for estimation. For this reason, we adopt MLE estimation strategy also in the case of the INGARCH$(1,1)$.

2.2. The fuzzy clustering method

In what follows we discuss the problem of measuring the dissimilarity among count time series data and the clustering method adopted. Let:

$$\mathbf{X} = \{x_{i,t} : i = 1,\dots,N; t = 1,\dots,T\} \tag{10}$$

be the matrix containing the $N(i = 1,\dots,N)$ count time series of length $T(t = 1,\dots,T)$. Following a model-based approach, we assume that the time series are conditionally Poisson distributed, so they are generated by distinct INGARCH processes that differentiate for their estimated parameters. Since the INGARCH resembles a standard GARCH model, it is reasonable to adopt a GARCH-type distance for measuring the dissimilarity among count time series.

Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be generic count time series belonging to $\mathbf{X}$. Suppose we fit an INGARCH$(1,1)$ model (7) to both time series, storing the estimated parameters in two vectors $\mathbf{T}_i = \left(\hat{\gamma}_{0,i}, \hat{\gamma}_{1,i}, \hat{\delta}_{1,i}\right)'$ and $\mathbf{T}_j = \left(\hat{\gamma}_{0,j}, \hat{\gamma}_{1,j}, \hat{\delta}_{1,j}\right)'$. Arguments justifying the use of a INGARCH$(1,1)$ model are given by its parsimonious parametrization and good performances in many real-world examples. Following the (Caiado & Crato, 2010) approach, we propose a Mahalanobis distance between the features of the count time series $x_{i,t}$ and $x_{j,t}$, called the INGARCH-based distance, which is defined as follows:

$$d_{\text{INGARCH}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{\left(\mathbf{T}_i - \mathbf{T}_j\right)' \boldsymbol{\Omega}_{i,j}^{-1}\left(\mathbf{T}_i - \mathbf{T}_j\right)}, \tag{11}$$

where $\boldsymbol{\Omega}_{i,j} = \hat{\mathbf{V}}_i + \hat{\mathbf{V}}_j$ is a weighting matrix, with $\hat{\mathbf{V}}_i$ and $\hat{\mathbf{V}}_j$ being associated to the variability of the estimated parameters for the time series i and j. In this paper, we consider $\hat{\mathbf{V}}_i$ and $\hat{\mathbf{V}}_j$ as two diagonal matrices containing the standard errors associated with the estimated parameters on the main diagonal. In this way, the matrix $\boldsymbol{\Omega}_{i,j}^{-1}$ weights the parameters by taking into account the uncertainty in their estimation. More in detail, the Mahalanobis distance (11) implicitly assigns lower weight to the parameters showing higher variability and higher weights for those with lower variability. Moreover, an important property of the distance (11) is that it does not require that the two time series are of equal length since it is based on estimated parameters.

Following a Partition Around Medoids (PAM) approach, we propose a clustering method called INGARCH-Fuzzy C-medoids (INGARCH-FCMd), that is based on the fuzzy C-medoids (FCMd) (Krishnapuram et al., 2001, 1999). The proposed clustering model can be formalized as follows:

$$\begin{cases} \min : \displaystyle\sum_{i=1}^{N}\sum_{c=1}^{C} u_{i,c}^{m} d_{\text{INGARCH}}^{2}\left(\mathbf{x}_i, \mathbf{x}_c j\right) \\ \quad = \displaystyle\sum_{i=1}^{N}\sum_{c=1}^{C} u_{i,c}^{m}\left[\left(\mathbf{T}_i - \mathbf{T}_c\right)' \boldsymbol{\Omega}_{i,c}^{-1}\left(\mathbf{T}_i - \mathbf{T}_c\right)\right] \\ \text{s.t.} \quad \displaystyle\sum_{c=1}^{C} u_{i,c} = 1 \quad \text{and} \quad u_{i,c} \geq 0, \quad \forall i, \forall c \end{cases} \tag{12}$$

where $d_{\text{INGARCH}}^{2}\left(\mathbf{x}_i, \mathbf{x}_c\right)$ is the squared Mahalanobis distance, as defined in (11), between the ith time series and the medoid time series of the cth cluster; $u_{i,c}$ denotes the membership degree of the ith time series to the cth cluster, the parameter $m > 1$ controls for the fuzziness of the partition. The optimal solution for the membership degree of the model (12) is given by Maharaj et al. (2019):

$$u_{i,c} = \frac{1}{\sum_{c'=1}^{C}\left[\frac{\left[(\mathbf{T}_i - \mathbf{T}_c)' \boldsymbol{\Omega}_{i,c}^{-1}(\mathbf{T}_i - \mathbf{T}_c)\right]}{\left[(\mathbf{T}_i - \mathbf{T}_{c'})' \boldsymbol{\Omega}_{i,c'}^{-1}(\mathbf{T}_i - \mathbf{T}_{c'})\right]}\right]^{\frac{1}{m-1}}} \tag{13}$$

Differently from the C-means clustering technique, the prototype of each cluster in the Fuzzy C-Medoids-based method is a real time series. Indeed, in the C−means algorithm, the cth cluster prototype is a fictitious time series, equal to the average of the real time series included in the cth cluster. This feature makes the partition obtained with the PAM algorithm more interpretable. Moreover, the same property makes the partition obtained with the PAM timid robust than c−means algorithm (D'Urso et al., 2018, 2021; Garcia-Escudero & Gordaliza, 2005; García-Escudero et al., 2003) because the real time series, i.e. the cth cluster medoid, is less influenced by outliers.

The main drawback of the non-hierarchical clustering algorithms lies in the prior selection of the number of clusters C. To overcome this limitation, following previous literature, we consider the use of cluster validity indices. Indeed, the cluster validity indices guide the users in selecting the number of clusters. Because of its particularly satisfactory results in recognizing the true number of clusters (for a reference, see the extensive simulations carried out in Arbelaitz et al., 2013), we select the optimal C according to the Fuzzy Silhouette criterion of Campello and Hruschka (2006), that is a fuzzy version of the Average Silhouette Width (ASW) criterion (Kaufman & Rousseeuw, 1990). The Silhouette measures the cohesion and separation of a partition, and it is computed as follows:

$$S_i = \frac{(b_i - a_i)}{\max\{b_i, a_i\}} \tag{14}$$

The value a_i is the average distance of the ith unit to the other units belonging to the same cluster. Then, letting $\bar{d}_{i,c'}$ be the average distance of i to all objects belonging to another cluster $c' \neq c$, we can define b_i as the minimum of the $\bar{d}_{i,c'}$ computed over $c' = 1,\dots,C; c' \neq c$, which represents the distance of i to others units belonging to the closest different cluster.

