Rodolfo Metulini*, Marica Manisera and Paola Zuccolotto

Modelling the dynamic pattern of surface area in basketball and its effects on team performance

https://doi.org/10.1515/jqas-2018-0041

Abstract: Because of the advent of GPS techniques, a wide range of scientific literature on Sport Science is nowadays devoted to the analysis of players' movement in relation to team performance in the context of big data analytics. A specific research question regards whether certain patterns of space among players affect team performance, from both an offensive and a defensive perspective. Using a time series of basketball players' coordinates, we focus on the dynamics of the surface area of the five players on the court with a two-fold purpose: (i) to give tools allowing a detailed description and analysis of a game with respect to surface areas dynamics and (ii) to investigate its influence on the points made by both the team and the opponent. We propose a three-step procedure integrating different statistical modelling approaches. Specifically, we first employ a Markov Switching Model (MSM) to detect structural changes in the surface area. Then, we perform descriptive analyses in order to highlight associations between regimes and relevant game variables. Finally, we assess the relation between the regime probabilities and the scored points by means of Vector Auto Regressive (VAR) models. We carry out the proposed procedure using real data and, in the analyzed case studies, we find that structural changes are strongly associated to offensive and defensive game phases and that there is some association between the surface area dynamics and the points scored by the team and the opponent.

Keywords: convex hulls; Markov Switching models; sensor data; team sports; Vector Auto Regressive models.

1 Introduction

Research in team sports science has gained relevance in the last few years, also because of the advent of new technologies and big data. Several methods borrowed from machine learning, network and complex systems, geographic information system (GIS), computer vision and statistics have been proposed to solve many different research questions meeting requirements of experts, coaches and analysts. Massive literature on this topic has been well summarized by some review papers (among others, Gudmundsson and Horton, 2016 and Stein et al., 2017).

One important line of research deals with players' spatial dynamics analysis, which is possible by means of data coming from video analysis technologies and global positioning systems (GPS). Teammates highly interact each other and this continuous interaction takes on the features of a complex system. Network theory has been first proposed by Wasserman and Faust (1994) in the context of team sports, to analyse the network of ball passings. Passos et al. (2011) used centrality measures with the aim of identifying the key interactions and the cooperation between team members in water polo. Besides network theory, a promising niche of literature intertwined with psychology, called ecological dynamics, expresses players on the court as agents who face external factors (Turvey and Shaw 1995; Araújo, Davids, and Hristovski 2006; Araújo et al. 2009; Travassos et al. 2012; Duarte et al. 2013; Araújo and Davids 2016). Multivariate data analysis tools can also be used (see for example Metulini, Manisera, and Zuccolotto 2017a; Metulini, Manisera, and Zuccolotto 2017b), where patterns of movements in basketball are identified by means of an integration of multidimensional scaling and cluster analysis). The complex system of the interactions among players is also studied by Richardson et al. (2012) using cluster phase analysis, and by Greihaine, Godbout, and Zerai (2011) from a psychological perspective.

The positioning of each single player in relation to teammates (and/or opponents) can play a crucial role in determining team performance. Coaching experience suggests that the team in possession of the ball should increase distance between players, while the opponent should defend by reducing distance between players (Araújo and Davids 2016). The overall spatial scattering of players on the court can be measured by several quantities (Araújo and Davids 2016): for example, the stretch index (or radius), the team spread and the effective playing space, known in the sport literature with the term

***Corresponding author: Rodolfo Metulini,** University of Brescia, Big & Open Data Innovation (BODaI) Laboratory, C.da S. Chiara, 50, Brescia IT 25122, Italy, e-mail: rodolfo.metulini@unibs.it
Marica Manisera and Paola Zuccolotto: University of Brescia, Big & Open Data Innovation (BODaI) Laboratory, C.da S. Chiara, 50, Brescia IT 25122, Italy

surface area, whose influence on team performance has been investigated by several authors, in the context of different team sports (see, for example, Frencken et al., 2011; Moura et al., 2012; Fonseca et al., 2012; Travassos et al., 2012; Goldfarb, 2014). Visual analysis is frequently used to find preliminary evidence of surface area patterns (Therón and Casares 2010; Kowshik, Chang, and Maheswaran 2012; Metulini 2016).

