Forecasting basketball players' performance using sparse functional data

G. Vinué$^{(1)\boxtimes}$, I. Epifanio$^{(2)}$

(1) Department of Statistics and O.R., University of Valencia, 46100 Burjassot, Spain.
(2) Dept. Matemàtiques and Institut de Matemàtiques i Aplicacions de Castelló.
Campus del Riu Sec. Universitat Jaume I, 12071 Castelló, Spain.

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: Forecasting, Functional data analysis, Archetypal analysis, Functional sparse data, Basketball

1. Introduction

Basketball analytics started to attract attention with the publications by [24] and [17]. More recently, other papers and books have been released [22, 34, 32]. Technological advances have made it possible to collect more data than ever about what is happening on the field, requiring new methods

Email address: guillermo.vinue@uv.es, epifanio@mat.uji.es (G. Vinué$^{(1)\boxtimes}$, I. Epifanio$^{(2)}$)

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. One of the main challenges in basketball analytics is to use past performance to predict future performance [32]. To address this open question, some forecasting methods have been developed. Following [19, 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: College Prospect Rating (CPR) is a score assigned to college basketball players that attempts to estimate their NBA potential [32, 33]. A methodology with a similar design to ours is the Career-Arc Regression Model Estimator with Local Optimization (CARMELO) method. For a player of interest, CARMELO identifies similar players throughout NBA history and uses their careers to forecast the future player's activity [35].

Regarding cross-sectional models, a Weibull-Gamma with covariates timing model is proposed in [18] 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 [30], 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 [37, 42]. A comparison between predictions based on NCAAB and NBA match data is discussed in [46]. A dynamic paired comparison model is described in [3] for the results of matches in two basketball and football tournaments. Furthermore, in [4] 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 analyses data drawn from continuous underlying processes, often time, i.e. a whole function is a datum. Let us assume that n smooth functions, $x_1(t)$, ..., $x_n(t)$, are observed, with the i-th function measured at t_{i1},..., t_{in_i} 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 [35], 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 $var_X(t) = (n-1)^{-1}\sum_{i=1}^{n}(x_i(t) - \bar{x}(t))^2$ and the covariance function $cov_X(t_h, t_l) = (n-1)^{-1}\sum_{i=1}^{n}(x_i(t_h) - \bar{x}(t_h))(x_i(t_l) - \bar{x}(t_l))$. An excellent overview of FDA is found in [28], while methodologies for studying functional data nonparametrically are found in [15]. [29] introduce related software and [27] present some interesting applications in different fields. Other recent applications include [9] and [23]. 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 since 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 by [21].

As regards the forecasting of functional time series, there is a body of research, such as [31, 2, 20], where functions are measured over a fine grid of points. However, only a few works deal with the problem of forecasting sparse functional data [11]. 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.

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 [14, 16]. To the best of our knowledge, there are only two references about sports analytics using FDA. In [43], FDA was introduced for the study of players' aging curves and both hypothesis testing and exploratory analysis were performed.

[40] extended archetypoid analysis (ADA) for sparse functional data (see also [41, 13]), showing the potential of FDA in sports analytics. In particular, 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 (averaged, for example).

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: (i) Regularized Optimization for Prediction and Estimation with Sparse data (ROPES), originally developed by Alexander Dokumentov and Rob Hyndman [11, 10]; (ii) Principal components Analysis through Conditional Expectation (PACE), originally developed by Fang Yao, Hans-Georg Müller and Jane-Lin Wang [44].

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 forecasting, 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 [25].

Forecasting future performance is also very relevant to other sports (see for instance [1]. 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** [5]) 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 three analyses: (i) A validation study is carried out to choose an optimal blend of tuning parameters which ROPES depends on; (ii) ROPES and PACE are compared with each other and with standard benchmarks; (iii) The reliability of ROPES predictions for current players using the method of analogues

[1] `https://www.basketball-reference.com/about/bpm.html`

[2] `https://www.basketball-reference.com/about/ws.html`

with FADA is shown. A discussion with CARMELO results is also provided. The paper ends with some conclusions in Section 5. An appendix shows how this methodology can also be proposed for forecasting international players.

2. Methodology

2.1. ROPES

The method ROPES (Regularized Optimization for Prediction and Estimation with Sparse data), proposed by Alexander Dokumentov and Rob Hyndman [11, 10], 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:

$$\{(U,V)\} = \underset{U,V}{argmin}\Big(||W \odot (Y - UV^T)||^2 + ||U||^2 + ||\mathrm{DIFF}_2(m,\lambda_2)V||^2 + ||\mathrm{DIFF}_1(m,\lambda_1)V||^2 + ||\mathrm{DIFF}_0(m,\lambda_0)V||^2\Big) \tag{1}$$

where:

