RESEARCH ARTICLE

Forecasting basketball players' performance using sparse functional data

Guillermo Vinué1 | Irene Epifanio2

1Department of Statistics and O.R., University of Valencia, Burjassot, Spain
2Departamento de Matemátiqus y Institut de Matemàtiques i Aplicacions de Castelló, Universitat Jaume I, Castelló, Spain

Correspondence Guillermo Vinué, Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Spain. Email: guillermo.vinue@uv.es

Funding information Ministerio de Ciencia, Innovación y Universidades (AEI/FEDER, UE), Grant DPI2017-87333-R; Universitat Jaume I, UJI-B2017-13

Abstract
Statistics and analytic methods are becoming increasingly important in basketball. In particular, predicting players' performance using past observations is a considerable challenge. The purpose of this study is to forecast the future behavior of basketball players. The available data are sparse functional data, which are very common in sports. So far, however, no forecasting method designed for sparse functional data has been used in sports. A methodology based on two methods to handle sparse and irregular data, together with the analogous method and functional archetypoid analysis is proposed. Results in comparison with traditional methods show that our approach is competitive and additionally provides prediction intervals. The methodology can also be used in other sports when sparse longitudinal data are available.

KEYWORDS
archetypal analysis, basketball, forecasting, functional data analysis, functional sparse data

1 | INTRODUCTION

The statistical analysis of sports is a fast-growing field. In particular, sports forecasting is a strongly expanding field. Proof of this is the increasing number of forecasting methods developed covering several sports and sportive topics [1]. Certainly, in professional sports, getting any notion in advance about the future performance of players can have important consequences on the roster composition in terms of renewing or buying players, or in terms of establishing levels of remuneration in function of that performance, for instance. One of the sports that has been revolutionized by data analytics is basketball. Basketball analytics started to attract attention with the publications in refs. [2,3]. More recently, other papers and books have been released [4–6]. Technological advances have made it possible to collect more data than ever about what is happening on the field, requiring new methods of analysis. There is currently a need for innovative methods that exploit the full potential of the data and that make it possible to generate additional value for athletes and technical staff. Here too, one of the main challenges is to use past performance to predict future performance [5]. To address this open question, some forecasting methods have been developed. Following ref. [7, Chapter 1.4], two main approaches can be distinguished based on the type of data used: time-series forecasting and cross-sectional forecasting. On the one hand, forecasting using data collected over time describes the likely outcome of the time series in the immediate future, based on knowledge of recent outcomes. On the other hand, cross-sectional forecasting methods use data collected at a single point in time. The goal here is to predict a target variable using some explanatory variables which are related to it.

Two well-elaborated methods can be found using historical time data: (a) College Prospect Rating, which is a score assigned to college basketball players that attempts to estimate their NBA potential [5,8]; (b) Career-Arc Regression Model Estimator with Local Optimization (CARMEL0), which, for a player of interest, identifies similar players throughout NBA history and uses their careers to forecast the future player’s activity [9].
Regarding cross-sectional models, a Weibull-Gamma with covariates timing model is proposed in ref. [10] to predict the points scored by players over time. In this case, the time variable is years playing in NBA. Another interesting approach is presented in ref. [11], where correlations and regression models are computed to figure out which foreign players will be successful in the NBA, by using their previous statistics in international competitions.

In addition to the effort of predicting individual performance, there have also been other approaches focusing on teams and other features of the game. Some models using simulation have been developed to forecast the outcome of a basketball match [12,13]. A comparison between predictions based on NCAAB and NBA match data is discussed in ref. [14]. A dynamic paired comparison model is described in ref. [15] for the results of matches in two basketball and football tournaments. Furthermore, in ref. [16] a process model is used with player tracking data for predicting possession outcomes.

We wish to consider a new perspective by using functional data analysis (FDA) in sports. FDA is a relatively new branch of Statistics that analyzes data drawn from continuous underlying processes, often time, that is, a whole function is a datum. Let us assume that $n$ smooth functions, $x_1(t), \ldots, x_n(t)$, are observed, with the $i$th function measured at $t_1, \ldots, t_m$ points. In our study, $x_i(t)$ represents the metric value of player $i$ for a certain age $t$. An important point we would like to emphasize here is that the time component of the FDA approach we are considering will represent players’ ages. As such, in this paper we propose different models for aging curves, which is a well-recognized and important topic within the more general area of forecasting player performance. As mentioned in ref. [9], the most important attribute of all, in terms of determining a player’s future career trajectory, is his/her age.

The goals of FDA coincide with those of any other branch of Statistics, and the classical summary statistics can be also defined, such as the mean function $\bar{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t)$, the variance function $\text{var}_X(t) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(t) - \bar{x}(t))^2$ and the covariance function $\text{cov}_X(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} (x_i(t_1) - \bar{x}(t_1))(x_i(t_2) - \bar{x}(t_2))$. An excellent overview of FDA is found in ref. [17], while methodologies for studying functional data nonparametrically are found in ref. [18]. Ramsay et al [19] introduce related software and Ramsay and Silverman [20] present some interesting applications in different fields. Other recent applications include refs. [21,22]. In all these problems, a continuous function lies behind these data even though functions are sampled discretely at certain points. The FDA framework is highly flexible because the sampling time points do not have to be equally spaced and both the argument values and their cardinality can vary across cases. When functions are observed over a relatively sparse set of points, we have sparse functional data. An excellent survey on sparsely sampled functions is provided in ref. [23].

As regards the forecast of functional time series, there is a body of research, such as refs. [24–26], where functions are measured over a fine grid of points. However, only a few works deal with the problem of forecasting sparse functional data [27]. Notice that when functions are observed over a dense grid of time points, it is possible to fit a separate function for each case using any reasonable basis. Nevertheless, in the sparse case, this approach fails and the information from all functions must be used to fit each function. This is because each individual has been observed at different time points. Therefore, any fixed grid that is formed will contain many missing observations for each curve. A very good explanation of this issue can be found in ref. [23].