This paper is concerned with basketball. Scientific literature on basketball covers several aspects of research, ranging from simply outlining the main features of a game by means of descriptive statistics (Kubatko et al. 2007) to investigating more complex problems, such as forecasting the outcome of a game or a tournament (West 2008; Loeffelholz, Bednar, and Bauer 2009; Brown et al. 2010; Gupta 2015; Lopez and Matthews 2015; Ruiz and Perez-Cruz 2015; Yuan et al. 2015; Manner 2016), analysing players' performance (Page, Fellingham, and Reese 2007; Cooper, Ruiz, and Sirvent 2009; Piette, Anand, and Zhang 2010; Fearnhead and Taylor 2011; Ozmen 2012; Page, Barney, and McGuire 2013; Deshpande and Jensen 2016), identifying optimal game strategies (Annis 2006) and describing the players' reactions to stressful moments (Crocker and Graham 1995; Zuccolotto, Manisera, and Sandri 2017).

Spatial dynamics analysis can be effectively applied in basketball; the most interesting issue is to determine how players' dynamics and team performance are related each other and if the presence of a specific player or a specific lineup on the court has an effect on performance. Indeed, to measure a player's performance is very different from assessing the impact of the player on the teammates' behavior and, ultimately, on the team's performance. While there are a lot of contributions in the literature about the former point, the latter is still little explored. Lamas et al. (2011), for example, define a set of "space creation dynamics" (states for the offense that produce a rupture of the defense, creating empty spaces for scoring opportunities) and show how they can be used for game analysis. In the discussion the authors point out that some players' profile may arguably influence the increase in the recurrence of some specific dynamic in a team game strategy. Fewell et al. (2012) describe basketball team offensive strategies by their network properties, in order to capture the interactions among individuals and determine whether these interactions could be associated with team advancement in play-offs. Although this is not the basic aim of their paper, the tool they propose is well suited to assess how single players or lineups affect team performance, as it is always possible to characterize the specific network occurring when they are on the court.

Additional investigations can be made by merging video and GPS data to play-by-play datasets. The practical possibility to do that depends of course on the availability of such data, that are usually collected by authorized operators within the activities of the basketball associations. For major competitions (NBA, FIBA) play-by-play data are largely available and characterized by an acceptable quality level. But there are other championships whose play-by-play data are not easily available, or affected by missing data and recording errors.

Limiting the focus to the analysis of players' positions, probabilistic models of players' coordinates have been used to split the match in phases (Perše et al. 2009; Perica, Trninić, and Jelaska 2011) according to the way players move on the court. Following this line of research, Metulini et al. (2017a,b) analyse some specific case studies focusing on surface areas and confirm the coaching knowledge that offensive and defensive phases exhibit different patterns. In this paper we investigate the association between surface area and team performance, starting from the idea of Frencken et al. (2011), who aim to define a collective variable able to measure the dynamics of team sports. In the cited paper, focused on soccer, the authors use the collective variable to identify an overall game pattern and then assume that deviations from these patterns are present in the build-up of goals. They find that surface area and centroid position may provide a sound basis for a collective variable that captures the dynamics of attacking and defending in soccer at team level. In a similar way, we translate this idea into basketball and examine surface area as a candidate for describing some basic game pattern, which remains valid whatever the game strategy. In this paper, according to Passos, Araújo, and Volossovitch (2016), we measure the surface area by the area within the convex hull of the players on the court. The centroid position is used to label the game phase (offense, defense, transition).

To this aim, we propose a structured three-step procedure, with a two-fold purpose: (i) to give tools allowing a detailed description and analysis of a game with respect to surface area dynamics and (ii) to investigate its influence on the points made by both the team and the opponent. In detail, using a time series of surface areas from a whole match as a dependent variable, we firstly fit data with a Markov Switching Model (MSM) in order to find two regimes that account for structural changes in the space among players. Secondly, we characterize each regime in terms of its association to game phases (offense, defense, transition) and to certain game situations (i.e. presence of a given player or a combination of players on the court,

implementation of a given playbook, ...), also defining some graphical tools able to describe the match dynamics. Finally, we resort to a Vector AutoRegressive model (VAR) in order to investigate whether changes in the regimes' probabilities affect the performance, in terms of points made during offense and points made by the opponents when the team is in defense. We check the functioning of the proposed procedure on three real datasets collected during games played by Italian professional basketball teams in a tournament which was held in February 2017.

