Quality and Reliability Engineering International manuscript No.
(will be inserted by the editor)

Clustering Functional Data Streams: Unsupervised Classification of Soccer Top Players based on Google Trends

Francesca Fortuna · Fabrizio Maturo · Tonio Di Battista

Received: date / Accepted: date

Abstract In the last decades, the analysis of web data has become crucial for a large number of companies. For many enterprises, the analysis of customer behaviour in surfing the world wide web is crucial for making or changing corporate strategies. Information about search engine queries, number of visits, bounce rate, time on site, traffic sources, and success of an online advertising are essential for understanding if and where firms are losing customers, forecasting future sales, understanding past performance, analyzing profile customer behaviour, monitoring buying patterns, measuring the impact of site changes, and removing barriers to sale. Therefore, it is widely recognized that discovering advanced methodologies for inspecting this kind of data and finding marketing strategies is a very relevant topic both in business administration and statistics. The study of this type of data often means to deal with big data and data streaming. This study focuses on a specific type of web data, i.e. trends of search engine queries. Specifically, focusing on daily google queries regarding 24 top football players, we propose a functional k-means approach for identifying specific patterns of their "google trends". Nowadays, many football teams have become large companies that produce very high incomes, and the

Francesca Fortuna
"G. d' Annunzio" University of Chieti-Pescara
DISFPEQ, Pescara, Italy
Tel.: +390854537551
E-mail: francesca.fortuna@unich.it

Fabrizio Maturo
"G. d' Annunzio" University of Chieti-Pescara
Department of Management and Business Administration, Pescara, Italy
Tel.: +390854537431
E-mail: f.maturo@unich.it

Tonio Di Battista
"G. d' Annunzio" University of Chieti-Pescara
DISFPEQ, Pescara, Italy
Tel.: +390854537551
E-mail: dibattis@unich.it

induced gain and interest that may arise from purchasing notorious players is become a crucial aspect for their commercial strategies. Thus, we believe that our approach could provide interesting insights for practitioners, scholars, and insiders.

Keywords clustering football players · data streaming · google query · google trends · FPCA

1 Introduction

During the last decades, sports analytics have been attracting the attention of worldwide practitioners and scholars due to economic and non-economic interests. Particularly, soccer has become tremendously popular and widely televised, and nowadays it generates huge amounts of revenue and inspires countless hours of debate and conversation among fans [1]. Soccer is discussed for fun or for making money in betting markets, and many football teams have become large companies that produce very high incomes (also independently of the team's performance).

Purchasing football players has become, unlike a few decades ago, a commercial event. Buying notorious players is not just a matter of football convenience, but it has become a fact of marketing and advertising. The acquisition of certain players may increase the interest around a team, number of subscribed fans and sold tickets, television rights, incomes from merchandising (football shirts, signed objects, gadgets, etc.), and the number of companies which may be ready to finance teams (e.g. the induced gain and interest that may arise from purchasing players such as Neymar, Ronaldo, or Messi is simply unquantifiable). Sometimes, the purchase of particular players may even attract the interest of followers and investors from foreign countries (e.g. purchase operations of players such as Keisuke Honda have brought thousands of Japanese fans to follow AC Milan). This kind of operations are able to move huge capitals from one country to another. Hence, due to the increasing commercialization of soccer development and its competitiveness, the demand for efficient use of resources within a football club is becoming more and more relevant [2] [3]. For these reasons, particularly in recent years, an increasing number of statistical models has been proposed to pursue game and market values predictions [4] [5]. Papers on this topic appear regularly in many leading journals, and most scholars agree that in view of its social, cultural and economic importance, professional sport is a legitimate area of interest for both theoretical and empirical researchers [5].

The enormous attention paid to football in today's society, where technology is king, is automatically reflected in internet traffic. Hence, football clubs (and professionals in general) need advanced statistical tools which are able to describe and anticipate the preferences and interests of the worldwide public. Thus, they need instruments for inspecting web data, and particularly, data streaming. In truth, the analysis of web data is crucial for every type of companies which needs to analyze consumer behaviour in surfing their web-sites.

Many information, such as number of visits, bounce rates, time on site, traffic sources, and success of an online advertising are essential for understanding if and where firms are losing customers, forecasting future sales, understanding past performances, analysing customer behaviour, and monitoring buying patterns. Analysing this type of data often means to deal with big data and data streaming. A data stream is an unbounded, ordered sequence of objects that can be read only once or a small number of times [6]. The main characteristics of data streaming are that data continuously flow, and their size is extremely large and potentially infinite. Examples of data stream include web searches, web traffic, phone conversations, ATM transactions, among the others.

In the context of soccer, football teams are certainly interested in understanding which players may bring high incomes and many followers due to their notoriety and potential appeal. Thus, these companies should ask a simple question: what are the most famous players, those who can bring fans and earnings to the company? In today's society, one possible answer is surely the following: the players who have the greatest potential to create gains in the short term (and not only) are surely those most sought by worldwide internet users. An excellent response for football clubs, is therefore contained in researches carried out by people in search engines. However, this kind of data evolves continuously over time. The simple total number of visits is not, in itself, a useful indicator of the fame and "charm" of a player. Indeed, research of players' names, in the short term, could be influenced by extreme and random events, such as an accident, a resounding individual error, an expulsion and so on. On the contrary, the analysis of the trend over time of player searches can hide interesting information for insiders.