Therefore, a considerable Silhouette value S_i means that the ith unit is closer to those belonging to its cluster than the others belonging to the nearest different cluster. By averaging the S_i, we obtain the Average Silhouette Width (ASW), a synthetic value used to choose the best partition. The Fuzzy Silhouette proposed by Campello and Hruschka (2006) is a fuzzy version of the ASW, which considers a weighted average for the Silhouettes S_i with the membership degrees $u_{i,c}$ used as weights:

$$FS = \frac{\sum_{i=1}^{N}(u_{i,c} - u_{i,c'})^{\alpha} S_i}{\sum_{i=1}^{N}(u_{i,c} - u_{i,c'})^{\alpha}} \tag{15}$$

where S_i is the Silhouette computed as in (14), $u_{i,c}$ and $u_{i,c'}$ are the first and second-largest elements of the ith row of the fuzzy partition matrix, respectively, and $\alpha \geq 1$ a positive constant. The higher the value of the FS, the better the partition. Therefore, following common practice, we use the FS for choosing the optimal number of clusters C.

3. Application to Italian football data

In this section we present an empirical experiment with football data. In particular, we aim at clustering football clubs in the Italian Serie A on the basis of the scored goals' time series. Section 3.1 describes the data, Section 3.2 presents two benchmark clustering algorithms, while Section 3.3 discusses the results.

Table 1
Sample of Italian football teams.

Team	Number of matches
Fiorentina	380
Juventus	380
Atalanta	380
Chievo	266
Genoa	380
Milan	380
Roma	380
Bologna	342
Torino	380
Cagliari	342
Inter	380
Lazio	380
Napoli	380
Parma	228
Sampdoria	380
Udinese	380
Verona	266
Sassuolo	342

3.1. Data

The number of scored goals is an essential variable in football as it is a proxy for the football team's performance over time and offencive power. In this context, clustering allows us to deeply understand the similarities of the football teams in terms of performances over time. Moreover, identifying similarities in terms of scored goals could also be exploited regarding betting strategies.

As highlighted by previous studies (e.g. see Greenhough et al., 2002; Groll et al., 2018; Maher, 1982) the number of scored goals can be modelled as Poisson distribution. The main problem of using static approaches is that we do not take into account the serial dependence, which is present when modelling the number of scored goals in a football match (e.g. Angelini & De Angelis, 2017; Koopman & Lit, 2015). Time dependence in the number of scored goals can be explained by the periods characterized by high team performances and vice-versa (e.g. phases of the team's sporting cycle Mourao, 2016).

As an empirical experiment, we consider the number of goals scored by the football teams that participated in the last 10 Serie A seasons.[2] Furthermore, we excluded the sample teams relegated or promoted multiple times during the previous ten years. Indeed, teams that are absent from the championship for many years show missing values which cannot be reasonably imputed or estimated. The final sample is shown in Table 1.

As an example of the collected time series, Fig. 1 shows the Roma Calcio time series of scored goals.

As shown in Fig. 1, the time series is characterized by serial dependence in the mean, which is reasonably not constant. Note also that the number of observed matches is not the same among the considered football teams (see Table 1). For this reason, we should find a suitable approach that allows clustering count time series with different lengths. The INGARCH-based clustering approach proposed so far fits well with this scope since, as already said, it does not require the time series to be of the same length.

3.2. Benchmark clustering methods

Many conventional approaches (e.g. raw-data based, such as the simple Euclidean distance or correlation-based) require the time series to have the same lengths. The model-based clustering approaches, based on the estimated parameters, overcome this issue. To the best of our knowledge, no fuzzy clustering algorithms are explicitly taught to deal with count time series data. None particularly deal with unequally sized count time series. However, some existing fuzzy clustering methods could be adopted for this aim.

First of all, considering the class of observation-based approaches, the standard FCMd algorithm (Krishnapuram et al., 2001) with Euclidean distance among the counts can be used in principle. However, the simple Euclidean distance is not feasible for studying the football dataset of this empirical experiment because the time series have different lengths. Therefore, a feasible observational-based fuzzy clustering method has to be based on the Dynamic Time Warping distance (DTW, see Berndt & Clifford, 1994).

The DTW distance allows considering similar two time series showing similar behaviour in different periods. Let $\mathbf{x}_i$ $(t_i = 1, \dots, T_i)$ and $\mathbf{x}_j$ $(t_j = 1, \dots, T_j)$ be two sequences with $T_i \gtreqless T_j$. A so-called warping path is used to align the elements of the sequences such that their distance is minimized. Let $d(x_{i,r}, x_{j,s})$ be the Euclidean distance between two points r and s of the sequences $\mathbf{x}_i$ and $\mathbf{x}_j$, the DTW distance is given by the optimal alignment obtained with the minimization of the following the cumulative distance:

$$\Delta(r, s) = d(x_{i,r}, x_{j,s}) + \min\left[\Delta(r-1, s-1), \Delta(r-1, s), \Delta(r, s-1)\right] \tag{16}$$

Thus, the first benchmark that we consider is the Fuzzy DTW-FMCd method (see Izakian et al., 2015). The clustering problem can be defined as follows:

$$\begin{cases} \min: \displaystyle\sum_{i=1}^{N}\sum_{c=1}^{C} u_{i,c}^{m} D_{\mathrm{DTW}}^{2}\left(\mathbf{x}_i, \mathbf{x}_c\right) \\ \text{s.t.} \displaystyle\sum_{c=1}^{C} u_{i,c} = 1 \quad \text{and} \quad u_{i,c} \geq 0, \quad \forall i, \forall c \end{cases} \tag{17}$$

with $D_{\mathrm{DTW}}^{2}\left(\mathbf{x}_i, \mathbf{x}_c\right)$ be the squared DTW distance between the ith time series and the cth cluster medoid. The main problem of DTW-based partitional clustering methods is that averaging – which is required for the definition of clusters' prototypes – is not straightforward (Petitjean et al., 2011) because it has to be consistent with temporal alignment provided by DTW distance. However, the DTW-FMCd overcomes this problem being the cth time series prototype a time-varying unit really observed in the dataset, so that averaging is not required (Izakian et al., 2015).