The paper is structured as follows. Section 2 introduces the basketball case studies, describes the data source and the datasets manipulation, while Section 3 is devoted to the proposed three-step procedure. For each step, once described the method, we immediately show its application to the data introduced in Section 2. The obtained results are discussed in Section 4, while concluding remarks and future research developments are in Section 5.

2 Case study

In this paper we deal with basketball sensor data. Basketball is a sport played by two teams of five players each on a rectangular court (28 m $\times$ 15 m). The match, according to International Basketball Federation (FIBA) rules, lasts 40 min, and is divided in four periods of 10 min each. The objective is to shoot a ball through a hoop 46 cm in diameter and mounted at a height of 3.05 m to backboards at each end of the court.

2.1 Data source

Data refer to players' coordinates collected during three matches played in February 2017 by Italian professional basketball teams, at the Italian Basketball Cup Final Eight. In the following, we will refer to the three case studies with CS1, CS2 and CS3. MYagonism (https://www.myagonism.com/) was in charge to set up a system to record the players' positioning on the court during the games. Each player worn a microchip that, having been connected with machines built around the court, collected the player's position (in pixels of 1 m^2) in both the x-axis and the y-axis, as well as in the z-axis (i.e. how high the player jumps). The positioning of the players has been detected with an average frequency of about 80 Hz. During the match, a total of 10 (for CS1 and CS3) and 11 (for CS2) players rotated on the court. The system recorded series of 4, 733, 124 observations for CS1, 4, 072, 227 for CS2 and 4, 906, 254 for CS3, each one referring to one among positioning, velocity or acceleration in one among x-, y- or z-axis, for a specific player in a specific time instant.

Play-by-play data are not available for this tournament, so, in order to recover at least one game variable, we collected the scores of the match at the end of every minute, by watching the video of the game. When merging these data to sensor data, the problem of time misalignment arises. The procedure we propose is also able to deal with this occurrence, as will be explained later.

2.2 The dataset

For each case study, we start from a data matrix $\mathbf{X}$ where each row corresponds to a sensor record, described by the variables time (in milliseconds), player, team, positioning, velocity and acceleration along the court length (x), the court width (y) and height (z). $\mathbf{X}$ needs several transformations in order to be ready for the analysis.

First, it is important to clarify that data are detected with a non-constant frequency; in addition, data of different players are recorded at different time instants. As a consequence, we let the dataset contain any detected time instant $\tilde{\tau}$, and we attribute the last datum available to players not detected in $\tilde{\tau}$. The times $\tilde{\tau}$ are not regularly spaced, so an adjustment procedure is needed to obtain the time series of observations drawn from regularly spaced processes.

Second, the players' positions are detected also during the moments when the game is off: these time instants have to be removed from the dataset. However, there is no variable labelling time instants when the game is off, so we needed rules to identify moments to be filtered out. We filter the rows of $\mathbf{X}$ in three consecutive steps, according to the following criteria:

- instants in which the game is off due to game intervals, time-outs, pre-game and post-game: delete periods when more than or less than exactly five players exhibit both x- and y-axes coordinates inside the court;
- moments when free throws are being shot: delete periods when at least one player remains inside the circle of the free throw line for at least h_1 consecutive seconds.
- other (fouls, violations, ball out, ...): delete periods in which all the five players' velocity is smaller than h_2 km/h for at least h_3 consecutive seconds.

Thresholds h_1, h_2 and h_3 are fixed according to reasonable criteria, and then tuned by checking the sensitivity

of results to different choices, by means of graphical representations. Some minor adjustments are made on h_2 and h_3 in order to obtain a total of 40 min (2,400,000 ms, subject to a certain margin of tolerance) for the final dataset. We set $h_1 = 10$, $h_2 = 10$ and $h_3 = 2.5$ in CS1, $h_1 = 10$, $h_2 = 9.4$ and $h_3 = 2.5$ in CS2 and $h_1 = 10$, $h_2 = 9$ and $h_3 = 2.5$ in CS3. Further details about this procedure can be found in Metulini (2017).