Sports data are sparse and irregular. They are sparse because most players do not have a very long career in the same league. And they are irregular because each player’s career lasts for a different length of time. Despite the fact that time series data or movement trajectories are very common in sports, FDA has been mostly used in sport biomechanics or medicine [28,29]. To the best of our knowledge, there are only two references about sports analytics using FDA. In ref. [30], FDA was introduced for the study of players’ aging curves and both hypothesis testing and exploratory analysis were performed. Vinué and Epifanio [31] extended archetypoid analysis (ADA) for sparse functional data (see also refs. [32,33]), showing the potential of FDA in sports analytics. In particular, in ref. [31, section 5.2] it was demonstrated that advanced analysis with FDA reveals patterns in the players’ trajectories over the years that could not be discovered if data were simply aggregated (eg, averaged). We take advantage of this fact and continue the work done in ref. [31, section 5.2] with a view to predict how players will evolve.

In this paper, we propose a methodology to predict player’s performances using sparse functional data. Two metrics will be analyzed: Box Plus/Minus (BPM)1 and Win Shares (WS).2 Analysis using BPM will allow us to establish a plausible comparison with CARMELO, while analysis with another variable such as WS will allow us to evaluate differences in career arcs. To that end, we will focus on two existing methods designed to handle sparse and irregular data: (a) regularized optimization for prediction and estimation with sparse (ROPES) data, originally developed by Alexander Dokumentov and Rob Hyndman [27,34]; (b) principal components analysis through conditional expectation (PACE), originally developed by Fang Yao, Hans-Georg Müller and Jane-Lin Wang [35].

1https://www.basketball-reference.com/about/bpm.html
2https://www.basketball-reference.com/about/ws.html
Our methodology will also involve using the method of analogues based on functional archetypoid analysis (FADA), which will allow us to refine predictions for the players of interest and to achieve a more reliable forecast, in line with the expectations of basketball analysts. We will apply them to a very comprehensive database of NBA players. Results will be obtained using the R software [36].

As noted earlier, forecasting future performance is also very relevant to other sports (see for instance ref. [37]). We would like to emphasize that our methodology can also be used in other sports when sparse longitudinal data are available. Data and R code (including a web application created with the R package shiny [38]) to reproduce the results can be freely accessed at https://www.uv.es/vivigui/software. The rest of the paper is organized as follows: Section 2 reviews ROPES, PACE, ADA and FADA. Section 3 will be concerned with the data and input variables used. Section 4 presents two analyses: (a) ROPES and PACE are compared with each other and with standard benchmarks and (b) the reliability of ROPES predictions for current players using the method of analogues with FADA is shown. A comparison with CARMELO results is also provided and the implications of the archetypoid coefficients are discussed. The paper ends with some conclusions in Section 5. An appendix, supporting information contains a validation study to choose an optimal blend of tuning parameters which ROPES depends on, which is crucial for good predictive activity. It also shows how this methodology can also be proposed for forecasting international players. The appendix is available as an online supplement.

2 | METHODOLOGY

2.1 | Regularized optimization for prediction and estimation with sparse

The method ROPES, proposed by Alexander Dokumentov and Rob Hyndman [27,34], solves problems involving decomposing, smoothing and forecasting two-dimensional sparse data. In practical terms, where the aim is to interpolate and extrapolate the sparse longitudinal data, made up of \( n \) observations, and presented over the time dimension with \( m \) time points, the following optimization problem is solved:

\[
\{(U, V)\} = \arg \min_{U,V} \|W \odot (Y - UV^T)\|^2 + \|U\|^2 \\
+ \|\text{DIFF}_2(m, \lambda_2)V\|^2 + \|\text{DIFF}_1(m, \lambda_1)V\|^2 \\
+ \|\text{DIFF}_0(m, \lambda_0)V\|^2,
\]

where:
- \( Y \) is an \( n \times m \) matrix.
- \( U \) is an \( n \times k \) matrix of “scores” (“coefficients”), \( k = \min(n, m) \).

\( \bullet \) \( V \) is an \( m \times k \) matrix of “features” (“shapes”).
\( \bullet \) \( \| \cdot \| \) is the Frobenius norm.
\( \bullet \) \( \odot \) is the element-wise matrix multiplication.
\( \bullet \) \( W \) is an \( n \times m \) “masking matrix” of weights.
\( \lambda_0, \lambda_1 \) and \( \lambda_2 \) are smoothing parameters.

2.2 | Principal components analysis through conditional expectation

Functional principal components analysis (FPCA) is a common tool to reduce the dimension of data when the observations are random curves. The usual computational methods for FPCA based on function discretization or basis function expansions are inefficient when data with only a few repeated and sufficiently irregularly spaced measurements per subject are available. As a quick reminder, when functions are measured over a fine grid of time points, a separate function for each individual can be used. In the sparse case, however, the information from all functions must be used to fit each function.

A version of FPCA, in which the FPC scores are framed as conditional expectations, was developed by Fang Yao, Hans-Georg Müller and Jane-Lin Wang to overcome this issue [35]. This method was referred to as PACE for sparse and irregular longitudinal data. In practice, the prediction for the trajectory \( X_i(t) \) for the \( i \)th subject, using the first \( p \phi_q \) eigenfunctions, is:

\[
\hat{X}_i^p(t) = \hat{\mu} + \sum_{q=1}^{p} \xi_{iq}\phi_q(t),
\]

where \( \hat{\mu} \) is the estimate of the mean function \( E(X(t)) = \mu(t) \) and \( \xi_{iq} \) are the FPC scores. PACE and its implementation in the R library fdapace [39] use local smoothing techniques to estimate the mean and covariance functions of the trajectories, specifically a local weighted bilinear smoother is used for estimating the covariance. Generalized cross validation is used for bandwidth choice, which is the default method.
for the FPCA function in the R library fdapace (default parameters are considered; eg, 10-folds and a Gaussian kernel are used). The number of components $p$ is determined using the fraction-of-variance-explained threshold (0.9999 by default) computed during the SVD of the fitted covariance function.