- Y is an $n \times m$ matrix.
- U is an $n \times k$ matrix of “scores” (“coefficients”), $k = min(n,m)$.
- V is a $m \times k$ matrix of “features” (“shapes”).
- $||.||$ is the Frobenius norm.
- $\odot$ is the element-wise matrix multiplication.
- W is an $n \times m$ “masking matrix” of weights.
- λ_0, λ_1 and λ_2 are smoothing parameters.

and where $\mathrm{DIFF}_i(m,\lambda)$ represents the discrete i times derivative operator multiplied by the scalar λ. In particular, $\mathrm{DIFF}_0(m,1)$ is the identity matrix $m \times m$; $\mathrm{DIFF}_1(m,1)$ is the matrix $(m-1) \times m$, with -1 values in the main diagonal, 1 values in the following upper diagonal and 0 otherwise;

$\mathrm{DIFF}_2(m,1)$ is the matrix $(m-2)\times m$, with 1 values in the main diagonal, -2 values in the following upper diagonal, 1 values in the following upper diagonal, and 0 otherwise.

ROPES is equivalent to maximum likelihood estimation with partially observed data, which allows the calculation of confidence and prediction intervals. They are estimated using a Monte-Carlo style method. The original two sources [11, 10] should be referred to for all the specific details.

2.2. PACE

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 discretizing the functions or basis by expanding the functions are inefficient when data with only a few repeated and sufficiently irregularly spaced measurements per subject are available. Note that when functions are measured over a fine grid of time points, it is possible to fit a separate function for each case using any reasonable basis. However, in the sparse case, this approach fails and 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 [44]. This method was referred to as Principal components Analysis through Conditional Expectation (PACE) for sparse and irregular longitudinal data. In practice, the prediction for the trajectory $X_i(t)$ for the ith subject, using the first p ϕ_q eigenfunctions, is:

$$\hat{X_i^p}(t) = \hat{\mu} + \sum_{q=1}^{p} \hat{\xi_{iq}} \hat{\phi_q}(t) \tag{2}$$

where $\hat{\mu}$ is the estimate of the mean function $E(X(t)) = \mu(t)$ and ξ_{iq} are the FPC scores. PACE and its implementation in the R library **fdapace** ([7]) 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; for example, 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 4.2. 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. [44] also explain the construction of asymptotic pointwise confidence intervals for individual trajectories and asymptotic simultaneous confidence bands.

2.3. ADA

Archetypoid analysis (ADA) was presented in [41] and is an extension of archetypal analysis defined by [6] (see [8, 26] for other derived methodologies). In ADA, archetypes correspond to real observations (the so-called archetypoids). Let $\mathbf{X}$ 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 $\mathbf{Z}$ that characterizes the archetypal patterns in the data. In ADA, the optimization problem is formulated as follows:

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{g} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{g} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \tag{3}$$

under the constraints

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

Archetypoids are computed with the R package **Anthropometry** [39].

2.3.1. ADA for sparse data with FDA

ADA was defined for functions in [13], 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 functional archetypoid analysis (FADA) is as follows. Based on the Karhunen-Loève expansion, the functions are approximated as in Eq. 2. Because the eigenfunctions are orthonormal, to obtain FADA we can apply ADA to the $n \times p$ matrix $\mathbf{X}$, with the scores (the coefficients in the Karhunen-Loève basis).

3. Data

We have used the R package **ballr** [12] to obtain the total advanced statistics for each NBA player from the 1973-1974 season to the latest season, 2017-2018, including the player's age on February 1st of that season. From the total set of statistics, we have focused on Box Plus/Minus (BPM) and Win Shares (WS).

BPM is a box score-based variable for estimating basketball players' quality and contribution to the team. It takes into account both the players' statistics and the team's overall performance. The final value enables us to evaluate the player's performance relative to the league average. BPM is a per-100-possession statistic and its scale is as follows: 0 is the league average, +5 means that the player has contributed 5 more points than an average player over 100 possessions, -2 is replacement level, and -5 is very bad. Replacement level players are those who replace a roster spot for short-term contracts, so they are not normal members of a team's rotation. We have chosen BPM because it was created to use only the information that is available historically. According to the website where BPM is explained [3], *"it is possible to create a better stat than BPM for measuring players, but difficult to make a better one that can also be used historically"*, which fits perfectly with the goal of our method. BPM is available from the 1973-1974 season.

We have chosen a second metric, which is also widely used: Win Shares (WS). It also has the advantage of taking the surrounding team into account. In particular, WS is a player statistic that distributes the team's success among the team players. It is calculated using player, team and league statistics. The sum of all the players' WS in a given team will be approximately equal to that team's total wins for the season. A player with negative WS means that the player took away wins that the teammates had generated.