A common way for analyzing this kind of data involves clustering techniques [7] [8] [9]. The clustering problem is that, for a given set of data points, we wish to partition them into one or more groups of similar objects. The similarity of an object with another one is typically obtained through the use of some distance measure or introducing an objective function. However, it is a challenging problem for the data stream domain due to the curse of dimensionality, which lead to difficulties in detecting clusters in the data set, clusters robustness over time, the large volume of data which makes most traditional algorithms inefficient, and computational time-consuming. A data stream should be viewed as an infinite process consisting of data which continuously evolves over time, and thus, also the underlying clusters may considerably change depending on the moment they are computed. For these reasons, a data stream clustering algorithm should provide the flexibility to compute clusters over user-defined time periods in an interactive fashion. In recent years, several clustering algorithms have been developed for data streams[8] [10] [9]. Although such methods address the scalability problems, they do not consider the following issues: the quality of clusters is poor when data evolves a lot over time, a data stream clustering algorithm should explore the data at different portions of the stream, and the algorithmic efficiency in terms of the number of computational resources. Several algorithms for clustering data streams are based on k-means [11], which has been proposed in several vari-

ants (e.g., see [12] [6]). Despite the successful application of these algorithms to many real-world problems, they have some major limitations. Indeed, the sampling units are observed in a finite set of time points that may be irregularly spaced and different for the same individuals, the number of clusters must be defined a priori by the user, the choice of the distance metric, and algorithm convergence due to possible local minimum and high dimensionality. To solve these issues, the functional data analysis (FDA) approach seems to be appropriate. The latter provides several advantages because it grants the evaluation of the curves behavior throughout the whole reference domain. For example, it may be helpful to improve the graphical interpretation of functions with additional functional tools, such as integrals, derivatives, and so on [13] [14] [15]. In addition, functional versions of classical statistical methods can be extended to different contexts with various applications [16, e.g.], [17], [18], [19], [20], [21], [22], [23]; particularly, we can exploit FDA also in datastreams analysis keeping crucial information. Data flowing from the server of a web site, such as the number of queries on google or clicks on hypertests, and the number of likes on different contents of a web page, can be viewed as a single continuous stream of data in the time domain. Thus, each single stream of observations can be seen as a function rather than vector.

The remainder of the paper is organized as follows. Sect. 2 presents the materials and methods. In particular, it illustrates the functional principal components approach and related k-means clustering algorithm for functional data. Sect. 3 shows the main results obtained by applying the proposed approach to a real data set concerning the daily google queries on 24 top football players. Finally, Sect. 4 presents the discussion and conclusions of this study.

2 Materials and Methods

High dimensional data monitoring has recently attracted increasing attention among researchers as well as practitioners. However, existing process monitoring methods fail to fully utilize the information of high dimensional data streams due to their complex characteristics including the large dimensionality, spatio-temporal correlation structure, and non-stationarity [24]. A possible solution to these issues may be provided by the FDA approach. Indeed, each stream can be seen as a function in a continuous domain, such as time. The central idea behind FDA is that each function is assumed to be a sample from a smooth function of time. However, despite the continuous nature of functional data, in real applications, curves are observed at a finite set of sampling points. One usual solution to reconstruct the functional form of the data, is to assume that sample paths belong to a finite-dimension space spanned by a basis [17]. In this context, a strem such as the temporal trend of a generic web data, $X(t)$, could be represented by a basis expansion as follows:

$$X_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t) \quad i = 1, ..., n \tag{1}$$

where $X_i(t)$ is the reconstructed function for the i-th unit; $\phi_k(t)$ are linearly independent and known basis functions; and a_{ik} are coefficients that link each basis function together in the representation of $X_i(t)$. Various basis systems can be adopted, depending on the characteristics of the curves [25] [26] [27]. Usual basis systems are B-spline functions for non periodic data; Fourier functions for periodic data; piecewise constant functions for counting processes; and wavelets bases for curves with strong local behavior. The basis coefficients can be obtained from the discrete observations either by interpolation, when data are observed without error, and least square approximation in the other case. In this work, we consider least squares approximation with B-splines basis for the functional representation of web-data.
In this context, one can be interested in optimal representation of curves into a function space of reduced (finite) dimension [28] for explaining the main features of the data. To address this issue, functional principal component analysis (FPCA) can be adopted. The latter allows us to explain the dependence structure of a functional data set in terms of a reduced set of uncorrelated variables [29] [30] [16]. In particular, let us assume that the observed curves are centered so that the sample mean is equal to zero. Then, for each unit $(i = 1, .., n)$, the j-th principal component score is given by:

$$\xi_{ij} = \int_T x_i(t) f_j(t)\, dt \tag{2}$$

where the weight functions or loadings $f_j(t)$ are the solutions of the eigenequation $C(f_j) = \int c(t,s) f_j(s)\, ds = \lambda_j f_j(t)$, where $c(t,s)$ is the sample covariance function and $\lambda_j = Var[\xi_j]$ [31] [32]. Then, the sample curves admit the following principal component decomposition:

$$x_i(t) = \sum_{j=1}^{p} \xi_{ij} f_j(t) \quad i01, ..., n \tag{3}$$

By truncating this representation in terms of the first q principal components $(q << p)$, we can obtain an approximation of the sample curves whose explained variance is given by $\sum_{j=1}^{q} \lambda_j$.
A common way for analyzing this kind of data involves clustering techniques. In this paper, clustering of functions is carried out in combination with dimension reduction in order to remove the effect of irrelevant functional information and detect the underlying cluster structure based on relevant (low-dimensional) functional information only [10]. In particular, we refer to FPCA for computing proximities between functional data as semi-metric distances based on the approach of [16]. Specifically, we focus on the semimetric PCA that computes distance between curves based on the FPCA method, as follows:

$$d_2^{(q)}\left(X_i, X_i'\right)_K \approx \left[\sum_{j=1}^{q} \left(\xi_{ij} - \xi_{i'j}\right)^2 \left\| f_j^{(q)} \right\|\right]^{1/2} \quad i \neq i' \tag{4}$$

where q denotes the reduced dimensional space at q components [16] [30].
The basic idea of this unsupervised clustering approach is to find a partition for which the variability within clusters is minimized. The most used algorithm, in this context, is the k-means. Starting from n functional observations, this method aims to group units into $G \leq n$ sets $S = S_1, S_2, ..., S_G$ so as to minimize the within-cluster sum of squares. The first step of this iterative procedure consists in fixing G initial centroids in S, $\{c_1^{(0)}(t), ..., c_G^{(0)}(t)\}$. Then, each function is assigned to the cluster whose centroid, at the previous iteration $(m-1)$ is the nearest according to the chosen distance:

$$C_g^{(m)} = \underset{g=1,2,..,G}{\arg\min} \left[\sum_{j=1}^{q} \sum_{i=1}^{n} \|\xi_{ij} - \psi_g^{m-1}(t)\|^2 \right] \quad m = 0, 1, ..., M \tag{5}$$

where $C_g^{(m)}$ is the m-th cluster assignment of the i-th function, $i = 1, 2, ..., n$. Once all the functions have been assigned to a cluster, the cluster means are updated as:

$$\psi_g^m(t) = \sum_{\xi_{ij} \in C_g} \frac{\xi_{ij}}{n_g} \tag{6}$$

where n_g is the number of functions in the g-th cluster, C_g. This procedure continues until no function changes cluster or a maximum number of iteration should be selected in advance. This process is completed with a strategy for dimensionality reduction consisting of a new projection onto a finite-dimensional Euclidean space which makes the distance between functions coincide with the Euclidean distance of the projected data [19].

3 Application

Following, we present an application in the field of web-traffic data analysis; particularly, we focus on the number of daily google queries of a sample of 24 football top players (https://trends.google.com/trends/). The latter have been selected according to the following criteria: market value higher than 50 million euros, valued on December 19, 2017 (https://www.transfermarkt.it/); and only players who participate in 2017-18 UEFA Champions League are considered. Table 1 describes some details of our statistical units; we can observe that 24 football players belong to the major European football clubs. Their marked values range between 60 millions and 150 millions of euros (Neymar).
Figure 1 shows the trends of the google queries containing the names of the 24 players during one day. Naturally, the graph is very confused because the time domain (day) is composed of 180 instant observations, starting at 13:24 on the day 2017.12.18 until 13.08 on the day 2017.12.19. The intervals among observations is quite 8 minutes, and thus we can easily observe strong variability both within and between players along the day. To overcome the illegibility of the chart, we refer to the FDA approach for creating functions as linear combinations of B-splines for each player (Eq. 1). Figure 2 shows

Player name	Nationality	Age	Footbal club	Market value (mln euros)
Neymar	Brasil	25	FC Paris Saint-Germain	150,00
Lionel Messi	Argentine/ Spain	30	FC Barcellona	120,00
Cristiano Ronaldo	Portugal	30	Real Madrid CF	100,00
Kylian Mbappé	France/Cameroon	18	FC Paris Saint-Germain	90,00
Luis Suárez	Uruguay	30	FC Barcellona	90,00
Robert Lewandowski	Poland	29	FC Bayern Monaco	80,00
Antoine Griezmann	France/Cameroon	26	Atlético de Madrid	80,00
Harry Kane	England	24	Tottenham Hotspur	80,00
Gareth Bale	Wales	28	Real Madrid CF	80,00
Eden Hazard	Belgium	26	Chelsea FC	75,00
Paul Pogba	France/Guinea	24	Manchester United	75,00
Kevin De Bruyne	Belgium	26	Manchester City	75,00
Romelu Lukaku	Belgium/RD of Congo	24	Manchester United	70,00
Toni Kroos	Germany	27	Real Madrid CF	70,00
Paulo Dybala	Argentine/Italy	24	Juventus FC	70,00
Gonzalo Higuaín	Argentine/France	30	Juventus FC	70,00
Sergio Agüero	Argentine/Spain	29	Manchester City	65,00
Pierre-Emerick Aubameyang	Gabon/France	28	Borussia Dortmund	65,00
Philippe Coutinho	Brasil	25	FC Liverpool	65,00
Dele Alli	England/Nigeria	21	Tottenham Hotspur	60,00
Karim Benzema	France/Algeria	30	Real Madrid CF	60,00
Marco Verratti	Italy	25	FC Paris Saint-Germain	60,00
Koke	Spain	25	Atlético de Madrid	60,00
Sergio Busquets	Spain	29	FC Barcellona	60,00

Table 1: Description of the 24 football players under analysis

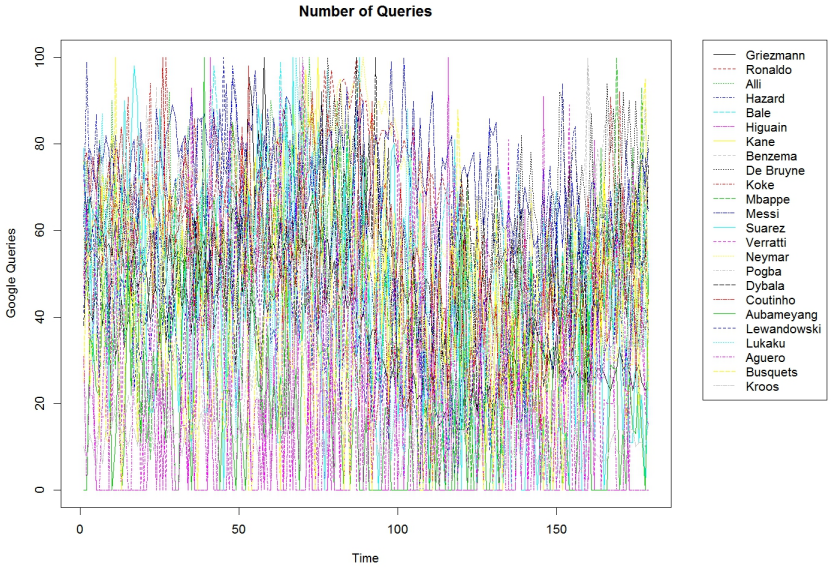

Fig. 1: Number of Google Queries for each Football Top Player (the domain is divided into 180 points starting at 13:14 on the day 2017.12.18 until 13.08 of the day 2017.12.19)