The eigenfunctions $\hat{\phi}_q(t)$ and the number $p$ are estimated with the training set, and they are used in the estimation of the scores for the test set. This is the procedure we will follow in Section 4.1. With the scores and the estimated eigenfunctions, we obtain an approximation of the trajectories and they can be used to predict unobserved portions of the functions. Yao et al. [35] also explain the construction of asymptotic pointwise confidence intervals for individual trajectories and asymptotic simultaneous confidence bands.

### 2.3 Archetypoid analysis

ADA was presented in ref. [33] and is an extension of archetypal analysis defined in ref. [40] (see refs. [41, 42] for other derived methodologies). In ADA, archetypes correspond to real observations (the so-called archetypoids). Let $X$ be an $n \times p$ matrix of real numbers representing a multivariate data set with $n$ observations and $p$ variables. For a given $g$, the objective of ADA is to find a $g \times p$ matrix $Z$ that characterizes the archetypal patterns in the data. In ADA, the optimization problem is formulated as follows:

$$RSS = \sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{g} \alpha_{ij}z_j \right\|^2 = \sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{g} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}x_l \right\|^2,$$

under the constraints

1. $\sum_{j=1}^{g} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, n$ and
2. $\sum_{i=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, g$, that is, $\beta_{jl} = 1$ for one and only one $l$ and $\beta_{jl} = 0$ otherwise.

Archetypoids are computed with the R package Anthropometry [43].

#### 2.3.1 ADA for sparse data with FDA

ADA was defined for functions in ref. [32], where it was shown that functional archetypoids can be obtained as in the multivariate case if the functions are expressed in an orthonormal basis, simply by applying ADA to the basis coefficients. When functions are measured over some sparse points, we have sparse functional data.

The basic idea of FADA is as follows. Based on the Karhunen-Loève expansion, the functions are approximated as in Equation (2). Because the eigenfunctions are orthonormal, to obtain FADA we can apply ADA to the $n \times p$ matrix $X$, with the scores (the coefficients in the Karhunen-Loève basis).
FIGURE 1   Box Plus/Minus (BPM) and Win Shares values for some players present in the database, at their corresponding ages. We have chosen these players because of two reasons: first, we wanted to represent the values of well-known players. Sabonis and Petrovic are two of the best international players of all time. Abdul-Jabbar and James are two of the best players in history. Embiid is one of the most promising players nowadays. Second, we wanted to fill the entire range of ages, both with short and long careers such as −86.7 for Gheorghe Muresan in 1998-1999\(^4\) or −49.3 for Mindaugas Kuzminskas in 2017-2018.\(^5\)

Figure 1 illustrates the type of data we are working with. It shows the observations of certain players, whose values are represented as connected points. Players’ ages will represent the time points to be used by our methodology. The initial range of ages in the database went from 18 to 44 years old. However, there were only a few measurements between ages 41 and 44, related to a few long-lasting players, so we have removed them. The age range finally considered is from 18 to 40 years old, that is, there are 23 time points. In total there are 3075 players.

4 | RESULTS

Section 4.1 contains a comparison analysis between ROPES, PACE and two benchmark methods. Section 4.2 discusses the implications of the archetypoid coefficients. Section 4.3 specifies the type of projections obtained. Section 4.4 presents an interactive web application.

4.1 | Comparison with other methods

In order to evaluate the usefulness of ROPES and PACE, we carry out a comparison with each other and with two benchmark methods, such as the average method and the naïve method. In the average method, the forecast of the next value is the mean of the previous values. In the naïve option, the forecast is the value of the last observation. They are two common simple alternatives to more advanced techniques [7, section 2.3].

In order to check the performance of all the methods, we have applied them to the test set of 385 players (see the validation study of the appendix to know how this set is created). Table 1 reports an extract of the results. It contains the following information for all players in the 2017-2018 season: (a) their age; (b) their actual BPM value; (c) the predictions with ROPES (using the optimal \(\lambda\) combination from the validation study of the appendix), PACE and the simple methods; (d) the squared difference between predictions and actual values (BPM\(_{\text{pr}}\) − BPM\(^2\)) and (e) the resulting total MSE (highlighted in bold). PACE obtains the smallest MSE, followed by ROPES. It is interesting to note that the mean BPM obtained with the naïve method is practically the same as the actual one (both rounded to \(-0.92\)). This result is most probably because over- and under-predictions cancel each other out.

Figure 2 displays the boxplots for the actual BPM values together with the BPM predictions for each method in different intervals.

The intervals \([-14, -9)\) and \([-9, -4)\) refer to players with a very bad performance (according to the BPM scale). We see in both cases that the predictions are far away from the true values. All the methods give a conservative forecast for such extreme values. PACE is the method that provides the closest results in these two intervals. ROPES is close to PACE in \([-9, -4)\). In the interval \([-4, 1)\), the four methods show similar values with respect to the actual ones. In

\(^4\)https://www.basketball-reference.com/players/m/muresgh01.html\(\backslash_{\text{all}}\)\(\backslash_{\text{advanced}}\)