Among the class of feature-based clustering methods, two commonly employed approaches are those based on the pairwise correlation coefficient (e.g. Mantegna, 1999) or on the Auto Correlation Function (ACF, e.g. see D'Urso & Maharaj, 2009). In our empirical setting, the approach based on the pairwise correlation coefficient is not feasible due to the difference in the time series lengths. Therefore, we consider the ACF-based clustering method (D'Urso & Maharaj, 2009) as the second benchmark of our application with real data. The autocorrelation at lth lag $(l = 1, \dots, L)$ of a time series $x_{i,t}$ $(t = 1, \dots, T)$ is computed as:

$$\hat{\rho}_{i,l} = \frac{\sum_{t=l+1}^{T}\left(x_{i,t} - \bar{x}_i\right)(x_{i,t-l} - \bar{x}_i)}{\sum_{t=1}^{T}\left(x_{i,t} - \bar{x}_i\right)^2} \tag{18}$$

The ACF-based distance among two time series x_{i,t_i} and x_{j,t_j} of different lengths can be defined as the Euclidean distance among the L estimated auto-correlations. Therefore, the ACF-FMCd method can be written as the solution of the following problem:

$$\begin{cases} \min: \displaystyle\sum_{i=1}^{N}\sum_{c=1}^{C} u_{i,c}^{m} \sum_{l=1}^{L}\left(\hat{\rho}_{i,l} - \hat{\rho}_{c,l}\right)^2 \\ \text{s.t.} \displaystyle\sum_{c=1}^{C} u_{i,c} = 1 \quad \text{and} \quad u_{i,c} \geq 0, \quad \forall i, \forall c \end{cases} \tag{19}$$

This distance can be computed for time series of different lengths as long as they have enough L auto-correlations. As the simulations showed in Díaz and Vilar (2010) suggests, $L = 50$ is a standard choice in this setting.

[2] Data used in our empirical study can be found at http://www.football-data.co.uk.

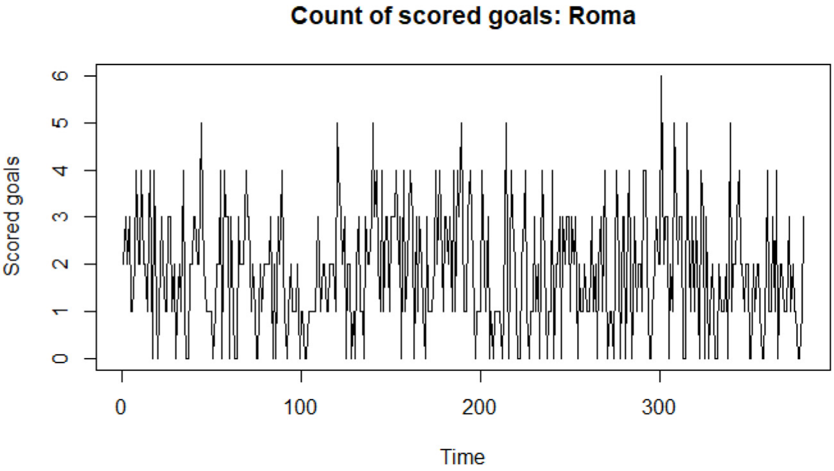

Fig. 1. Roma Calcio: number of scored goals' time series. The representation of the discrete scores through lines leads to an intuitive visualization of the plot.

3.3. Clustering results

For clustering count time series with the proposed INGARCH-based FCMd method, following the GARCH-based approach proposed in Caiado and Crato (2010), we have to estimate first the parameters of the INGARCH$(1,1)$ processes. To further justify our modelling choice, we evaluated for each football team 25 INGARCH(p,q) of different orders in terms of Bayesian Information Criteria (BIC). The model with the lowest BIC provides the best fit. Tables 2 and 3 show that the INGARCH$(1,1)$ model has the best fit for all the football teams since it has the lowest BIC among the alternative orders.

The estimation results are shown in Table 4.

From Table 4 we note that the mean autoregressive often deviates from zero for most teams. In contrast, the component related to the past observed process is tiny for many teams. Moreover, some football teams (e.g. Fiorentina, Juventus, Bologna) are characterized by constant terms larger than one, while only a few (e.g. Inter, Verona, Genoa) show a very small value of the constant term. Then, it is interesting to highlight that some teams have a very persistent mean (e.g. Atalanta, Inter, Verona) while others (e.g. Juventus, Bologna, Torino) show the opposite characteristic. Therefore, exploring similarities by simply looking at the results shown in Table 1 is not easy. For this reason, we employ the proposed INGARCH-FCMd clustering. In defining the dissimilarity (11), it is important the estimation of the matrix $\boldsymbol{\Omega}_{i,j} = \hat{\mathbf{V}}_i + \hat{\mathbf{V}}_j$. We consider the standard errors to include the uncertainty in the parameters estimation step. The standard errors are shown in the parenthesis of Table 4. The models are all overall significant – not all the parameters are statistically equal to zero – but some differences can be highlighted. For example, some football teams with very small estimated parameters δ_1 or γ_1 show no statistically significant values (e.g. Fiorentina, Bologna, Cagliari), but an important difference can still be highlighted in terms of their magnitude and parameter γ_0. With the INGARCH-FCMd method, we explore such differences to identify the clusters' composition.

The INGARCH-FCMd clustering method requires the a-priori selection of the number of clusters C. As explained in Section 2.2, following previous studies, we choose the optimal number of clusters by maximizing the Fuzzy Silhouette (Campello & Hruschka, 2006). The values of the Fuzzy Silhouette for a different number of clusters are shown in Fig. 2.

Accordingly, we choose $C = 4$ groups. The Silhouette value is also entirely satisfactory – suggesting that the partition well explains the differences in the dataset – since with $C = 4$ we obtain a value of $FS_{C=4} = 0.864$. The Silhouette value is satisfactory. Then, we compare the partitions obtained with the benchmark clustering methods previously discussed, i.e. the DTW-FMCd and the ACF-FMCd. The Fuzzy Silhouette associated with the benchmarks are shown in Fig. 3.

In both the cases the Silhouette is maximized with $C = 2$ clusters. However, we should note that the Silhouette values are shallow, meaning that the partitions are of low quality. Indeed, the Fuzzy Silhouette is close to zero for both the alternative clustering methods. In particular, the value obtained with the DTW-FCMd with $C = 2$ clusters is 0.015, while the one obtained with the ACF-FCMd is 0.056. These results confirm that the considered benchmarks are not well suited for clustering count time series as those included in this dataset.

The partition resulting from using the INGARCH-based FCMd with $C = 4$ clusters is shown in Table 5. The clusters' medoids are highlighted in bold font.