[3] https://www.basketball-reference.com/about/bpm.html

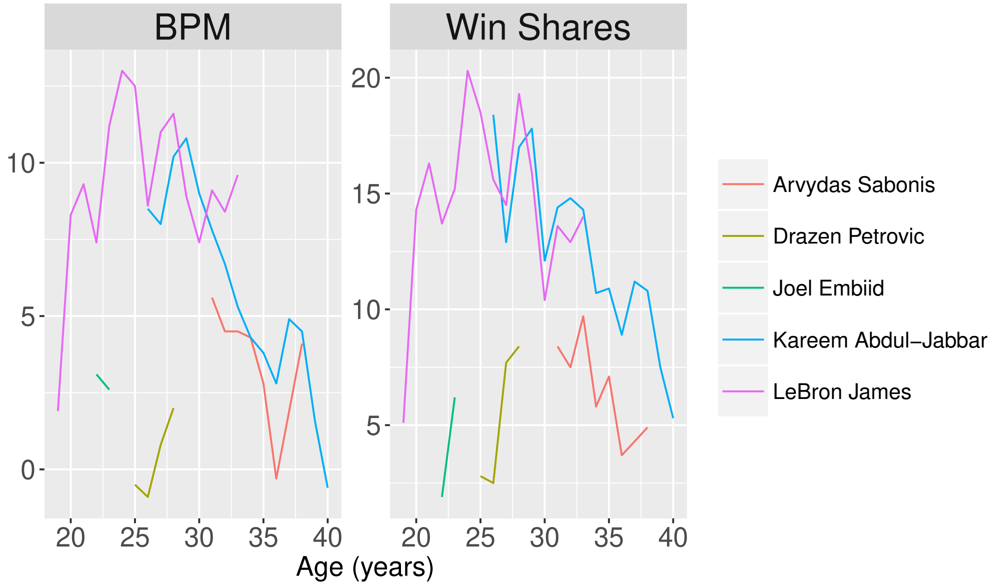

Figure 1: BPM and Win Shares values for some players present in the database, at their corresponding ages (colors in the online version).

The reason for analyzing two variables is to investigate differences in career arcs for different aspects of skill. This allows us to highlight the power of our approach and could be of interest to basketball fans/analysts. Any other statistic can be chosen.

We have removed the observations with fewer than five games played. They were related to very extreme BPM values, such as -86.7 for Gheorghe Muresan in 1998-1999[4] or -49.3 for Mindaugas Kuzminskas in 2017-2018[5].

Fig. 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,

[4] https://www.basketball-reference.com/players/m/muresgh01.html#all_advanced

[5] https://www.basketball-reference.com/players/k/kuzmimi01.html#all_advanced

related to a few long-lasting players, so we have removed them. The age range finally considered is from 18 to 40 years old, i.e., there are 23 time points. In total there are 3075 players.

4. Results

In Section 4.1 a validation study is proposed to choose an appropriate combination of ROPES tuning parameters, which is crucial for good predictive activity. Section 4.2 contains a comparison analysis between ROPES, PACE and two benchmark methods. Section 4.3 specifies the type of projections obtained.

4.1. Selection of parameters

ROPES depends on three tuning parameters (λ_2, λ_1, λ_0), which have to be chosen to guarantee that the model itself returns predictions with enough accuracy. We evaluate the precision of the model's prediction in terms of the mean squared error (MSE). MSE measures the average of the squares of the differences between the predicted values $\hat{y}$ and the true values y across all individual estimates i, as shown in Eq. (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{4}$$

We adopt MSE since ROPES uses it to measure the error term. In order to select the parameters, we proceed as follows: our goal will be to predict the BPM in the 2017-2018 season, for the players who played at least in one season before the 2017-2018 season and who also played in the 2017-2018 season itself. The justification for doing this is related to sporting reasons. In sports, when coaches and managers are building their rosters, it is highly important for them to have a basic idea about how players will perform during the following season. Of course, they would also like to know the players' performance in the long term, but most rosters are built according to the most immediate season. This would allow them to decide whether the current roster should remain the same for the next season or whether some players should be replaced. This procedure makes sense because we will consider the previous performance of all the players selected, but we are only interested in predicting their BPM for the next season, by taking into account each player's data and the information about the other players. This procedure is more computationally efficient than the leave-one-out approach.

From the 3075 players, there are 385 who played in the 2017-2018 season and at least another season before that. Firstly, we split our data into a training+validation set ($TrVaSet$) with 2690 ($3075-385$) players and a test set with the 385 aforementioned players. No test set player belongs to $TrVaSet$. We will select the optimal combination of λ's with $TrVaSet$. To do this, we use a 10 fold cross-validation procedure. Inside each fold, we use 60% of the data for training and then calculate the performance using the 40% remaining validation data. For the validation players, their BPM value is replaced by NA in the Y matrix. In the W masking matrix the 1 value is then replaced by 0.