\(^5\)https://www.basketball-reference.com/players/k/kuzmimi01.html\(\backslash_{\text{all}}\)\(\backslash_{\text{advanced}}\)
| Player        | Age | BPM  | ROPES (900, 10, 10) | PACE | Average | Naïve |
|--------------|-----|------|---------------------|------|---------|-------|
|              |     |      | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  | BPM<sub>pr</sub>  |
|              |     |      | BPM<sub>pr</sub>  | (BPM<sub>pr</sub> – BPM)<sup>2</sup> | BPM<sub>pr</sub>  | (BPM<sub>pr</sub> – BPM)<sup>2</sup> | BPM<sub>pr</sub>  | (BPM<sub>pr</sub> – BPM)<sup>2</sup> | BPM<sub>pr</sub>  | (BPM<sub>pr</sub> – BPM)<sup>2</sup> | BPM<sub>pr</sub>  | (BPM<sub>pr</sub> – BPM)<sup>2</sup> |
| Aaron Brooks | 33  | −4.3 | −4.14              | 0.03  | −3.32   | 0.96  | −2.21   | 4.37  | −4.6  | 0.09  |
| Josh Richardson | 24  | 1.4  | 0.66               | 0.55  | 0.56    | 0.71  | 0.40    | 1.00  | 0.2   | 1.44  |
| Zaza Pachulia | 33  | 0.8  | 1.00               | 0.04  | 0.54    | 0.07  | −0.75   | 2.40  | 2.7   | 3.61  |
| Mean (MSE)   |     | −0.92| −0.61              | 6.73  | −0.87   | 3.24  | −1.15   | 7.59  | −0.92 | 7.11  |
| SD           |     | 3.31 | 3.01               | 16.28 | 2.35    | 7.6   | 2.72    | 15.41 | 3.22  | 16.53 |
| Mean ± SD    |     | (2.39, 4.23) | (2.4, 3.62) | (1.48, 3.22) | (−3.87, 1.57) | (−4.14, 2.3) |

Note: The difference between actual and predicted values and MSE are also provided. MSE is highlighted in bold. Extract of the results.
Abbreviations: BPM, Box Plus/Minus; PACE, principal components analysis through conditional expectation; ROPES, regularized optimization for prediction and estimation with sparse.
the interval \([1, 6)\), ROPES gives the most similar predictions with respect to the actual ones. In the interval \([6, 11)\), again ROPES and PACE give the most accurate predictions. Remarkably, the naïve method shows outliers in all the intervals. Figure 2 is showing that PACE and ROPES tend to overestimate bad players and underestimate good ones. This can very helpful for team managers in their search for future stars (this additional interpretation was suggested by one of the referees).

Overall, PACE is the method that performs best. ROPES is able to beat the simple benchmark methods, showing an improvement with respect to them. The main drawback of the current PACE implementation is the lack of prediction intervals. The main goal of this paper is to draw attention to the added value of using an FDA approach to forecast players’ performance, which has not been done so far. Therefore, even though PACE should give somewhat more accurate predictions than ROPES, in next section we will use ROPES to forecast future players’ activity because it does provide prediction intervals. Prediction intervals are very helpful and important because they express how much uncertainty is associated with the forecast.

4.2 Implications of the archetypoid coefficients

We have analyzed a total of eight players from different teams, namely Devin Booker, Clint Capela, Joel Embiid, Nikola Jokić, Tyus Jones, Zach LaVine, Donovan Mitchell and Jayson Tatum. They are representing several career statuses. Embiid, Jokić, Mitchell and Tatum are already established figures (especially Embiid and Jokić). Booker and Capela are a step below the super stars but they are also very good players. Something similar could be said about Tyus Jones, who is constantly improving his skills. Finally, LaVine is an offensive specialist.

In a first attempt to compute predictions using all the players of the data set, we realized that the ROPES method had some pull towards the mean of the entire sample (like the other methods discussed in Section 4.1 but not as strong as them). This gave unrealistic performance predictions for both the best and most promising players. Therefore, in order to refine predictions, it is much more suitable to use the so-called “method of analogues.” The idea is to find players related to the one of interest and then use their documented activity to obtain the predictions. We know how other players already performed, so we can use their information to gain an approximate idea about the future performances of others. The method of analogues has been used for years in fields such as climatology [45] and epidemiology [46]. Recently, an R package has been released that contains analogue methods for paleoecology [47]. The CARMELO method is also based on this scheme.

In order to find related players, we use ADA (see ref. [33] for theoretical details). ADA searches for extreme observations (the so-called archetypoids) to describe the frontiers of the data. In this technique, the BPM (and WS) function of a player is approximated by a mixture of archetypoids, which are themselves functions of boundary players (outstanding—positive or negative—performers). Archetypoids are specific players and the \(\alpha\) coefficients represent how much each archetypoid contributes to the approximation of each individual. The most comparable archetypoid should be the one corresponding to the largest value of the \(\alpha\) coefficients for the player of interest.

We choose the number of archetypoids for each metric following the screeplot explained by [33]. Five are selected for BPM and four for WS.
The archetypoids for the BPM metric are (their career BPM shown in brackets): Devin Gray (−8.4), Darryl Dawkins (−2.52), Diamond Stone (−24.1), Eddy Curry (−6.5) and LeBron James (9.21).

LeBron James is the representative of super star players. He is one of the best players in history. This is in line with the expected results since James has achieved the highest BPM values.

Darryl Dawkins represents the replacement level players (as a reminder, −2 is replacement level). Dawkins had a long NBA career. He was selected with the fifth pick in the 1975 NBA draft and played for 14 seasons, where he averaged double figures in scoring in nine of them. He leads the league in fouls committed in three seasons. In his case, his performance does not fit exactly with the replacement level description, but his average BPM does.

Eddy Curry, Devin Gray and Diamond Stone are representatives of players with a short-term career or with overall poor performance. Eddy Curry was selected fourth overall in the 2001 NBA draft and had a long NBA career. He led the NBA in field goal percentage in the 2002-2003 season but he did not really meet the expectations that his talent was indicating. Devin Gray had an irrelevant NBA career, playing a total of 27 games in two NBA seasons. Diamond Stone only played seven games in the NBA. From the basketball point of view, Devin Gray and Diamond Stone are related to the "bad" archetypal player profile. Both players played very little in the NBA. However, from the mathematical point of view, they are not exactly representing the same profile, since Devin Gray has a BPM of −8.4, while Diamond Stone has a BPM of −24.1 (the third worst BPM value in our database, the range of BPM values goes from −31.5 to 21). Therefore, Stone is representing even a much more extreme bad pattern than Gray. We acknowledge that the five archetypoids computed for BPM are not exactly representing all the players' typologies available in the database. In some cases, a greater number of archetypoids is needed to capture other players' profiles. Even though we have determined the optimal number of archetypoids with the screeplot, in a real situation it would be up to the analyst to decide how many representative cases to consider.