First of all, we note that the clusters are quite balanced. The most numerous is cluster 3 (Milan medoid), with six teams, while cluster 1 (Parma medoid) and cluster 2 (Lazio medoid) both include four teams. The cluster with fewer teams is cluster 4 (Udinese medoid) with three teams.

For a deeper understanding of the differences in the clusters, we first analyse the time series of clusters' prototypes (see Figs. 4 and 5).

From Figs. 4 and 5 we observe that the Parma and Udinese, the Cluster 1 and Cluster 4 prototypes, are two teams scoring a lower amount of goals in a match. Indeed, both Parma and Udinese never scored more than five goals, even if Parma scored three or four goals more frequently than Udinese. Conversely, Lazio – i.e. the prototypes of Cluster 2 – scored many times more than five goals, with six matches scoring six goals and one match with seven goals. In the end, Milan – the Cluster 3 prototype – scored only in one match more than 5 goals, but the number of matches with four or five goals is much larger than those of Parma and Udinese. Hence, we argue that Cluster 2 and Cluster 3 include the teams scoring more goals in a single match than the other two clusters.

To further validate these arguments, Table 6 shows the unconditional mean of the processes, calculated according to Eq. (2).

Indeed, from Table 6 we notice that the mean of Cluster 1 (Parma prototype) equals 1.22, the mean of Cluster 2 (Lazio prototype) is 1.81, the mean of Cluster 3 (Milan prototype) is 1.44, and the mean of Cluster 4 (Udinese prototype) is 1.17. More in detail, Cluster 2 includes three teams with very high offencive performance – i.e. scoring a large number of goals per match – such as Lazio (1.75 goals), Napoli (2 goals) and Juventus (1.95 goals). In contrast, cluster 4 includes the Chievo, which is the one with the lowest average number of goals per match (0.92). These arguments allow a clear ranking in terms of offencive performances, which can be outlined as follows: (1) Cluster 2, (2) Cluster 3, (3) Cluster 1, (4) Cluster 4. In other words, Cluster 2 represents the superior set of football teams in terms of offencive performance.

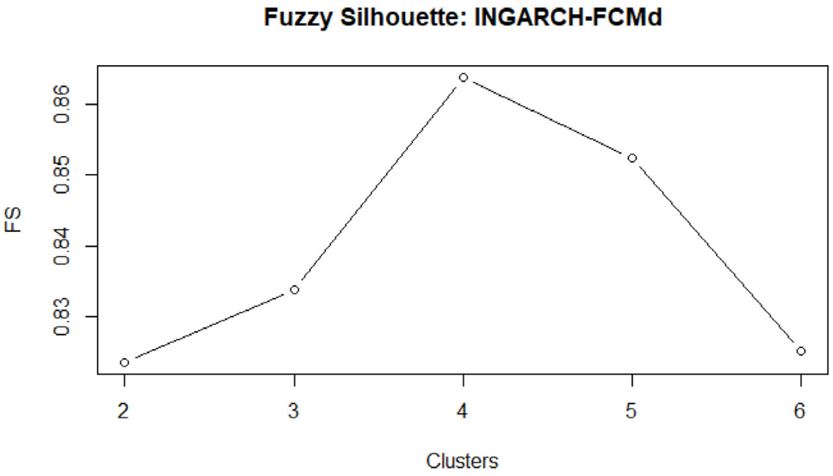

Fig. 2. INGARCH-FCMd: Fuzzy Silhouette.

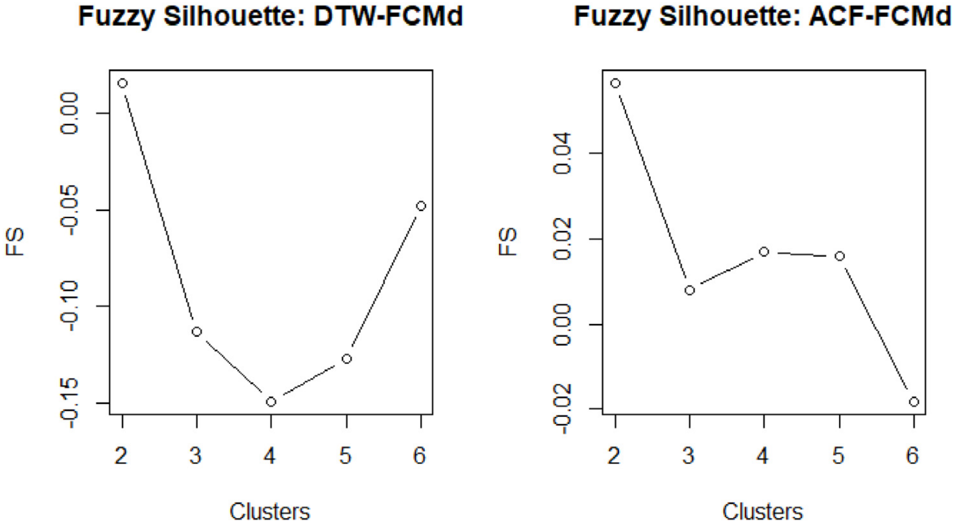

Fig. 3. Benchmark clustering methods: Fuzzy Silhouette.

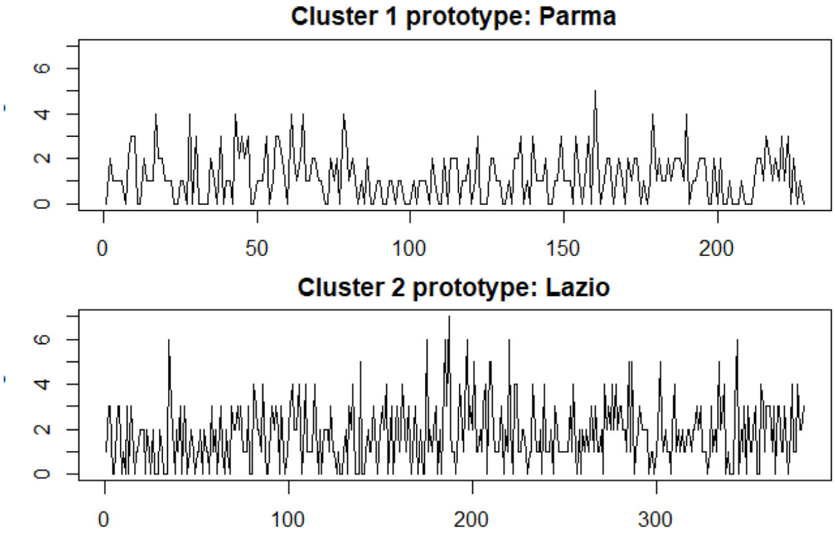

Fig. 4. Number of scored goals' time series: prototypes Cluster 1 and Cluster 2. The representation of the dots of the point process through lines goes in the direction of an easy reading of the plot.