Finally, for CS1 the filtered matrix $\mathbf{X}_F$ counts for a total of effective 2,415,336 ms (40 min and 15.336 s), and contains 206,332 observations at not-regularly spaced times $\tilde{\tau}$ (average frequency of 85.426 Hz). For CS2, $\mathbf{X}_F$ counts for a total of effective 2,425,346 ms (40 min and 25.346 s), and contains 232,544 observations at not-regularly spaced times $\tilde{\tau}$ (average frequency of 95.881 Hz). For CS3, $\mathbf{X}_F$ counts for a total of effective 2,402,458 ms (40 min and 2.458 s), and contains 201,651 observations at not-regularly spaced times $\tilde{\tau}$ (average frequency of 83.935 Hz). Considering CS1, apart from one player ($p9$), who remained most time on the bench, the effective time (minutes:seconds) on the court of the other nine players ranges from 12:21 (player $p2$) to 32:17 (player $p1$). Only two lineups played for more than 5 min: $p1$, $p3$, $p4$, $p6$ and $p10$ (8:21) and $p2$, $p3$, $p5$, $p6$ and $p8$ (5:09).

With regard to CS2, players $p1$, $p2$, $p5$ and $p6$ remained on the court for a large amount of time (respectively, 37:44, 27:14, 34:55 and 30:16 minutes:seconds). Totally, seven players remained in the court for at least 18 min, while other four players remained most of the time on the bench. Only two lineups played for more than 5 min: $p1$, $p2$, $p4$, $p5$ and $p6$ (8:59) and $p1$, $p2$, $p5$, $p6$ and $p8$ (5:16).

Finally, in CS3 nine players remained on the court for a considerable amount of time, ranging from 12:57 (player $p3$) to 29:49 (player $p5$). Only one lineup played for more than 5 min: $p2$, $p5$, $p6$, $p9$ and $p10$ (6:23).

We also computed new variables useful for our analysis. We first defined whether each time instant corresponds to an offensive, defensive or transition moment, depending on whether the centroid position (average coordinates of the five players on the court) on the x-axis lies on the back court (defense – D), in the front court (offense – O) or in between (transition – Tr, in detail, in within $[-4,+4]$ meters from the half court).[1] This generates the categorical variable $P_{\tilde{\tau}}$, denoting the game phase, with categories $p = O, D, Tr$ (offense, defense and transition, respectively).

The instant classified as transition moments ($P_{\tilde{\tau}} = Tr$) were 441,068 effective milliseconds (7 min and 21 s) in CS1, 433,005 (7 min and 10 s) in CS2 and 403,334 (6 min and 44 s) in CS3. Offensive phases ($P_{\tilde{\tau}} = O$) were detected for 1,019,255 ms (16 min and 56 s) in CS1, 962,883 (15 min and 52 s) in CS2 and 951,171 (15 min and 48 s) in CS3. Lastly, defensive phases ($P_{\tilde{\tau}} = D$) accounted for 955,013 ms (15 min and 54 s) in CS1, 1,029,458 (16 min and 58 s) in CS2 and 1,047,953 (17 min and 28 s) in CS3.

Finally, we regularized the space between consecutive observations by selecting a row every 100 effective milliseconds. We denote with $\tilde{\mathbf{X}}_F$ the final data matrix, completed with the new variables and composed of the observations recorded at regularly spaced times $\tilde{t} = 1, 2, \cdots, \tilde{T}$, with a constant frequency of 10 Hz.

3 Methods and analysis

Let $\mathcal{C}_{\tilde{t}}$ be the convex hull of the five players on the court at time $\tilde{t}$, and $A_{\tilde{t}}$ the corresponding area, measuring what in the sport literature is usually called surface area, as explained in the Introduction section. Basic summary statistics give a rough support to the idea (common in the coaching experience) that the surface area switches from narrow to large when moving from defense to offense phases [e.g. the median areas (in m^2) of $A_{\tilde{t}}|P_{\tilde{t}} = D$ and $A_{\tilde{t}}|P_{\tilde{t}} = O$ are respectively 22.6 and 52.3 in CS1, 25.8 and 44.7 in CS2, 24.4 and 44.2 in CS3].