Regarding the WS metric, the archetypoids are (their career WS shown in brackets): Steve Burtt (0), Ben Wallace (5.84), Otis Birdsong (4.03) and LeBron James (14.6). Again, as expected, James is the representative of super star players. The fact that James is selected as the “best” archetypoid in both metrics is indicating how he can excel in many aspects of the game.

Otis Birdsong and Ben Wallace represent very good players. Otis Birdsong played 12 NBA seasons and appeared in 4 NBA All-Star Games. He was selected with the second pick of the 1977 NBA draft. Ben Wallace was very good at grabbing rebounds and blocking opponent shots. He won the NBA Defensive Player of the Year Award four times and won a championship with the Pistons in 2004.

Steve Burtt represents ordinary players. He played 101 games in four NBA seasons between 1984-1985 and 1992-1993.

Table 2 shows the $\alpha$ values for the eight players selected for the BPM and WS archetypoids.

As mentioned before, in ADA each datum is expressed as a mixture of actual observations (archetypoids). In particular, the $\alpha$ coefficients of each player are of great utility because they allow us to determine the composition of each player according to the archetypoid players, and to establish a clustering of similar players [31]. A discussion of the implications of the resulting archetypoid coefficients is given next. The BPM composition of Embiid is as follows. Embiid’s profile matches 49% of James’, 23% of Gray’s, 15% of Dawkins’ and 13% of Curry’s, so this is reflecting the fact that Embiid is on his way to become a super star like James is, but he still has some room of improvement. Regarding the WS composition, Embiid’s profile is 42% explained by Burtt’s, 29% by Birdsong’s, 18% by Wallace’s and 11% by James’. In this case, the Embiid’s room for development is even more evident. The archetypoid coefficients for Jokic are quite impressive. His highest $\alpha$ is with James in both

| Player | BPM archetypoids | WS archetypoids |
|--------|------------------|-----------------|
|        | D. Gray | D. Dawkins | D. Stone | E. Curry | L. James | S. Burtt | B. Wallace | O. Birdsong | L. James |
| D. Booker | 0.32 | 0.12 | 0.07 | 0.13 | 0.36 | 0.70 | 0.12 | 0.08 | 0.10 |
| C. Capela | 0.00 | 0.08 | 0.33 | 0.10 | 0.49 | 0.00 | 0.14 | 0.76 | 0.10 |
| J. Embiid | 0.23 | 0.15 | 0.00 | 0.13 | 0.49 | 0.42 | 0.18 | 0.29 | 0.11 |
| N. Jokic | 0.06 | 0.19 | 0.00 | 0.07 | 0.68 | 0.19 | 0.03 | 0.35 | 0.43 |
| T. Jones | 0.31 | 0.14 | 0.07 | 0.12 | 0.36 | 0.58 | 0.12 | 0.24 | 0.06 |
| Z. LaVine | 0.39 | 0.15 | 0.05 | 0.11 | 0.30 | 0.77 | 0.10 | 0.07 | 0.06 |
| D. Mitchell | 0.37 | 0.12 | 0.00 | 0.12 | 0.39 | 0.59 | 0.08 | 0.12 | 0.21 |
| J. Tatum | 0.42 | 0.06 | 0.00 | 0.09 | 0.43 | 0.59 | 0.06 | 0.00 | 0.35 |

Abbreviations: BPM, Box Plus/Minus; WS, Win Shares.
variables. His BPM profile matches 68% of James' BPM profile (the highest similarity to James in this analysis by far) and his WS profile matches 43% of James' WS profile (only Tatum has a close value) and 35% of Birdsong's. This high similarity with respect to James implies that Jokic’s performance is already very good. Proof of his remarkable activity is that he has received his first All-Star and All-NBA First Team selections in 2018-2019 season. In addition, his 12 triple-doubles ranked second on the season behind only Russell Westbrook [6]. Capela is another player worth highlighting. His highest BPM coefficient is also with James, though his profile still matches 33% of Stones’, which indicates some shortcomings in his performance. Regarding his WS composition, his profile matches 76% of Birdsong’s, 14% of Wallace's and 10% of James’, which indicates a remarkable team productivity. On the other hand, the BPM profiles for Booker, Jones, Tatum and Mitchell are as close to James' as to Gray's. In terms of James’ coefficient, it is very difficult to approach his excellence, so we should not expect Booker, Jones, Tatum and Mitchell to achieve such an incredible level. However, the Gray's coefficient is also indicating that their performance is not very far away from an average player yet, so they still need to make further efforts to display a difference from competitors. Their WS profiles reaffirm this claim. Finally, the LaVine's BPM and WS profiles are most similar to Gray's and Burtt’s profiles, respectively, so he is not showing an outstanding performance at all.

4.3 Projections of future performance with ROPES and the method of analogues. Comparison with CARMELO

In order to select the cluster of analogous players, we first choose the archetypoid with the highest $\alpha$. As an illustration, Embiid's greatest similarity for BPM is with James. Then, the group of BPM similar players to Embiid is made up of James, together with the other players whose largest $\alpha$ coefficient is also for James and who have an $\alpha$ value greater than Embiid's $\alpha$. We will use this cluster to forecast the Embiid's future career arc. Current stars such as Chris Paul or Kevin Durant and stars of previous seasons such as Michael Jordan or Charles Barkley belong to this set. Likewise for WS.