Table 2
BIC for alternative INGARCH(p, q) models – I.

Fiorentina	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1205.12634	1211.066517	1217.006682	1222.946853	1228.887025
$q = 2$	1211.066511	1217.006728	1222.946856	1228.887036	1234.827196
$q = 3$	1217.006683	1222.946855	1228.887025	1234.827196	1240.767367
$q = 4$	1215.519735	1222.318805	1229.399837	1235.323987	1240.818663
$q = 5$	1226.897521	1233.12359	1239.028139	1244.941894	1250.944122
Juventus	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1216.012488	1221.952659	1227.89283	1233.833002	1239.773173
$q = 2$	1221.0382	1227.888597	1231.383248	1237.631807	1245.709124
$q = 3$	1226.60943	1233.131651	1239.479659	1245.482996	1248.831168
$q = 4$	1232.661139	1238.413143	1245.712282	1251.206361	1257.593687
$q = 5$	1236.825902	1243.234527	1246.576684	1255.114805	1261.054945
Atalanta	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1221.253702	1235.756572	1250.473749	1259.090317	1262.473558
$q = 2$	1227.27417	1238.751205	1244.656393	1259.659219	1259.413495
$q = 3$	1234.010356	1247.956953	1254.422938	1261.06524	1265.610191
$q = 4$	1239.562421	1251.835582	1270.424789	1270.065528	1274.398914
$q = 5$	1245.16602	1259.261117	1264.265269	1284.413117	1290.707092
Chievo	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	674.0309837	679.6155871	685.197973	690.7814694	696.3649657
$q = 2$	679.6144768	685.197973	690.7814758	696.3650036	701.9484732
$q = 3$	685.197973	689.6638416	696.3649656	701.9484631	707.1155622
$q = 4$	690.7814693	696.3649673	701.9484631	707.5319648	713.1154547
$q = 5$	695.7252252	701.6694112	706.4353003	712.0152645	718.4199005
Genoa	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1042.670296	1050.821696	1055.461137	1061.920625	1060.834907
$q = 2$	1048.539198	1057.01183	1062.951998	1068.89217	1074.832341
$q = 3$	1057.011832	1062.952011	1068.89217	1074.832341	1080.718779
$q = 4$	1057.081277	1064.612543	1073.536665	1079.45845	1085.417007
$q = 5$	1063.76988	1073.447328	1078.108344	1085.32813	1090.131392
Milan	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1178.778536	1184.421419	1191.786696	1197.885541	1199.815142
$q = 2$	1184.519114	1191.091595	1197.111426	1202.032724	1208.746792
$q = 3$	1191.151615	1196.452868	1203.854386	1207.795496	1217.388172
$q = 4$	1195.932671	1203.323498	1210.511038	1216.594449	1222.277004
$q = 5$	1201.332482	1208.152618	1216.075552	1221.610799	1227.955886
Roma	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1258.304123	1264.970041	1270.265691	1277.168938	1281.258509
$q = 2$	1265.095575	1271.229744	1277.53674	1283.402973	1289.427331
$q = 3$	1271.454966	1277.561311	1283.867768	1289.139963	1295.780317
$q = 4$	1277.67811	1283.608791	1289.101119	1295.532853	1301.473025
$q = 5$	1283.792635	1289.598083	1295.951046	1301.891217	1307.596612
Bologna	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	959.9776279	965.8124387	971.6472505	977.4820601	983.3168709
$q = 2$	962.253917	964.8594815	970.5143862	978.4733953	983.8478885
$q = 3$	960.4638301	972.8429106	980.6411529	983.2673475	991.9936658
$q = 4$	967.6692306	976.5700522	979.3790299	989.1279808	995.3996041
$q = 5$	978.0182296	986.0590947	989.2432864	999.1743291	1004.38162
Torino	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1146.379532	1152.319703	1158.259874	1164.200045	1170.140273
$q = 2$	1152.25065	1157.502203	1164.130987	1168.978334	1176.011321
$q = 3$	1158.259874	1164.200046	1170.140222	1176.080388	1182.020559
$q = 4$	1162.122857	1168.063026	1174.003193	1179.943369	1185.883542
$q = 5$	1170.140217	1176.080388	1182.020559	1187.960731	1193.900902

4. Conclusions

Clustering time series is essential for pattern recognition. Although there have been many contributions on the topic, the problem of clustering count time series is still poorly explored. Count time series arise in many scientific areas, from medicine to epidemiology, engineering, business, and finance.

The main contribution of this paper is the proposal of a new fuzzy clustering method for count time series data. Assuming that the observed time series processes are conditionally Poisson distributed, we developed a model-based fuzzy clustering method based on the parameters estimated by the INGARCH processes. A Mahalanobis distance is proposed to account for uncertainty in the estimation process, such that the parameters with high variability are weighted less than those with low variability. To the best of our knowledge, the so-called INGARCH-based fuzzy C-medoids (INGARCH-FCMd) method is the first attempt to develop partitional clustering methods for count time series data.

An essential property of the proposed clustering method is that, as most model-based clustering methods (e.g. D'Urso et al., 2016; Otranto, 2008; Piccolo, 1990), it does not require the time series to have the same lengths. However, we should note that the time series length affects the quality of the parameter estimates. In other words, if the length is too short, the estimated parameters would be inflated, and the resulting clustering would be wrong or less accurate. Although the relevance of this problem, in the proposed application to football data, the time series lengths are enough to get reasonably good estimates. Such a claim is confirmed by the obtained clustering results, which are easily explainable.

Table 3
BIC for alternative INGARCH(p, q) models — II.