the 24 reconstructed functional data (one for each player). In the first part of the domain, we observe that Gonzalo Higuain, Sergio Agüero, and Antoine Griezmann are the less researched names on google. On the contrary, Koke, Neymar, Paulo Dybala, and Eden Hazard are the most searched in the same part of the domain (13:24- 14:20 on 2017.12.18). In the middle part of the domain, we note high peaks for google queries regarding Cristiano Ronaldo, Sergio Busquets, Harry Kane, Neymar, Eden Hazard and Kevin De Bruyne. In the final part of the domain, there is a slight decrease in the average number of total google queries for every football player. However, at the end of the domain, Figure 2 highlights that the curves tend to increase once again, e.g. see Karim Benzema, Kevin De Bruyne and Eden Hazard. Many important players are not very requested by the world wide web despite their notoriety, e.g. Sergio Agüero, Pierre-Emerick Aubameyang, Karim Benzema. This may be an interesting insight for football clubs' marketing strategies because, although the above players are very expensive, they do not arouse the same interest of other colleagues from the world wide web. A further fascinating and visible detail is that some players such as Antoine Griezmann and Harry Kane, are characterized by a very hight variability during the day whereas the great part of the other curves shows a more constant behaviour over the whole domain.
Figure 3 shows the functional mean and variance of the 24 curves. The func-

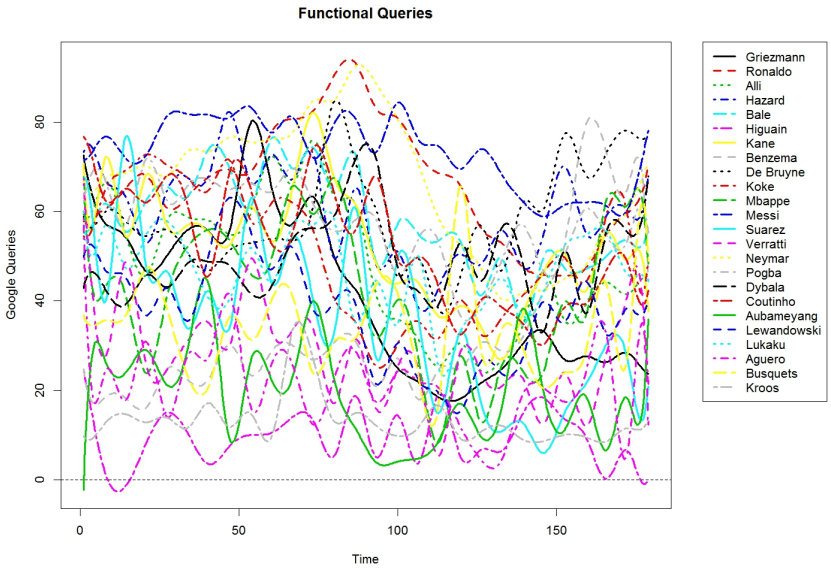

Fig. 2: Spline Approximation of the Google Queries for each Football Top Player

tional mean oscillates between 31.79 and 54.55 and the maximum is achieved

at 21:00 (Italian time zone); conversely, the lowest values correspond to the night hours, i.e. from 3:32 to 7:32 am. Considering that we are evaluating football players, who play in an European competition (UEFA Champion League), these results are are in line with our expectations. Regarding the functional variance, we observe an hight peak (570) in the middle part of the domain; also this circumstance has an immediate interpretation because in the first part of the night (23:00-2:00) many people are sleeping but others surf the web.
Figures 4 and 5 provide the first and second derivatives of the spline approx-

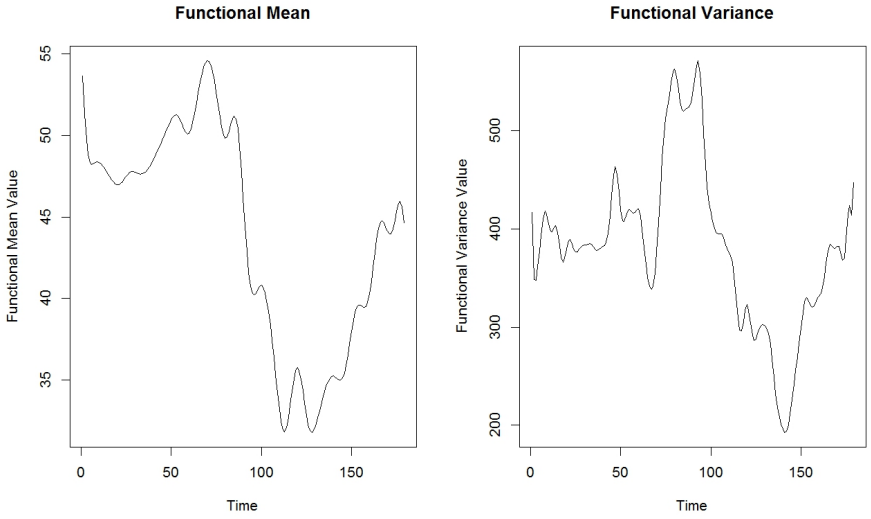

Fig. 3: Functional Mean and Functional Variance of Functional Data

imations of the google queries for each football player. The interest in these two plots relies on their different interpretations with respect to Figure 2. Indeed, the latter shows the total number of queries for each player whereas the derivatives highlight the velocity and acceleration in searching players' names on google search engines, respectively. For both derivatives, we observe that in the extreme parts of the domain, there are natural strong increase or decrease of curves due to the spline approximation. In both derivatives graphs, we can check that Sergio Busquets, Luis Suárez, Pierre-Emerick Aubameyang and Sergio Agüero are characterized by first and second derivatives oscillations which are higher than other players. This emphasizes that, in specific part of the day, there is a growing interest (followed by decrease of interest) for these players on the web. This aspect underlies that the velocity or acceleration in the number of queries do not denote a general average interest in these players but probably depends on news or information which have been spread on the world wide web in previous hours. An interesting clue for demonstrating

our interpretation is that some top players such as Lionel Messi and Cristiano Ronaldo, are characterized by first and second derivatives that are almost flat (probably because world wide web users constantly look for their names all over the day).