The ROPES algorithm (with the lambda combination obtained in the validation study) is used to obtain $p$-forecasting intervals, where $p = 0.05$ is the selected significance level. We will discuss the ROPES predictions with the ones that the CARMELO 2018-2019 version provides.\(^6\) CARMELO is a basketball forecasting system released in the 2015-2016 season. Successive versions present some improvements [9]. To the best of our knowledge, it is the only publicly available projection system to compare our approach against. For each player of interest, CARMELO computes the similarity scores between that player and all historical players. To that end, it uses a number of statistics and players' attributes and a version of a nearest neighbor algorithm. The wins above replacement (WAR) metric is computed for all historical players with a positive similarity score. The forecast is given by averaging these WAR values.

WAR reflects a combination of a player's projected playing time and his projected productivity while on the court. Productivity is measured by a blend of two-thirds Real Plus-Minus (RPM) and one-third BPM. BPM was solely used to make the 2016–2017 forecasts, but the combination of RPM and BPM is used for the 2018–2019 forecasts (as in 2015-2016 and 2017-2018). According to the developers of CARMELO, the RPM/BPM blend seems to outperform BPM alone. The RPM statistic quantifies how much a player hurts or helps his/her team when (s)he is on the court. There has been some controversy regarding the validity of RPM, since the computations are not detailed.\(^7\) In fact, the CARMELO methodology cannot be replicated either. In addition, for seasons before 2000-2001, no RPM is available and CARMELO uses BPM only. The final point that needs to be made is that RPM is not available in our database. Therefore, we would like to draw the reader’s attention to the fact that our results are not directly equivalent to those of CARMELO, because the target variable is not exactly the same. However, both approaches should be complementary.

Figure 3 shows the ROPES forecast obtained for the eight players selected. In addition, for the sake of a convenient comparison with CARMELO, Figure 4 shows the screenshots of the CARMELO curves for the same eight players.

In Figure 3 we see that the predictions for Jokic show that his BPM and WS are expected to increase in the next three seasons and will remain quite high for several seasons (though they will go down from age 28). His lower and upper predictions are also high values. The width of the prediction intervals is constant over seasons. In general, the width of the intervals remains quite stable for most players. As a referee rightly mentioned, the intervals show several possible scenarios for some players, going from high to low values, so there is some uncertainty related to the point predictions. The CARMELO forecast for Jokic indicates some decrease in his performance, but still keeping high numbers (Figure 4D). The width of the CARMELO intervals fluctuates, the uncertainty in the ’20 season is bigger than in the ’19 and ’21 seasons. The BPM-ROPES forecast for Embiid shows that he will improve his BPM in the two coming seasons and then his performance will slowly decline. Regarding WS, it indicates a constant decrease over time. CARMELO also indicates that Embiid’s performance will increase within two seasons and

\(^6\)https://projects.fivethirtyeight.com/carmelo/

\(^7\)https://www.boxscoregeeks.com/articles/rpm-and-a-problem-with-advanced-stats
FIGURE 3  Box Plus/Minus (BPM) and Win Shares (WS) predictions for Devin Booker, Clint Capela, Joel Embiid, Nikola Jokic, Tyus Jones, Zach LaVine, Donovan Mitchell and Jayson Tatum using only the set of analogue players. Past values are in red and predictions are in green. The shaded area indicates the limits of the prediction intervals

then his values will decrease (the intervals move between narrow and wide intervals, Figure 4C). The BPM-ROPES forecast for Capela shows a constant increase in the coming years, keeping good values for many seasons. On the other hand, his WS values will constantly decrease. In this case, the intervals are widening over the years (especially in the BPM facet). The CARMELO forecast for Capela is a bit more conservative. Its prediction intervals are a bit wider in the '20 and '21 seasons than in the '19 season and then narrow a bit again (Figure 4B). The BPM-ROPES forecast for Devin Booker and Tyus Jones is somewhat similar to Capela’s. However, their WS-ROPES forecast is a flat arc around 0. The CARMELO predictions for these two players show a certain increase in their activity (Figure 4a,e). The width of the prediction interval for Booker remains quite constant. For Jones the width interval fluctuates between some wide and narrow ranges. The BPM-ROPES prediction for Donovan Mitchell and Jayson Tatum are similar to Jokic’s, though their values are not so outstanding. Their WS prediction is a flat arc around 0. The CARMELO predictions for Mitchell and Tatum also show a constant increase in their activity (Figure 4g,h). In both cases, the width of the prediction interval also fluctuates between some wide and narrow ranges. Finally, the BPM-ROPES forecast for LaVine shows a constant increase but keeping negative numbers, especially for the next 4 years. His WS prediction moves around 0. CARMELO also suggests an ordinary and flat performance in the coming seasons (Figure 4f). It is worth mentioning that some of the WS predictions stop at age 33 because this is the last age for which the set of analogous players shows values. For Mitchell, Booker and Tatum, WS-ROPES has been too conservative. Higher values would have been a bit more realistic, since everything seems to indicate that these three players will become very good players in the near future. Another aspect that demands a careful examination of the results is that the WS prediction for Booker climbs to about 0.5 in the last 2 years. These are some pitfalls worth highlighting for end-users of this methodology.

As a final point, it is important to remember that statistical models are not completely reliable for long-term forecasts, because the assumption that the future looks similar to the past slowly breaks down the further we go into the future. So the predictions should be constantly updated as new data becomes available.

4.4  Web application

Additionally, an interactive web application available at https://www.uv.es/vivigui/AppPredPerf.html allows the user
to represent the BPM and WS forecasting plots for every player in the 2017-2018 season under the age of 24 (154 players in total). A link to the CARMELO forecast for every player is also provided for easy comparison. The app gives some basic information about the way it works. It can also be generated from R with these two commands: library (shiny) ; runUrl('http://www.uv.es/vivigui/softw/AppPredPerf.zip').
5 | CONCLUSIONS

Basketball, like any other sport, contains a lot of uncertainty. A central issue is to predict future players’ performance using past observations. In spite of the fact that basketball data continue to expand and there is a constant demand for new techniques that provide objective information to help understand the game, there are not many publicly available projection systems. In this paper we have presented a methodology to deal with sparse functional data in order to forecast the basketball players’ performance. This has been done by analyzing ROPES and PACE and by including the method of analogues together with functional archetypoid analysis.