Cagliari	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	948.5274456	954.3622555	960.1970662	966.031877	971.8666877
$q = 2$	948.0152007	953.1419301	959.6848625	965.5196483	971.3544346
$q = 3$	960.1970663	966.031877	971.8666877	977.7015037	983.5363094
$q = 4$	964.7036387	970.5322353	976.3732023	982.2080614	988.042824
$q = 5$	971.8666877	977.7014991	983.5363092	989.3711199	995.2059346
Inter	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1255.704549	1261.283298	1268.218688	1274.885665	1285.229461
$q = 2$	1262.079181	1273.726364	1277.75341	1285.838192	1291.391725
$q = 3$	1268.341881	1273.985701	1282.491349	1289.457557	1298.702953
$q = 4$	1273.927957	1286.200848	1286.90349	1298.081014	1297.514239
$q = 5$	1285.105285	1284.708401	1296.985626	1299.566353	1308.865968
Lazio	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1277.733	1283.67317	1289.613342	1295.553513	1301.493684
$q = 2$	1283.673171	1289.613342	1295.553513	1301.493684	1307.433855
$q = 3$	1285.564813	1295.515173	1301.455247	1307.289682	1312.040099
$q = 4$	1293.225135	1299.165306	1298.001482	1308.337403	1313.001856
$q = 5$	1301.493684	1307.433855	1313.374028	1319.314198	1325.254369
Napoli	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1314.709295	1320.664937	1324.465048	1332.545277	1338.316678
$q = 2$	1322.032183	1326.782	1333.912526	1339.490499	1345.74476
$q = 3$	1328.065712	1334.005883	1339.946062	1345.886225	1351.826772
$q = 4$	1331.995495	1337.481807	1343.653016	1349.79681	1355.75618
$q = 5$	1339.946054	1345.886225	1351.826396	1357.766568	1363.706739
Parma	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	656.6918123	662.1211587	667.5505033	672.9801667	678.4091945
$q = 2$	660.0743654	665.1321421	670.7273248	677.2843472	682.7056238
$q = 3$	666.6270086	672.6187616	678.3955424	683.7164785	686.8908731
$q = 4$	671.7193097	676.8282829	682.5659827	687.3377689	693.4304615
$q = 5$	678.4091946	683.4466038	688.1752756	694.0292924	700.1265771
Sampdoria	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1124.51423	1130.454401	1136.394572	1142.334745	1148.274914
$q = 2$	1130.4544	1136.394572	1142.334743	1148.274915	1154.215085
$q = 3$	1136.392379	1142.334743	1148.274914	1153.59899	1158.079296
$q = 4$	1142.21947	1148.159642	1153.539125	1159.994163	1165.21376
$q = 5$	1148.274914	1154.215104	1160.155257	1166.095428	1172.035601
Udinese	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1082.092294	1087.340427	1094.304004	1099.274857	1106.184342
$q = 2$	1088.726001	1094.69178	1100.080052	1106.572122	1110.505032
$q = 3$	1095.192307	1100.124012	1106.979875	1112.785653	1118.860225
$q = 4$	1101.132488	1107.072649	1113.012821	1118.952993	1124.893164
$q = 5$	1105.483619	1111.270386	1117.363987	1123.304133	1129.244306
Verona	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	760.6692811	767.7627475	776.5753058	779.9606719	786.9314912
$q = 2$	766.0128276	773.4627754	776.7878021	783.709982	791.784007
$q = 3$	770.408404	775.9768816	781.6217238	787.204346	790.3805757
$q = 4$	780.1695401	785.5885431	793.6503286	798.3617633	802.8750701
$q = 5$	787.872534	792.9135303	794.9359317	802.5115581	809.9182815
Sassuolo	p=1	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$q = 1$	1029.76582	1036.767953	1042.655495	1048.490307	1054.325133
$q = 2$	1036.211227	1042.147149	1047.981935	1053.816751	1059.651554
$q = 3$	1042.644114	1050.226612	1056.061426	1061.896246	1067.731043
$q = 4$	1050.454091	1056.288901	1062.123712	1067.958523	1073.793334
$q = 5$	1056.288901	1062.123712	1067.958523	1070.22793	1079.628144

In particular, for the experiment with real data we compared the proposed INGARCH-based clustering algorithm (INGARCH-FCMd) with two observation- and feature-based benchmarks, namely the DTW-FCMd (Izakian et al., 2015) and the ACF-FCMd (D'Urso & Maharaj, 2009). Both benchmark approaches can deal with time series of different lengths. The results show that the proposed INGARCH-FCMd method provides a much better clustering quality – measured in terms of Fuzzy Silhouette (Campello & Hruschka, 2006) – with respect the benchmarks.

Because of the novelty of the problem, there are many future possible research directions. First of all, we should note that distance (11) is based on INGARCH models of the same order. Although this choice can be reasonably justified – parsimony, fast estimation, well-understood statistical properties, substantial empirical evidence – it is interesting extending the proposed clustering framework to account for INGARCH models of different orders. It is known that the INGARCH process, as happens for the standard GARCH, is also an ARMA process. Therefore, it admits an $AR(\infty)$ representation. A Euclidean distance based on the $AR(\infty)$ representation can be considered for clustering count time series with different INGARCH orders. In other words, this alternative method should mimic the ARMA-based one proposed by Piccolo (1990).

The second aspect of improvement is that the proposed clustering method, as the classical model-based clustering methods such as ARIMA-based or GARCH-based, assumes that the football teams' time series are independent of each other. This assumption does not hold in all real-world examples. A way of alleviating the independence problem, especially in the contexts of sports data applications as the one presented in this paper, can be the inclusion of covariates in the statistical model. With this respect, we have to note that the INGARCH-FCMd method can be extended to include additional variables because the INGARCH model allows for exogenous covariates (INGARCH-X or

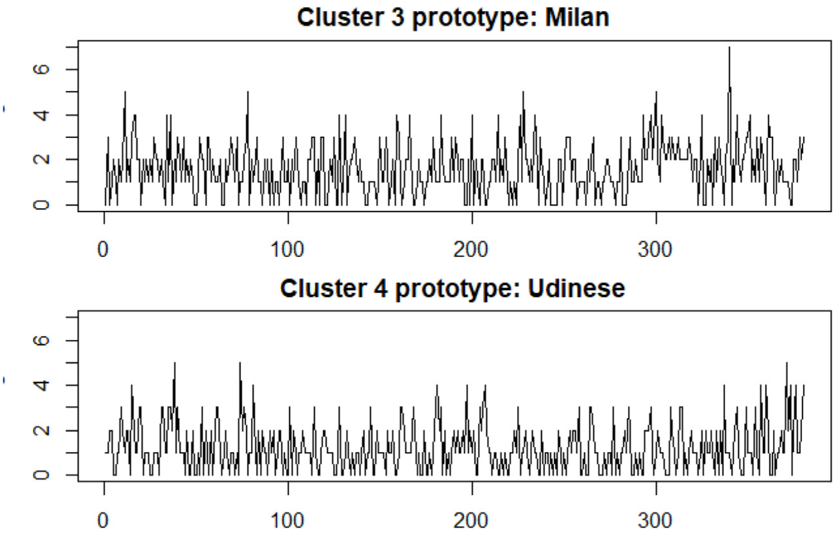

Fig. 5. Number of scored goals' time series: prototypes Cluster 3 and Cluster 4. The representation of the dots of the point process through lines allows an easy reading of the plot.