Functional principal components (FPCs) are computed using Equation (3)

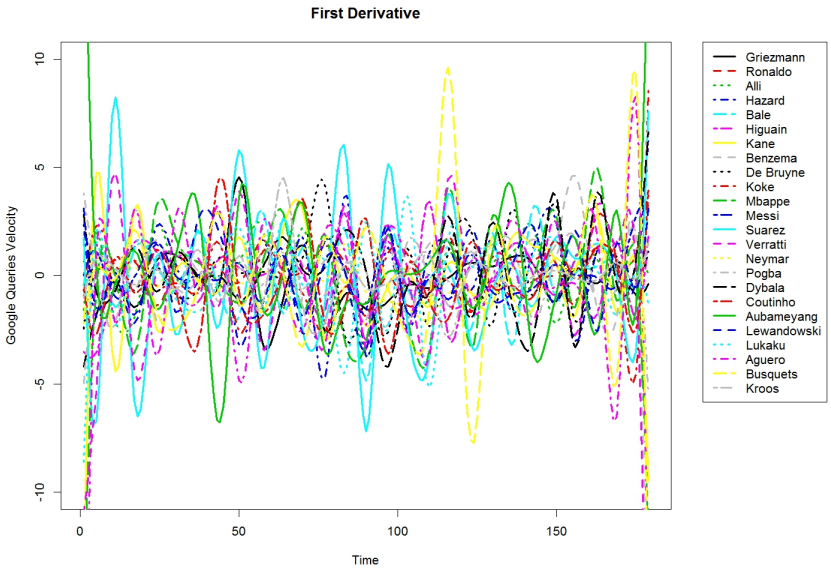

Fig. 4: First Derivative of the Spline Approximation of Google Queries for each Football Top Player

and showed in Figure 6. We observe that the first three FPCs explain 89.13 % of the total variability of the functional data. The graph displays that each FPC allows us interpreting a specific part of the time domain. Indeed, the first FPC mainly represents afternoon and evening (the first half of the domain- from 13:24 to 00:52), the second FPC is indicative of morning (from 5:40 to 13:16) and, finally, the third FPC explains the deep night hours (from 1:00 to 5:00). This figure also presents the bi-plot graphs among principal components for interpreting the relationship between players and couples of FPCs.

Figures 7, 8, 9, and Table 2 display the results of the functional k-means clustering according to the semi-metric FPC distance. The latter is computed by considering the FPCs decomposition of the spline approximation, first and second derivatives, respectively. The k-means algorithm is implemented by choosing three initial random centroids that are computed following Equation (6). Every clustering result has its own interpretation. Specifically, when we focus on the spline approximations of the functions, we obtain three different groups that can be viewed as expression of a ranking among players according

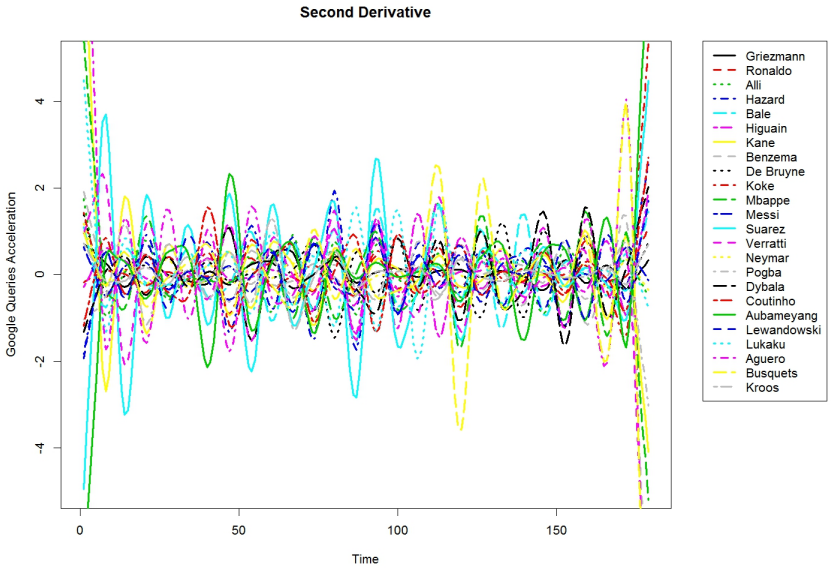

Fig. 5: Second Derivative of the Spline Approximation of Google Queries for each Football Top Player

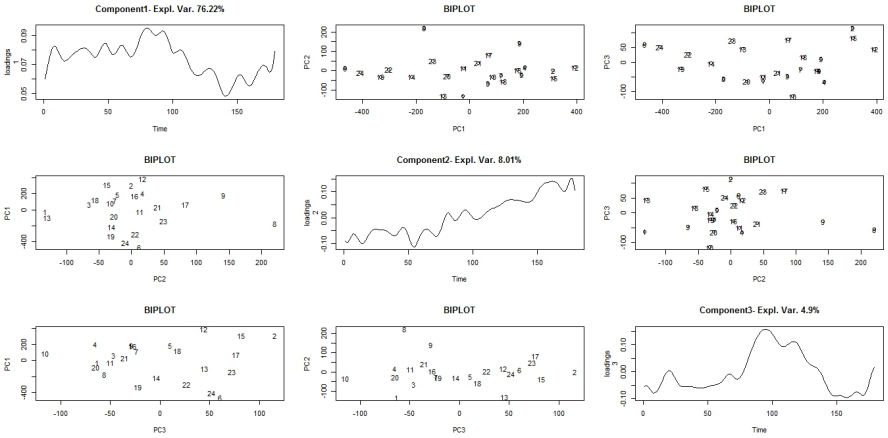

Fig. 6: First Three Functional Principal Components based on the Functional Principal Component Decomposition of the Spline Approximation of the Google Queries for each Football Top Player