We propose a three-step procedure to analyse the dynamic of $\{A_{\tilde{t}}\}_{\tilde{t}=1,2,\cdots,\tilde{T}}$ and assess its influence on the points made by the team and the opponent.

Step 1: we assume a regime-switching model for the process $\{A_{\tilde{t}}\}$, in order to detect structural changes in the surface areas;

Step 2: for offense and defense separately, we analyse the identified regimes dynamics and we relate them to other game variables such as, for example, the presence of a specific player on the court or the whole lineup, the implementation of a specific playbook, etc., by means of contingency tables and visual tools;

Step 3: we assume a multivariate model able to assess the relation between the regimes and the points scored in offense and defense phases.

In the next subsections, for each step we will describe in detail the proposed procedure and the results obtained with the analysed data.

1 We obviously considered that teams change court side after the half-time interval.

3.1 Step 1: regime-switching surface areas

3.1.1 Method

The first step assumes that the surface areas' stochastic process is affected by recurrent structural changes involving its mean. In order to investigate this issue, we borrow models and terminology from the fairly different context of econometrics, where time series with structural changes are often analysed with regime-switching models. In that context, regimes are defined as states involving different parameters for the stochastic process under study, and they are often found to correspond to specific economic situations (e.g. expansion or recession). Translating the idea into our case study, we assume that surface area dynamics are characterized by different regimes involving different mean levels of the process, and we fit a Markov Switching Model (MSM; see Hamilton 2010) to the data coming from the process $\{A_{\tilde{t}}\}$, observed at the equally-spaced times $\tilde{t} = 1, 2, \cdots, \tilde{T}$. The model is able to detect if a regime-switching dynamic is present, estimate the parameters of the process in the different regimes and, for each observation time, the probability of being in one regime or the other. The rationale behind this first step is built upon team sports technical considerations: it is an established fact that the space among players tends to switch from narrow to large when moving from defense to offense phases. Nevertheless, we cannot assume a strict matching between the area $A_{\tilde{t}}$ and the game phase. Quite the opposite, we are just interested to the moments when the surface area is different from what we would expect on the basis of the game phase, because the most intriguing game situations are hidden right there. So, the idea is to separate, in this step, regimes of narrow and large surface areas using a model not considering the game phase and, in step 2, investigate the regimes' dynamics for offense and defense phases separately, highlighting when the regime deviates from what expected.

Let us assume that the surface area switches between two regimes characterized by different mean levels. We call the two regimes N and L, standing for "narrow" and "large", respectively (N is the regime characterized by the lower mean level, and L the other one). Let $R_{\tilde{t}}$ be the (unobserved) random variable denoting the regime at time $\tilde{t}$:

$$E(A_{\tilde{t}}|R_{\tilde{t}} = r) = \alpha^{(r)}, \qquad r = N, L. \tag{1}$$

The probabilistic model describing the regimes' dynamics is assumed to be a two-state Markov chain (Baum et al. 1970; Lindgren 1978; Hamilton 1989),

$$\Pr(R_{\tilde{t}}|R_{\tilde{t}-1}, R_{\tilde{t}-2}, \ldots) = \Pr(R_{\tilde{t}}|R_{\tilde{t}-1}) \tag{2}$$

and we denote with $\pi_{NN} = \Pr(R_{\tilde{t}} = N|R_{\tilde{t}-1} = N)$ and $\pi_{LL} = \Pr(R_{\tilde{t}} = L|R_{\tilde{t}-1} = L)$ the two-state transition probabilities, recalling that $\pi_{NL} = \Pr(R_{\tilde{t}} = N|R_{\tilde{t}-1} = L) = 1 - \pi_{LL}$ and $\pi_{LN} = \Pr(R_{\tilde{t}} = L|R_{\tilde{t}-1} = N) = 1 - \pi_{NN}$. Formulation (1) is perhaps the simplest assumption in the family of regime-switching models, which often assume also the presence of autoregressive components or the effect of some exogenous variables.