ROPES depends on several parameters, so we have carried out a validation study to choose an optimal combination that provides smooth curves and avoids overfitting (included in the appendix). The combination obtained works well to avoid narrow intervals and overconfident inferences. A comparison study has also been carried out to compare ROPES with PACE, and with simple alternatives, such as the average and naïve methods. PACE performed best overall and also in terms of runtime with respect to ROPES. However, unlike ROPES, it is not possible to obtain prediction intervals with its current computational implementation. In addition, ROPES also performed better than simple methods. Therefore, we have applied ROPES in the real case using data between 1973-1974 and 2017-2018 NBA regular seasons.

In the sparse case, information from all functions is used to fit each function, so all individuals contribute to a greater or lesser degree to form the estimations. To overcome this problem and to refine the predictions, we have used the so-called “method of analogues.” The idea is to relate a player’s curve to one of the possible types of players and then to predict his performance using only the information about these comparable athletes. In our case, the types of players are given by the archetypoids of the data set.

Once the computations are finished, an interactive web application shows the plots with the past and future behavior of 2017-2018 NBA players under the age of 24. Two variables have been analyzed: on the one hand, BPM is recognized as the most suitable metric to carry out an analysis involving historical data; on the other hand, WS is another widely used advanced metric. Adding a second variable allows us to examine differences in career arcs for different aspects of skill. Any other variable can be used. The predictions for eight players have been presented and a comparison with CARMELO has been done. The implications of the archetypoid coefficients have also been interpreted.

Player forecasting systems are important as a means of summarizing the overall match performance of individual players. Any forecasting method is limited because some aspects such as injury risk or work ethic, which influence future performance, are very difficult to quantify. However, coaches and experts can use these systems to review performances of their own players as well as tracking the performance levels of potential acquisitions. We hope that the approach presented here will provide valuable information about players’ overall ability to support decision making. Sparse functional data are very common in sports. Therefore, it is very reasonable to bring methods developed to deal with this kind of data to the field of sports. This methodology can serve as a starting point for further efforts in the same direction. One of the referees suggested us to remark the following two situations that our analysis has not considered: (a) the different amounts of playing time going into each averaged BPM and WS data points. In mathematical terms, this is a case of unequal variances, also called heteroscedasticity; (b) the pattern of sparsity in the data is not random, since players retiring or leaving the NBA should have low BPM and WS values in these age intervals. Both situations were formulated by the referee. We will consider them in future work. Following another referee’s suggestion, we will also try to compare the players’ forecasts using relative rankings in terms of their coefficients from a common archetypoid. The data and all R code are freely available at https://www.uv.es/vivigui/software for reproducibility and further exploration of the results.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Alexander Dokumentov and Rob Hyndman for kindly providing the R code to run the ROPES algorithm. The authors are also grateful to the anonymous referees and associated editor for their helpful criticism. G.V. worked on the first two revisions of the manuscript as a postdoctoral scholar holder in international mobility at KU Leuven, Department of Computer Science, Belgium. This work has been partially supported by the Spanish Ministerio de Ciencia, Innovación y Universidades (AEI/FEDER, UE) Grant DPI2017-87333-R and Universitat Jaume I, UJI-B2017-13.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

G.V. planned the design of the research, collected the data, performed the computations, interpreted the results and wrote the manuscript. I.E. planned the design of the research, performed the computations and wrote the manuscript.

ORCID

Guillermo Vinué https://orcid.org/0000-0003-2083-8276
REFERENCES

1. H. O. Stekler and L. Vaughan Williams, *Sports forecasting (special issue)*, Int. J. Forecast. 26(3) (2010), 1–3. https://doi.org/10.1016/j.ijforecast.2009.12.005.

2. D. Oliver, *Basketball on paper: Rules and tools for performance analysis*, Potomac Books, Inc, Washington, DC, 2004.

3. J. Hollinger, *Pro basketball forecast*, Potomac Books, Inc, Washington, DC, 2005.

4. J. Kubatko et al., *A starting point for analyzing basketball statistics*, J. Quant. Anal. Sports 3(3) (2007), 1–10. https://doi.org/10.2202/1559-0410.1070.

5. S. M. Shea, *Basketball analytics: Spatial tracking*, Createspace, Lake St. Louis, MO, 2014.

6. S. M. Shea and C. E. Baker, *Basketball analytics: Objective and efficient strategies for understanding how teams win*, Advanced Metrics, LLC, Lake St. Louis, MO, 2013.

7. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and practice*, OTexts, 2013, available at https://www.otexts.org/book/fpp.

8. S. M. Shea and C. E. Baker, 2013, available at http://www.basketballanalyticsbook.com/.

9. N. Silver, *CARMELO NBA player projections*, 2018, available at https://fivethirtyeight.com/features/our-nba-player-projections-are-ready-for-2018-19/, https://fivethirtyeight.com/features/whats-new-in-our-nba-player-projections-for-2017-18/, https://fivethirtyeight.com/features/whats-new-in-our-nba-projections-for-2016-17/, https://fivethirtyeight.com/features/how-were-predicting-nba-player-career/, https://projects.fivethirtyeight.com/carmelo/.

10. D. Hwang, 2012. *Forecasting NBA player performance using a Weibull-Gamma statistical timing model*, MIT Sloan Sports Analytics Conference, Boston, MA, 2012, pp. 1–10.

11. K. Salador, *Forecasting performance of international players in the NBA*, MIT Sloan Sports Analytics Conference, Boston, MA, 2011, pp. 1–18, available at http://www.sloansportsconference.com/wp-content/uploads/2011/08/Forecasting-Performance-of-International-Players-in-the-NBA.pdf.