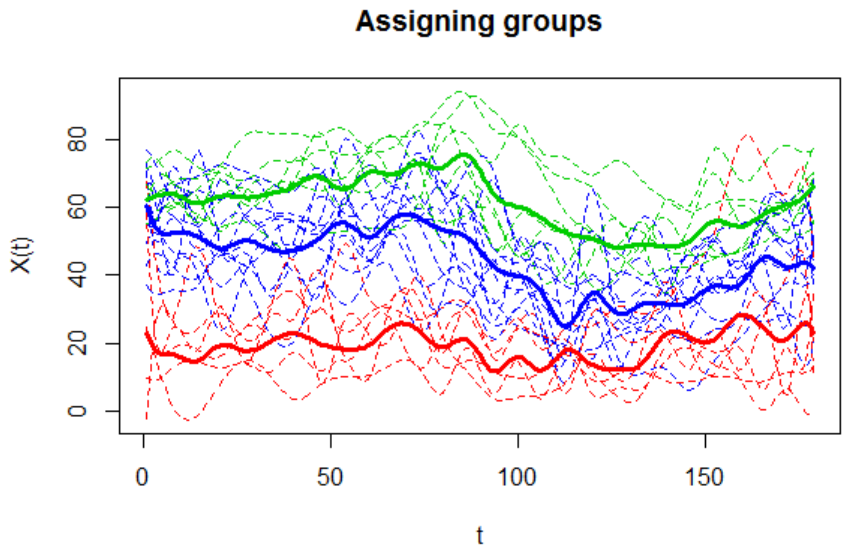

Fig. 7: Results of k-means Clustering of the Spline Approximation of the Google Queries for each Football Top Player according to the Functional Principal Components Decomposition. Red curves: group n.1; Green curves: group n. 2, Blue curves: group n. 3

to the number of visit during the day. In effect, the first group (red curves in Figure 7) is composed by those players which are less searched in google search engines (e.g. Gonzalo Higuaín, Karim Benzema, Marco Verratti, Pierre-Emerick Aubameyang, Sergio Agüero and Toni Kroos). Conversely, the second group (green curves in Figure 7) represents the most "fashionable" players, such as Cristiano Ronaldo, Eden Hazard, Gareth Bale, Kevin De Bruyne, Lionel Messi, Neymar, and Paul Pogba. Finally, the third group (blue curves in Figure 7) is made up by intermediate "interesting players" (see the first column of Table 2). These results underline significant insights. As we could expect, top players, such as Ronaldo, Messi, Neymar, Bale, etc., are important investments for their hight impact from the communication and merchandising perspectives because they are very requested by users. However, our findings stress that many players with high market values do not present the same powerful influence on the web. For example, Tony Kross, whose market value is equal to 70.000 millions of euros, and is an important player of Real Madrid CF, has received little interest in google queries.
Figures 8 and 9 have different interpretations with respect to Figure 7 because they are based on derivatives. Thus, they do not consider the total number of searches but the velocity and acceleration in google queries trends over the day. This means that higher values (positive or negative) of curve derivatives

Player name	First Group membership	Second Group membership	Third Group membership
Antoine Griezmann	3	2	2
Cristiano Ronaldo	2	3	2
Dele Alli	3	3	2
Eden Hazard	2	3	3
Gareth Bale	2	3	2
Gonzalo Higuaín	1	2	2
Harry Kane	3	2	1
Karim Benzema	1	2	1
Kevin De Bruyne	2	2	2
Koke	3	3	2
Kylian Mbappé	3	2	1
Lionel Messi	2	3	3
Luis Suárez	3	2	3
Marco Verratti	1	2	3
Neymar	2	1	2
Paul Pogba	2	3	2
Paulo Dybala	3	1	2
Philippe Coutinho	3	2	2
Pierre-Emerick Aubameyang	1	3	2
Robert Lewandowski	3	3	3
Romelu Lukaku	3	1	1
Sergio Agüero	1	1	1
Sergio Busquets	3	1	1
Toni Kroos	1	3	3

Table 2: Group Membership of to 24 Football Top Players

provide interesting information about possible anomalies of trends. Indeed, continuous fluctuations of these curves reveal that the corresponding football player is searched on the web only in specific instants of time. On the contrary, flat curves reflect a constant interest during the whole day. Hence, the cluster analysis based on first and second derivatives, identifies groups that are similar according to the above aspect. Indeed, we can observe that players such as Hazard, Lewandowski and Kross, belong to the group number three (blue curves in Figures 8 and 9), which is characterized by functions that are more flat than those of the group number one (red curves in Figures 8 and 9). This circumstance may be interpreted as follows: players in the third group are constantly (intensely or not) searched on the web during the day without shocks of the corresponding curves. On the contrary, players in the first group are characterized by high variability in the trend of the google queries. Therefore, the clustering membership based on derivatives is expression of similar patterns of google trends according to the continuity/discontinuity of interest that players arouse in the world wide web.