The first step is to optimize the parameter λ_2, setting $\lambda_1 = 0$ and $\lambda_0 = 10$. The parameter λ_2 takes values in a sequence from 0 to 1000 in increments of 100. In this way, the first blend is $(\lambda_2 = 0, \lambda_1 = 0, \lambda_0 = 10)$, the second is $(\lambda_2 = 100, \lambda_1 = 0, \lambda_0 = 10)$ and so on. We are looking for smooth curves, so we place more emphasis on λ_2 because it is related to the second derivative and this derivative is strongly related to the smoothness of the curve. This is justified because if the second derivative is a smooth curve, both the first derivative and the original function will also be smooth. From the definition of derivative, it directly follows that if a function has a first derivative at any point, then it does not have a sharp bend (v-shape) at that point (the same can be said of the second derivative with respect to the first derivative). See for example [28, Section 5.2.2] for further insights. The opposite is not always true.

Fig. 2 shows the averaged MSE across folds for every combination when only λ_2 is moving. The smallest MSE was for the combination with $\lambda_2 = 900$.

Once the optimal λ_2 has been found, we then adjust λ_1 and λ_0 as well. Both λ_0 and λ_1 take values in a sequence from 0 to 10 in increments of 5. Fig. 3 shows the averaged MSE across folds for every combination when both λ_0 and λ_1 are moving. The smallest MSE was for the combination ($\lambda_2 = 900$, $\lambda_1 = 10$, $\lambda_0 = 10$).

The grid search procedure is chosen since it is a traditional way of performing hyperparameter optimization and can return results in a reasonable amount of time. This second point was particularly important because ROPES is computationally expensive.

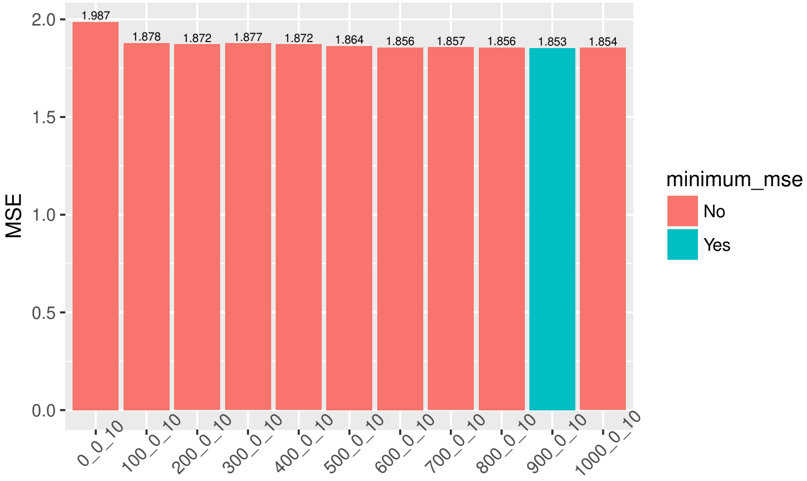

Figure 2: Averaged MSE across folds for every combination of lambdas when λ_2 is moving and λ_0, λ_1 are fixed. The combination for the smallest MSE is highlighted in green (colors in the online version).

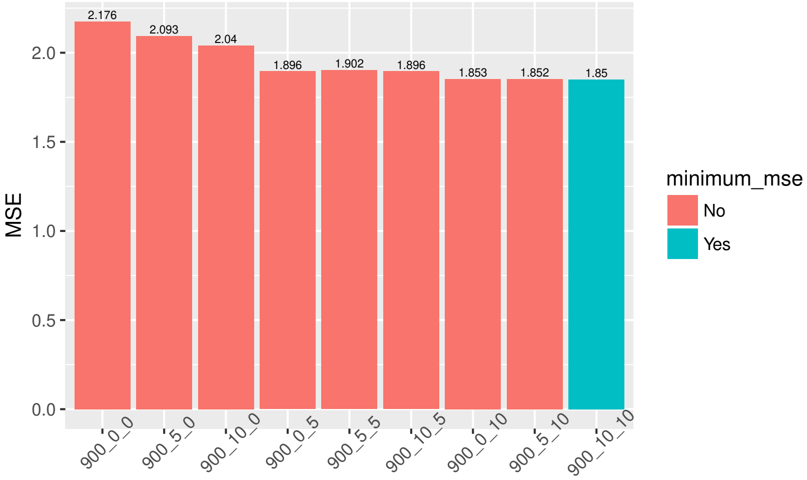

Figure 3: Averaged MSE across folds for every combination of lambdas when λ_2 is fixed and λ_0, λ_1 are moving. The combination for the smallest MSE is highlighted in green (colors in the online version).

Table 1: Actual and predicted BPM values for the 2017-2018 season with ROPES, PACE, the average method and the naïve method, for the test set of players who played during the 2017-2018 season and at least one season before. The difference between actual and predicted values and MSE are also provided. MSE is highlighted in bold. Extract of the results.

			ROPES (900,10,10)		PACE		Average		Naïve	
Player	Age	BPM	BPM_{pr}	Dif.^{2}	BPM_{pr}	Dif.^{2}	BPM_{pr}	Dif.^{2}	BPM_{pr}	Dif.^{2}
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)	

4.2. 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 [19, Section 2.3].