Table 4
INGARCH(1, 1) estimates for Italian football teams.

Team (j)	$\hat{\gamma}_{0,j}$	$\hat{\gamma}_{1,j}$	$\hat{\delta}_{1,j}$
Fiorentina	1.5236^{***}	0.0000	0.0001
	(0.0633)	(0.0319)	(0.0416)
Juventus	1.8752^{***}	0.0000	0.0384
	(0.0689)	(0.0305)	(0.0353)
Atalanta	0.0000^{**}	0.0356^{***}	0.9643^{***}
	(0.0000)	(0.0000)	(0.0000)
Chievo	0.6896^{***}	0.0000	0.2543^{***}
	(0.0440)	(0.0377)	(0.0475)
Genoa	0.0643^{***}	0.0308^{***}	0.9102^{***}
	(0.0047)	(0.0042)	(0.0043)
Milan	0.1386^{***}	0.0366^{***}	0.8761^{***}
	(0.0079)	(0.0050)	(0.0050)
Roma	0.3483^{***}	0.0453^{***}	0.7659^{***}
	(0.0163)	(0.0086)	(0.0088)
Bologna	1.1138^{***}	0.0000	0.0029
	(0.0570)	(0.0368)	(0.0510)
Torino	1.3692^{***}	0.0000	0.0051
	(0.0599)	(0.0331)	(0.0435)
Cagliari	1.1050^{***}	0.0000	0.0002
	(0.0568)	(0.0370)	(0.0514)
Inter	0.0161^{***}	0.0202^{***}	0.9707^{***}
	(0.0016)	(0.0009)	(0.0009)
Lazio	1.7455^{***}	0.0000	0.0011
	(0.0677)	(0.0304)	(0.0388)
Napoli	1.7420^{***}	0.0616	0.0743
	(0.0674)	(0.0286)	(0.0334)
Parma	1.1900^{***}	0.0000	0.0025
	(0.0722)	(0.0444)	(0.0605)
Sampdoria	0.9981^{***}	0.0000	0.2383^{***}
	(0.0447)	(0.0282)	(0.0341)
Udinese	0.6092^{***}	0.0474	0.4464^{***}
	(0.0311)	(0.0231)	(0.0259)
Verona	0.0345^{***}	0.0392^{***}	0.9344^{***}
	(0.0041)	(0.0033)	(0.0033)
Sassuolo	0.6452^{***}	0.0800^{**}	0.4581^{***}
	(0.0345)	(0.0225)	(0.0247)

Note: Parameters are estimated with MLE and the associated standard errors are reported in parenthesis under the estimates. ***, ** and * indicate significance at 1%, 5% and 10% confidence levels, respectively.

Table 5
INGARCH-FCMd clustering: membership degrees.

Team	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Fiorentina	0.1540	0.8408	0.0012	0.0041
Juventus	0.0013	0.9984	0.0001	0.0002
Atalanta	0.0000	0.0000	0.9999	0.0001
Chievo	0.0084	0.0006	0.0084	0.9826
Genoa	0.0000	0.0000	1.0000	0.0000
Milan	0.0000	0.0000	**1.0000**	0.0000
Roma	0.0001	0.0000	0.9678	0.0320
Bologna	0.9993	0.0002	0.0001	0.0004
Torino	0.9433	0.0520	0.0009	0.0039
Cagliari	0.9990	0.0003	0.0001	0.0006
Inter	0.0000	0.0000	0.9999	0.0001
Lazio	0.0000	**1.0000**	0.0000	0.0000
Napoli	0.0002	0.9998	0.0000	0.0000
Parma	**1.0000**	0.0000	0.0000	0.0000
Sampdoria	0.6640	0.0136	0.0202	0.3022
Udinese	0.0000	0.0000	0.0000	**1.0000**
Verona	0.0000	0.0000	0.9999	0.0001
Sassuolo	0.0000	0.0000	0.0001	0.9999

Table 6
Unconditional mean of the process.

Team	Unconditional mean	Crisp assignment
Fiorentina	1.5237	2
Juventus	1.9500	2
Atalanta	1.0360	3
Chievo	0.9247	4
Genoa	1.0889	3
Milan	1.5870	3
Roma	1.8445	3
Bologna	1.1170	1
Torino	1.3763	1
Cagliari	1.1053	1
Inter	1.7771	3
Lazio	1.7474	2
Napoli	2.0160	2
Parma	1.1929	1
Sampdoria	1.3105	1
Udinese	1.2034	4
Verona	1.3083	3
Sassuolo	1.3966	4

PARX model, e.g. see Agosto et al., 2016; Angelini & De Angelis, 2017; Lee & Lee, 2019).

A third interesting future direction is the development of a clustering method which handles count series characterized by overdispersion, i.e. the variance of the process is greater than the mean. A simple way to deal with overdispersion is to consider conditionally Negative Binomial distributed time series.

Ultimately, it is interesting to highlight that the INGARCH model belongs to the more comprehensive Generalized Linear Model (GLM) class for time series data. GLM have great potential in developing count time series clustering methods since GLM represents a general way for introducing alternative conditional distributions, nonlinear models and exogenous covariates. With this respect, we are currently working on developing a clustering method based on nonlinear time series models, which possibly include the information of some exogenous covariates, and that can generalize the INGARCH-based clustering approach presented in this paper in the directions suggested above.

CRediT authorship contribution statement

Roy Cerqueti: Conceptualization, Methodology, Supervision, Formal analysis, Visualization. **Pierpaolo D'Urso:** Conceptualization, Methodology, Supervision, Visualization. **Livia De Giovanni:** Conceptualization, Methodology, Software, Supervision, Visualization. **Raffaele Mattera:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Software, Validation, Formal analysis, Data Curation. **Vincenzina Vitale:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Software, Validation, Formal analysis, Data Curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agosto, A., Cavaliere, G., Kristensen, D., & Rahbek, A. (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *Journal of Empirical Finance*, *38*, 640–663.

Agosto, A., & Giudici, P. (2020). A Poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks*, *8*(3), 77.