4 Discussion and Conclusions

In the last decades, the analysis of web data has become a very important task for a large number of companies, including football clubs. Also because soccer has become very popular and generates huge amounts of revenue and inspire countless hours of debate and conversation among fans and insiders, soccer analytics have been attracting the attention of worldwide practitioners and scholars due to many economic and non-economic interests. Purchasing notorious players is no more a simple matter of football but it has become a strategy of marketing. The acquisition of certain players may stimulate the

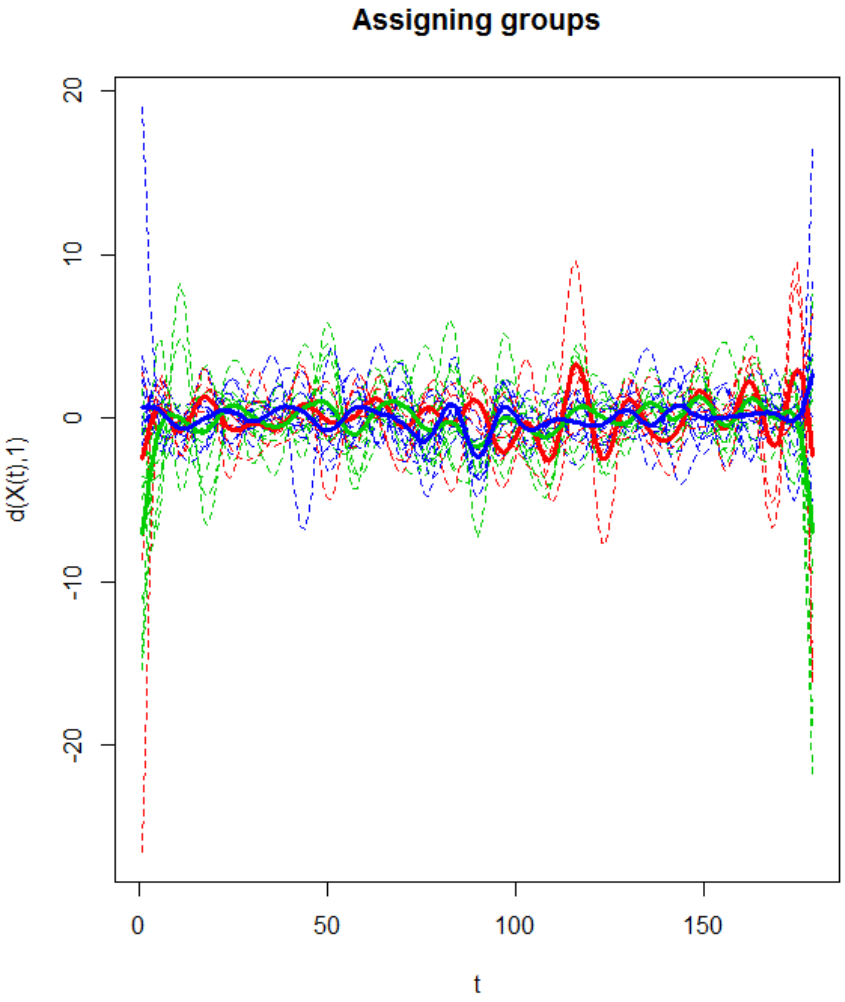

Fig. 8: Results of k-means Clustering of the First Derivatives of the Google Queries for each Football Top Player according to the Functional Principal Components Decomposition. Red curves: group n.1; Green curves: group n. 2, Blue curves: group n. 3

interest around a team and is able to move huge capitals. Due to the increasing commercialization of soccer, research on this topic appears regularly in the recent literature.

The enormous attention paid to soccer and football players in today's society, where technology is king, is automatically reflected in search engine queries. However, internet traffic is generated by people all around the world,

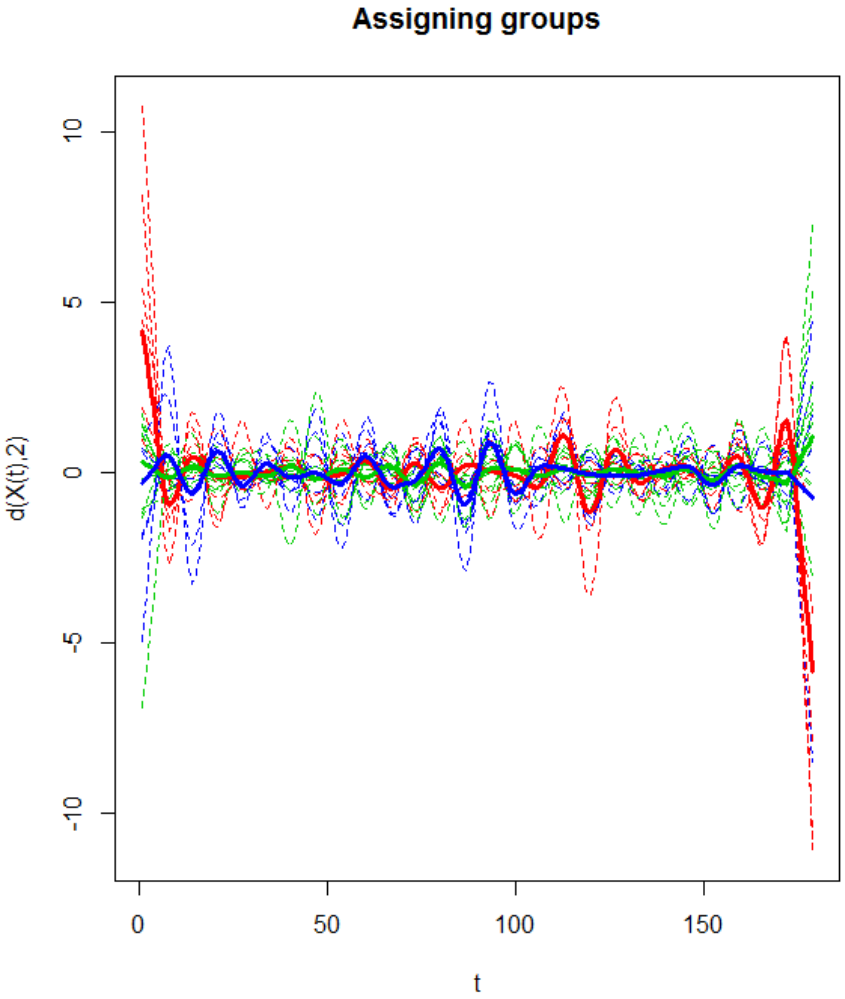

Fig. 9: Results of k-means Clustering of the Second Derivatives of the Google Queries for each Football Top Player according to the Functional Principal Components Decomposition. Red curves: group n.1; Green curves: group n. 2, Blue curves: group n. 3

and thus analysing this kind of data means to deal with data streaming and big data. Hence, football clubs and insiders need advanced statistical tools which are able to capture, synthesize, and illustrate the huge amount of information that is generated by the world wide web.