In order to check the performance of all the methods, we have applied them to the test set of 385 players. Table 1 reports an extract of the results. It contains the following information for all players in the 2017-2018 season: (i) their age; (ii) their actual BPM value; (iii) the predictions with ROPES (using the optimal λ combination), PACE and the simple methods; (iv) the squared difference between actual values and predictions (denoted as Dif.^{2}); (v) 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 may indicate that in theory the player's performance in the next year should not be far from the previous one. However, this is not always the case in practice.

Fig. 4 displays the boxplots for the actual BPM values together with the BPM predictions for each method in different intervals.

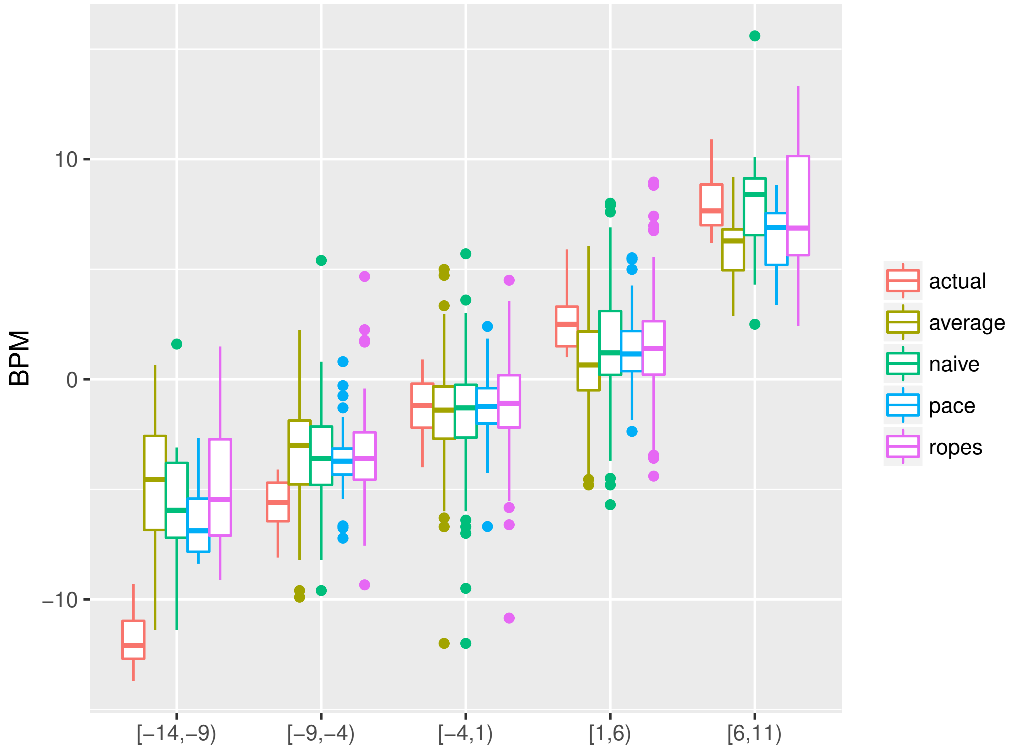

Figure 4: Boxplots for the actual BPM values together with the BPM predictions for each method in different intervals.

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 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.

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.

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

4.3.1. Case study: Joel Embiid

Joel Embiid is a Cameroonian player for the Philadelphia 76ers. He was selected with the third overall pick in the 2014 NBA draft and made his debut in the NBA in the 2016-2017 season after two years of injuries and problems. During his rookie season, Embiid was selected for the NBA All-Rookie First Team, even though he only played 31 games. In 2017-18, he was named an All-Star and was member of both the All-NBA Second Team and the NBA All-Defensive Second Team. The widespread view is that he will be a super star player on both ends of the court for years to come. Hence, it is interesting to see his forecasting.

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 4.2 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 [38]. Recently, an R package has been released that contains analogue methods for palaeoecology [36]. The CARMELO method is also based on this scheme.

In order to find related players, we use archetypoid analysis (see [41] for theoretical details). In this technique, the BPM (and WS) function of a player is approximated by a mixture of archetypoids, which are themselves functions of extreme players. Archetypoids are specific players and the α 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 α coefficients for the player of interest.

We choose the number of archetypoids for each metric following the screeplot explained by [41]. 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 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 LeBron 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 lead 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.

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, LeBron is the representative of super star players.

Otis Birdsong and Ben Wallace represent very good players. Otis Birdsong played twelve NBA seasons and appeared in four 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 α values for Embiid for the BPM and WS archetypoids.

In order to select the group of analogous players, we choose the archetypoids with the highest α. Embiid's greatest similarity for BPM is with Le-

Table 2: Similarity of Joel Embiid to the BPM and WS archetypoids according to the α coefficients.