Aknouche, A., Almohaimeed, B. S., & Dimitrakopoulos, S. (2021). Forecasting transaction counts with integer-valued GARCH models. *Studies in Nonlinear Dynamics & Econometrics*, http://dx.doi.org/10.1515/snde-2020-0095.

Angelini, G., & De Angelis, L. (2017). PARX model for football match predictions. *Journal of Forecasting*, *36*(7), 795–807.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256.

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, *25*(3), 2499–2511.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD Workshop, vol. 10, no. 16* (pp. 359–370). Seattle, WA, USA:.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Caiado, J., & Crato, N. (2010). Identifying common dynamic features in stock returns. *Quantitative Finance*, *10*(7), 797–807.

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, *50*(10), 2668–2684.

Caiado, J., Crato, N., & Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification*, *14*(1), 117–146.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, *157*(21), 2858–2875.

Cerqueti, R., D'Urso, P., De Giovanni, L., Giacalone, M., & Mattera, R. (2022). Weighted score-driven fuzzy clustering of time series with a financial application. *Expert Systems with Applications*, *198*, Article 116752.

Cerqueti, R., Giacalone, M., & Mattera, R. (2021). Model-based fuzzy time series clustering of conditional higher moments. *International Journal of Approximate Reasoning*, *134*, 34–52.

Chen, C. W., & Lee, S. (2017). Bayesian causality test for integer-valued time series models with applications to climate and crime data. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *66*(4), 797–814.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., & Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

Díaz, S. P., & Vilar, J. A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: A simulation study. *Journal of Classification*, *27*(3), 333–362.

D'Urso, P., De Giovanni, L., & Massari, R. (2016). GARCH-based robust clustering of time series. *Fuzzy Sets and Systems*, *305*, 1–28.

D'Urso, P., De Giovanni, L., & Massari, R. (2018). Robust fuzzy clustering of multivariate time trajectories. *International Journal of Approximate Reasoning*, *99*, 12–38.

D'Urso, P., De Giovanni, L., Massari, R., D'Ecclesia, R. L., & Maharaj, E. A. (2020). Cepstral-based clustering of financial time series. *Expert Systems with Applications*, *161*, Article 113705.

D'Urso, P., De Giovanni, L., & Vitale, V. (2022). A robust method for clustering football players with mixed attributes. *Annals of Operations Research*, 1–28.

D'Urso, P., García-Escudero, L. A., De Giovanni, L., Vitale, V., & Mayo-Iscar, A. (2021). Robust fuzzy clustering of time series based on B-splines. *International Journal of Approximate Reasoning*.

D'Urso, P., & Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, *160*(24), 3565–3589.

Ferland, R., Latour, A., & Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis*, *27*(6), 923–942.

Fokianos, K., Rahbek, A., & Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, *104*(488), 1430–1439.

Garcia-Escudero, L. A., & Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, *22*(2), 185–201.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, *12*(2), 434–449.

Greenhough, J., Birch, P., Chapman, S., & Rowlands, G. (2002). Football goal distributions and extremal statistics. *Physica A: Statistical Mechanics and its Applications*, *316*(1–4), 615–624.

Groll, A., Kneib, T., Mayr, A., & Schauberger, G. (2018). On the dependency of soccer scores–a sparse bivariate Poisson model for the UEFA European Football Championship 2016. *Journal of Quantitative Analysis in Sports*, *14*(2), 65–79.

Izakian, H., Pedrycz, W., & Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, *39*, 235–244.

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data. An introduction to cluster analysis. In *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167–186.

Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Transactions on Fuzzy Systems*, *9*(4), 595–607.

Krishnapuram, R., Joshi, A., & Yi, L. (1999). A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *FUZZ-IEEE'99. 1999 IEEE international fuzzy systems. conference proceedings (Cat. No. 99CH36315), vol. 3* (pp. 1281–1286). IEEE.

Lafuente-Rego, B., D'Urso, P., & Vilar, J. A. (2020). Robust fuzzy clustering based on quantile autocovariances. *Statistical Papers*, *61*(6), 2393–2448.

Lee, Y., & Lee, S. (2019). On causality test for time series of counts based on Poisson INGARCH models with application to crime and temperature data. *Communications in Statistics. Simulation and Computation*, *48*(6), 1901–1911.

Liao, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognition*, *38*(11), 1857–1874.

López-Oriona, Á., & Vilar, J. A. (2021). Quantile cross-spectral density: A novel and effective tool for clustering multivariate time series. *Expert Systems with Applications*, *185*, Article 115677.

López-Oriona, Á., Vilar, J. A., & D'Urso, P. (2022). Quantile-based fuzzy clustering of multivariate time series in the frequency domain. *Fuzzy Sets and Systems*.

Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). *Time series clustering and classification*. CRC Press.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B*, *11*(1), 193–197.

Mattera, R. (2021). Forecasting binary outcomes in soccer. *Annals of Operations Research*, 1–20.

Matteson, D. S., McLean, M. W., Woodard, D. B., & Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, *5*(2B), 1379–1406.

Mourao, P. R. (2016). Soccer transfers, team efficiency and the sports cycle in the most valued European soccer leagues–have European soccer teams been efficient in trading players? *Applied Economics*, *48*(56), 5513–5524.

Narizuka, T., & Yamazaki, Y. (2019). Clustering algorithm for formations in football games. *Scientific Reports*, *9*(1), 1–8.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, *52*(10), 4685–4698.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, *44*(3), 678–693.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, *11*(2), 153–164.

Rydberg, T. H., & Shephard, N. (2000). A modelling framework for the prices and times of trades made on the New York stock exchange. In *Nonlinear and Nonstationary Signal Processing* (pp. 217–246). Cambridge University Press Cambridge.

Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, *93*, Article 101562.

Savvides, A., Promponas, V. J., & Fokianos, K. (2008). Clustering of biological time series by cepstral coefficients based distances. *Pattern Recognition*, *41*(7), 2398–2412.

Ulas, E. (2021). Examination of National Basketball Association (NBA) team values based on dynamic linear mixed models. *PLoS One*, *16*(6), Article e0253179.

Vilar, J. A., Lafuente-Rego, B., & D'Urso, P. (2018). Quantile autocovariances: A powerful tool for hard and soft partitional clustering of time series. *Fuzzy Sets and Systems*, *340*, 38–72.

Xiong, L., & Zhu, F. (2019). Robust quasi-likelihood estimation for the negative binomial integer-valued GARCH (1, 1) model with an application to transaction counts. *Journal of Statistical Planning and Inference*, *203*, 178–198.