The main characteristics of data streaming are that data continuously flow, and their size is extremely large and potentially infinite. For this reasons, this

research proposes an original approach for clustering daily google queries using the functional principal component decomposition and functional k-means. Given that the simple total number of visits is not a useful indicator of the "charm" of a player because it could depend on extreme and random events, we believe that the analysis of trends over time bring more interesting insights. Moreover, this approach has also the advantage of reducing the dimension of the huge amount of data with the conversion of data vectors into functions, and allow us clustering statistical units by considering the whole domain of reference (e.g. the day). Naturally, extending the time horizon being analyzed, the advantage of this approach increases (e.g. estimating data of months or years).

Future developments of this research will certainly focus on a dynamic approach and the elaboration of an algorithm for the automatic choice of the optimal number of clusters, which should be able of self-adapting in a dynamic framework.

References

1. Howington E, Moates K. Is there a bye week advantage in college football? Electronic Journal of Applied Statistical Analysis 2017; 10(3)
2. Kern A, Schwarzmann M, Wiedenegger A. Measuring the efficiency of english premier league football. Sport, Business and Management: An International Journal 2012; 2(3): 177-195
3. Zelenkov Y, Solntsev I. Measuring the effciency of russian football premier league clubs. Electronic Journal of Applied Statistical Analysis 2017; 10(3)
4. Fin F, Iannario M, Piccolo D, Simone R. The effect of uncertainty on the assessment of individual performance: empirical evidence from professional soccer. Electronic Journal of Applied Statistical Analysis 2017; 10(3)
5. Dobson S, Goddard J. The Economics of Football. Cambridge University Press. 2011
6. Guha S, Meyerson A, Mishra N, Motwani R, O'Callaghan L. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering 2003; 15(3):515-528
7. Ailon N, Jaiswal R, Monteleoni C. Streaming k-means approximation. In Advances in Neural Information Processing Systems, Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A (eds) 2009; 22: 10-18
8. Gama J. Knowledge discovery from data streams. Chapman Hall/CRC 2010
9. Silva J, Faria E, Barros R, Hruschka E, Carvalho A, Gama J. Data stream clustering: A survey. ACM Computing Surveys 2013; 46(1):13:1-13:31
10. Bouveyron C, Girard S, Schmid C. High-Dimensional Data Clustering. Computational Statistics and Data Analysis 2007; 52(1):502-519
11. Jain A. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 2009; 31:651-666
12. Aggarwal C, Han J, Wang J, Yu P. A framework for clustering evolving data streams. In Proceedings of the 29th International Conference on Very Large Data Bases 2003; 29: 81-92
13. Di Battista T, Fortuna F, Maturo F. BioFTF: An R package for biodiversity assessment with the functional data analysis approach. Ecological Indicators 2017; 73:726-732
14. Maturo F, Di Battista T. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. Ecological Indicators 2018; 84(C):70-81
15. Maturo F, Di Battista T, Fortuna F. R package BioFTF: Biodiversity assessment using functional tools 2016
16. Ferraty F, Vieu P. Nonparametric functional data analysis. Springer, New York 2006

17. Ramsay JO, Silverman BW. Functional Data Analysis, 2nd edn. Springer, New York, 2005
18. Di Battista T, Fortuna F, Maturo F. Parametric functional analysis of variance for fish biodiversity assessment. Journal Of Environmental Infor- matics 2016; 28(2):101-109
19. García MLL, García-Ródenas R, Gómez AC. K-means algorithms for functional data. Neurocomputing 2015; 151:231-245
20. Di Battista T, Fortuna F. Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). Statistical Analysis and Data Mining 2017; 10:21-28
21. Maturo F, Migliori S, Paolone F. Do Institutional or Foreign Shareholders Influence National Board Diversity? Assessing Board Diversity Through Functional Data Analysis. Springer International Publishing 2017: 199-217.
22. Di Battista T, Fortuna F, Maturo F. Parametric functional analysis of variance for fish biodiversity. In International Conference on Marine and Freshwater Environments, iMFE 2014
23. Di Battista T, Fortuna F. Clustering dichotomously scored items through functional data analysis. Electronic Journal of Applied Statistical Analysis 2016; 9:433-450
24. Yan H, Paynabar K, Shi J. Real-time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition. Technometrics 2017
25. Ramsay JO. Monotone regression splines in action (with discussion). Statistical Science 1988; 3:425-461
26. Ramsay JO., Winsberg S. Maximum marginal likelihood estimation for semiparametric item analysis. Psychometrika 1991; 56:365-379
27. Ramsay JO. Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika 1991; 56:611-630
28. Gattone SA, Rocci R. Clustering curves on a reduced subspace. Journal of Computational and Graphical Statistics 2012; 21(2):361-379
29. Argiento R, Bissiri PG, Pievatolo A, Scrosati C. Multilevel functional principal component analysis of faAxade sound insulation data. Quality and Reliability Engineering International 2015; 31(7):1239-1253
30. Febrero-Bande M, de la Fuente M. Statistical computing in functional data analysis: The r package fda.usc. Journal of Statistical Software 2012; 51(4):1-28
31. Escabias M, Aguilera A, Valderrama M. Principal Component estimation of functional logistic regression: discussion of two different approaches. Nonparametric Statistics 2004; 16(3-4):365-384
32. Aguilera A, Aguilera-Morillo M. Penalized PCA approaches for b-spline expansions of smooth functional data. Applied Mathematics and Computation 2013; 219(14):7805-7819