	BPM Archetypoids				
Player	D. Gray	D. Dawkins	D. Stone	E. Curry	L. James
J. Embiid	0.23	0.15	0	0.13	0.49
	WS Archetypoids				
Player	S. Burtt	B. Wallace	O. Birdsong	L. James	
J. Embiid	0.42	0.18	0.29	0.11	

Bron and for WS is with Burtt. Then, the group of BPM analogous players to Embiid is made up of LeBron James, together with the other players whose largest α coefficient is also for LeBron and who have an α value greater than Embiid's α. 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. Fig. 5 shows the forecasting obtained for Embiid. Regarding BPM, it shows that Embiid 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. The WS prediction stops at age 33 because this is the last age for which the set of analogous players shows values. In general, the curves are smooth and the intervals are wide. These prediction intervals are very useful to assess the uncertainty of forecasts.

4.3.2. Discussion with respect to CARMELO and the web application

CARMELO is a basketball forecasting system released in the 2015-2016 season. Successive versions present some improvements [35]. 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

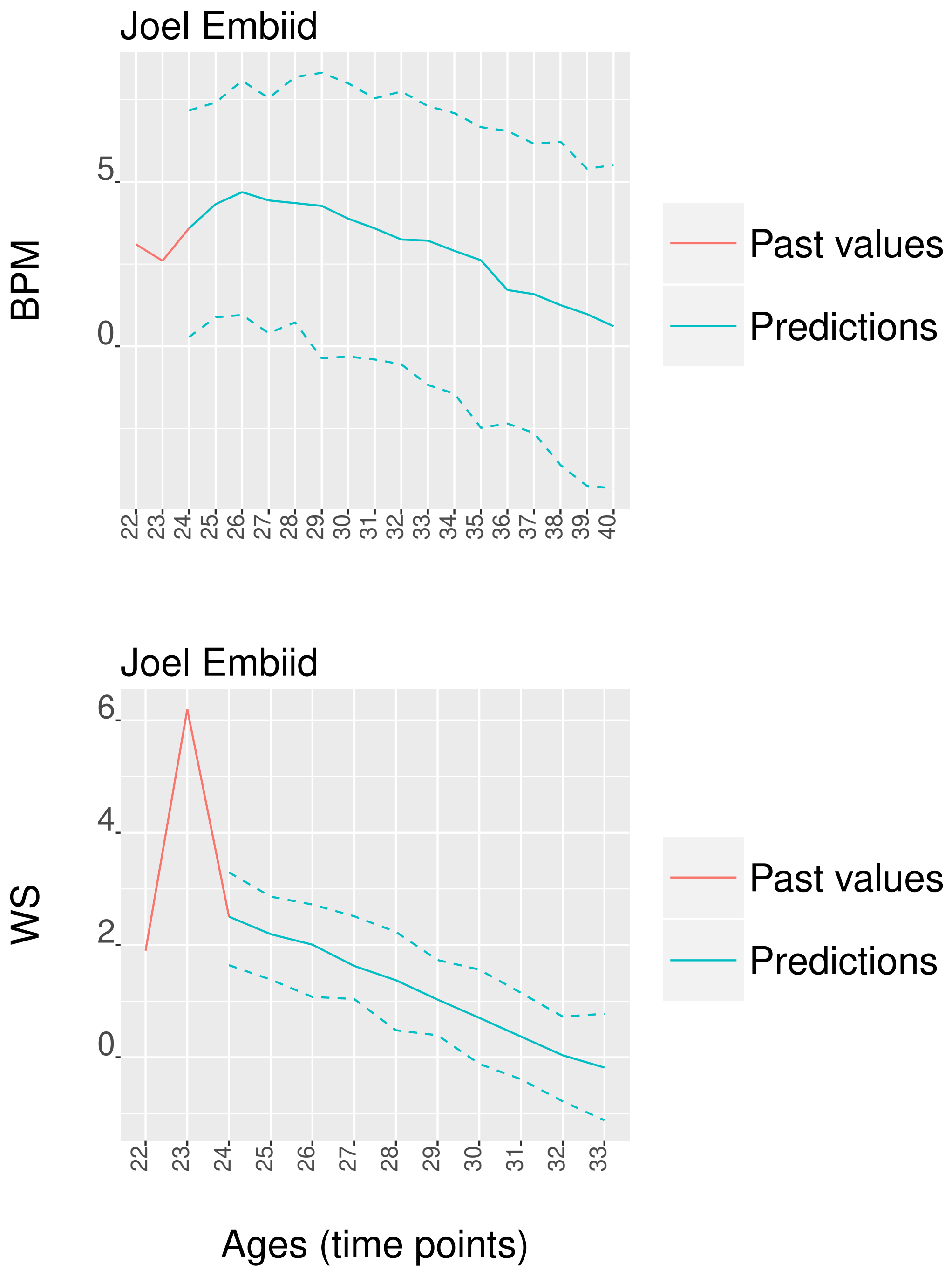

Figure 5: BPM and WS predictions for Joel Embiid using only the set of analogue players (colors in the online version).

blend of two-thirds Real Plus-Minus (RPM) and one-third Box Plus/Minus (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 [6]. In fact, the CARMELO methodology cannot be replicated either. In addition, for seasons before 2000-01, no RPM is available and CARMELO uses BPM only.