12. E. Strumbelj and P. Vračar, *Simulating a basketball match with a homogeneous Markov model and forecasting the outcome*, Int. J. Forecast. 28(2) (2012), 532–542. https://doi.org/10.1016/j.ijforecast.2011.01.004.

13. P. Vračar, E. Strumbelj, and I. Kononenko, *Modeling basketball play-by-play data*, Expert Syst. Appl. 44 (2016), 58–66. https://doi.org/10.1016/j.eswa.2015.09.004.

14. A. Zimmermann, *Basketball predictions in the NCAAB and NBA: Similarities and differences*, Stat. Anal. Data Min. 9 (2016), 350–364. https://doi.org/10.1002/sam.11319.

15. M. Cattelan, C. Varin, and D. Firth, *Dynamic Bradley-Terry modelling of sports tournaments*, J. R. Stat. Soc. Ser. C. Appl. Stat. 62(1) (2013), 135–150. https://doi.org/10.1111/j.1467-9876.2012.01046.x.

16. D. Cervone et al., *A multiresolution stochastic process model for predicting basketball possession outcomes*, J. Amer. Statist. Assoc. 111(514) (2016), 585–599. https://doi.org/10.1080/01621459.2016.1141685.

17. J. O. Ramsay and B. Silverman, *Functional data analysis*, 2nd ed., Springer, New York, USA, 2005.

18. F. Ferraty and P. Vieu, *Nonparametric functional data analysis: Theory and practice*, Springer, New York, USA, 2006.

19. J. O. Ramsay, G. Hooker, and S. Graves, *functional data analysis with R and MATLAB*, Springer, New York, USA, 2009.

20. J. O. Ramsay and B. Silverman, *Applied functional data analysis. Methods and case studies*, Springer, New York, USA, 2002.

21. T. Di Battista and F. Fortuna, *Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy)*, Stat. Anal. Data Min. 10(1) (2017), 21–28. https://doi.org/10.1002/sam.11334.

22. H. D. Nguyen et al., *Whole-volume clustering of time series data from zebrasibra brain calcium images via mixture modeling*, Stat. Anal. Data Min. 11(1) (2018), 5–16. https://doi.org/10.1002/sam.11366.

23. G. James, *Sparseness and functional data analysis*, in *The Oxford handbook of functional data analysis*, Oxford University Press, Oxford, UK, 2010, 298–326.

24. A. Aue, D. Dubart Norinho, and S. Hörmann, *On the prediction of stationary functional time series*, J. Amer. Statist. Assoc. 110(509) (2015), 378–392. https://doi.org/10.1080/01621459.2014.909317.

25. R. J. Hyndman and M. Shahid Ullah, *Robust forecasting of mortality and fertility rates: A functional data approach*, Comput. Stat. Data Anal. 51(10) (2007), 4942–4956. https://doi.org/10.1016/j.csda.2006.07.028.

26. H. L. Shang and R. J. Hyndman, *Grouped functional time series forecasting: An application to age-specific mortality rates*, J. Comput. Graph. Stat. 26(2) (2017), 330–343. https://doi.org/10.1080/10618600.2016.1237877.

27. A. Dokumentov and R. J. Hyndman, *Low-dimensional decomposition, smoothing and forecasting of sparse functional data*, Working paper, 1-31, 2016, available at http://robjhyndman.com/papers/ROPES.pdf.

28. I. Epifanio et al., *Analysis of multiple waveforms by means of functional principal component analysis: Normal versus pathological patterns in sit-to-stand movement*, Med. Biol. Eng. Comput. 46(6) (2008), 551–561. https://doi.org/10.1007/s11517-008-0339-6.

29. A. J. Harrison, *Applications of functional data analysis in sports biomechanics*, 32 International Conference of Biomechanics in Sports, 2014, pp. 1–9.

30. A. Wakim and J. Jin, 2014. *functional data analysis of aging curves sports*, 2014, pp. 1–25, available at http://arxiv.org/abs/1403.7548.

31. G. Vinué and I. Epifanio, *Archetypoid analysis for sports analytics*, Data Min. Knowl. Discov. 31(6) (2017), 1643–1677. https://doi.org/10.1007/s11806-017-0514-1.

32. I. Epifanio, *Functional archetype and archetypoid analysis*, Comput. Stat. Data Anal. 104 (2016), 24–34. https://doi.org/10.1016/j.csda.2016.06.007.

33. G. Vinué, I. Epifanio, and S. Alemany, *Archetypes: A new approach to define representative archetypal data*, Comput. Statist. Data Anal. 87 (2015), 102–115. https://doi.org/10.1016/j.csda.2015.01.018.

34. A. Dokumentov, *Smoothing, decomposition and forecasting of multidimensional and functional time series using regularisation*, Ph.D. Thesis, Monash University. Faculty of Business and Economics. Econometrics and Business Statistics, 2016, available at http://arrow.monash.edu.au/vital/access/manager/Repository/monash:165926.

35. F. Yao, H.-G. Müller, and J.-L. Wang, *Functional data analysis for sparse longitudinal data*, J. Amer. Statist. Assoc. 100(470) (2005), 577–590. https://doi.org/10.1198/016214504000001745.
AUTHOR BIOGRAPHIES

Guillermo Vinué was born in Valencia (Spain) in 1985. He graduated in Mathematics (2008) and completed his PhD in Statistics and Optimization (2014), both at the University of Valencia. His research interests are concerned on computational statistics with applications in industry and sports analytics.

Irene Epifanio was born in Valencia (Spain) in 1975. She graduated in Mathematics (1997) and received her PhD in Statistics (2002), both at the University of Valencia. She is currently a Titular Professor at the Department of Mathematics at the Universitat Jaume I (Spain). Her current research interests include statistical learning, image analysis and functional data analysis.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Vinué G, Epifanio I. Forecasting basketball players’ performance using sparse functional data. Stat Anal Data Min: The ASA Data Sci Journal. 2019;12:534–547. https://doi.org/10.1002/sam.11436