We have checked our BPM prediction for Embiid with the one that the CARMELO 2018-2019 version provides[7]. 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, since the target variable is not exactly the same. However, both approaches should be complementary. We see that CARMELO also indicates that Embiid's performance will increase within two seasons and then his values will decrease. This is in line with our forecast.

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')
```

As a final point, it is important to remember that statistical models are not completely reliable for long-term forecasting, 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.

[6] https://www.boxscoregeeks.com/articles/rpm-and-a-problem-with-advanced-stats

[7] https://projects.fivethirtyeight.com/carmelo/joel-embiid/

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 continues 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. 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. In order 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.

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. Two complicating factors that our analysis is currently not considering are as follows: (i) heteroscedasticity (unequal variances) caused by different amounts of playing time going into each averaged BPM and WS data points; (ii) the pattern of sparsity in the data is not random, since players retiring or leaving the NBA should indicate that their BPM and WS would be low in these intervals. We will consider these matters in future work. The data and all R code are freely available for reproducibility and further exploration of the results.

6. Acknowledgements

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. Guillermo Vinué worked on the first revision of the manuscript as part of KU Leuven, Department of Computer Science, Belgium. Funding: 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.

7. Data Accessibility

The authors are making the data associated with this paper available at `https://www.uv.es/vivigui/software`.

References

[1] Arndt, C., Brefeld, U., 2016. Predicting the future performance of soccer players. Statistical Analysis and Data Mining: The ASA Data Science Journal 9, 373–382, `http://dx.doi.org/10.1002/sam.11321`.

[2] Aue, A., Dubart Norinho, D., Hörmann, S., 2015. On the Prediction of Stationary Functional Time Series. Journal of the American Statistical Association 110 (509), 378–392, http://dx.doi.org/10.1080/01621459.2014.909317.

[3] Cattelan, M., Varin, C., Firth, D., 2013. Dynamic Bradley-Terry modelling of sports tournaments. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62 (1), 135–150, http://dx.doi.org/10.1111/j.1467-9876.2012.01046.x.

[4] Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K., 2016. A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes. Journal of the American Statistical Association 111 (514), 585–599, http://dx.doi.org/10.1080/01621459.2016.1141685.

[5] Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2015. shiny: Web Application Framework for R. R package version 0.12.2.
https://CRAN.R-project.org/package=shiny.

[6] Cutler, A., Breiman, L., 1994. Archetypal Analysis. Technometrics 36 (4), 338–347, http://dx.doi.org/10.2307/1269949.

[7] Dai, X., Hadjipantelis, P., Ji, H., Mueller, H.-G., Wang, J.-L., 2016. fdapace: Functional Data Analysis and Empirical Dynamics. R package version 0.2.5.
https://CRAN.R-project.org/package=fdapace.

[8] D'Esposito, M. R., Palumbo, F., Ragozini, G., 2012. Interval Archetypes: A New Tool for Interval Data Analysis. Statistical Analysis and Data Mining 5 (4), 322–335, http://dx.doi.org/10.1002/sam.11140.

[9] Di Battista, T., Fortuna, F., 2017, Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). Statistical Analysis and Data Mining: The ASA Data Science Journal 10 (1), 21–28, https://doi.org/10.1002/sam.11334.

[10] Dokumentov, A., 2016. 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, http://arrow.monash.edu.au/vital/access/manager/Repository/monash:165926.

[11] Dokumentov, A., Hyndman, R. J., 2016. Low-dimensional decomposition, smoothing and forecasting of sparse functional data, `http://robjhyndman.com/papers/ROPES.pdf`. Working paper, 1-31.

[12] Elmore, R., 2018. `ballr`: Access to Current and Historical Basketball Data. R package version 0.1.1.
`https://CRAN.R-project.org/package=ballr`.

[13] Epifanio, I., 2016. Functional archetype and archetypoid analysis. Computational Statistics & Data Analysis 104, 24–34, `http://dx.doi.org/10.1016/j.csda.2016.06.007`.

[14] Epifanio, I., Ávila, C., Page, Á., Atienza, C., 2008. Analysis of multiple waveforms by means of functional principal component analysis: normal versus pathological patterns in sit-to-stand movement. Medical & Biological Engineering & Computing 46 (6), 551–561, `http://dx.doi.org/10.1007/s11517-008-0339-6`.

[15] Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer.

[16] Harrison, A. J., 2014. Applications of functional data analysis in sports biomechanics. In: 32 International Conference of Biomechanics in Sports. pp. 1–9.

[17] Hollinger, J., 2005. Pro basketball forecast. Potomac Books, Inc., Washington, D.C.

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

[19] Hyndman, R. J., Athanasopoulos, G., 2013. Forecasting: Principles and Practice. OTexts, `https://www.otexts.org/book/fpp`.

[20] Hyndman, R. J., Shahid Ullah, M., 2007. Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis 51 (10), 4942–4956, `http://dx.doi.org/10.1016/j.csda.2006.07.028`.

[21] James, G., 2010. The Oxford handbook of functional data analysis. Oxford University Press, Ch. Sparseness and functional data analysis, pp. 298–326.

[22] Kubatko, J., Oliver, D., Pelton, K., Rosenbaum, D. T., 2007. A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports 3 (3), 1–10, http://dx.doi.org/10.2202/1559-0410.1070.

[23] Nguyen, H. D., Ullmann, J. F. P., McLachlan, G. J., Voleti, V., Li, W., Hillman, E. M. C., Reutens, D. C., Janke, A. L., 2018. Whole-volume clustering of time series data from zebrafish brain calcium images via mixture modeling. Statistical Analysis and Data Mining: The ASA Data Science Journal 11 (1), 5–16, https://doi.org/10.1002/sam.11366.

[24] Oliver, D., 2004. Basketball on paper: Rules and tools for performance analysis. Potomac Books, Inc., Washington, D.C.

[25] R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[26] Ragozini, G., Palumbo, F., D'Esposito, M. R., 2017. Archetypal analysis for data-driven prototype identification. Statistical Analysis and Data Mining: The ASA Data Science Journal 10 (1), 6–20, http://dx.doi.org/10.1002/sam.11325.

[27] Ramsay, J. O., Silverman, B., 2002. Applied Functional Data Analysis. Methods and Case Studies. Springer.

[28] Ramsay, J. O., Silverman, B., 2005. Functional Data Analysis, 2nd Edition. Springer.

[29] Ramsay, J. O., Hooker, G., Graves, S. 2009. Functional Data Analysis with R and MATLAB. Springer.

[30] Salador, K., 2011. Forecasting Performance of International Players in the NBA. In: MIT Sloan Sports Analytics Conference. Boston, MA, USA, pp. 1–18, http://www.sloansportsconference.com/wp-content/uploads/2011/08/Forecasting-Performance-of-International-Players-in-the-NBA.pdf.

[31] Shang, H. L., Hyndman, R. J., 2017. Grouped Functional Time Series Forecasting: An Application to Age-Specific Mortality Rates. Journal of Computational and Graphical Statistics 26 (2), 330–343, http://dx.doi.org/10.1080/10618600.2016.1237877.

[32] Shea, S. M., 2014. Basketball analytics: Spatial tracking. Createspace, Lake St. Louis, MO.

[33] Shea, S. M., Baker, C. E., 2013. http://www.basketballanalyticsbook.com/.

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

[35] Silver, N., 2018. CARMELO NBA player projections. 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/.

[36] Simpson, G., Oksanen, J., 2016. analogue: Analogue and Weighted Averaging Methods for Palaeoecology. R package version 0.17-0. https://CRAN.R-project.org/package=analogue.

[37] Strumbelj, E., Vračar, P., 2012. Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. International Journal of Forecasting 28 (2), 532–542, http://dx.doi.org/10.1016/j.ijforecast.2011.01.004.

[38] Viboud, C., Boelle, P.-Y., Carrat, F., Valleron, A.-J., Flahault, A., 2003. Prediction of the Spread of Influenza Epidemics by the Method of Analogues. American Journal of Epidemiology 158 (10), 996–1006, https://doi.org/10.1093/aje/kwg239.

[39] Vinué, G., 2017. Anthropometry: An R Package for Analysis of Anthropometric Data. Journal of Statistical Software 77 (6), 1–39, https://doi.org/10.18637/jss.v077.i06.

[40] Vinué, G., Epifanio, I., 2017. Archetypoid analysis for sports analytics. Data Mining and Knowledge Discovery, 1–35, https://doi.org/10.1007/s10618-017-0514-1.

[41] Vinué, G., Epifanio, I., Alemany, S., 2015. Archetypoids: A new approach to define representative archetypal data. Computational Statistics and Data Analysis 87, 102–115, http://dx.doi.org/10.1016/j.csda.2015.01.018.

[42] Vračar, P., Strumbelj, E., Kononenko, I., 2016. Modeling basketball play-by-play data. Expert Systems with Applications 44, 58–66, http://dx.doi.org/10.1016/j.eswa.2015.09.004.

[43] Wakim, A., Jin, J., 2014. Functional Data Analysis of Aging Curves in Sports, http://arxiv.org/abs/1403.7548, 1-25.

[44] Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional Data Analysis for Sparse Longitudinal Data. Journal of the American Statistical Association 100 (470), 577–590, http://dx.doi.org/10.1198/016214504000001745.

[45] Zorita, E., Von Storch, H., 1999. The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods. Journal of Climate 12, 2474–2489, http://dx.doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.

[46] Zimmermann, A., 2016. Basketball predictions in the NCAAB and NBA: Similarities and differences. Statistical Analysis and Data Mining: The ASA Data Science Journal 9, 350–364, http://dx.doi.org/10.1002/sam.11319.