After specifying Gaussian densities $\mathcal{N}(\alpha^{(N)}, \sigma_N^2)$ and $\mathcal{N}(\alpha^{(L)}, \sigma_L^2)$ under the two regimes, the parameter vector

$$\boldsymbol{\theta} = (\alpha^{(N)}, \alpha^{(L)}, \sigma_N^2, \sigma_L^2, \pi_{NN}, \pi_{LL})'$$

is estimated via EM algorithm, as the regime is unobserved. The estimation algorithm also returns the so-called "filtered" probabilities

$$\xi_{\tilde{t}|\tilde{t}}^{(r)} = \Pr(R_{\tilde{t}} = r|\mathcal{I}_{\tilde{t}}, \boldsymbol{\theta}) \tag{3}$$

where $\mathcal{I}_{\tilde{t}}$ denotes the information available up to time $\tilde{t}$, and the corresponding "smoothed" probabilities

$$\xi_{\tilde{t}}^{(r)} = \Pr(R_{\tilde{t}} = r|\mathcal{I}, \boldsymbol{\theta}) \tag{4}$$

obtained using all the set of information $\mathcal{I}$ up to time $\tilde{T}$, by means of the algorithm developed by Kim (1994).

If the Markov chain is presumed to be ergodic, we can compute the unconditional probabilities π_r of the two regimes

$$\pi_N = \frac{1 - \pi_{LL}}{2 - \pi_{NN} - \pi_{LL}} \quad \text{and} \quad \pi_L = 1 - \pi_N \tag{5}$$

and their average persistence $\delta^{(r)}$

$$\delta^{(N)} = \frac{1}{1 - \pi_{NN}} \quad \text{and} \quad \delta^{(L)} = \frac{1}{1 - \pi_{LL}}. \tag{6}$$

3.1.2 Data analysis

We fit the MSM of equation (1) to data from the process $\{A_{\tilde{t}}\}$, in order to detect structural changes in the expected value of the surface area.[2] We used the R package MSwM (Sanchez-Espigares and Lopez-Moreno 2014).

2 We preliminarily performed an Augmented Dickey-Fuller (ADF) test for unit roots, in order to guarantee the stationary condition. ADF test reports values equal to −6.88 for CS1, −6.52 for CS2 and −6.39 for CS3, exceeding the threshold −3.43, corresponding to $\alpha = 0.01$.

Table 1: Markov Switching Model results (A), 95% confidence intervals for the estimated intercepts (B) and transition probabilities (C).

	CS1 Coef (S.e.)	CS2 Coef (S.e.)	CS3 Coef (S.e.)
(A) Markov Switching Model results			
Regime N			
Intercept	22.060***	24.448***	21.940***
	(0.114)	(0.123)	(0.112)
Residual standard error	9.156	9.328	8.760
Regime L			
Intercept	62.897***	60.857***	56.114***
	(0.221)	(0.265)	(0.220)
Residual standard error	21.087	20.256	18.133
(B) 95% Confidence intervals for the estimated intercepts			
Regime N	$\left[21.835;\ 22.284\right]$	$\left[24.207;\ 24.689\right]$	$\left[21.721;\ 22.160\right]$
Regime L	$\left[62.465;\ 63.329\right]$	$\left[60.337;\ 61.376\right]$	$\left[55.683;\ 56.545\right]$
(C) Transition probabilities			
	$\begin{bmatrix} 0.986 & 0.013 \\ 0.014 & 0.987 \end{bmatrix}$	$\begin{bmatrix} 0.987 & 0.018 \\ 0.013 & 0.982 \end{bmatrix}$	$\begin{bmatrix} 0.986 & 0.013 \\ 0.014 & 0.987 \end{bmatrix}$

Signif. codes: $\cdot p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Considering CS1, and consistently with the notation used in SubSection 3.1.1, we find that the estimated intercept of regime $r = L$ is larger than that of regime $r = N$ (Table 1, box A), with $\alpha^L = 62.897$ and $\alpha^N = 22.060$. These estimated parameters are highly statistically significant, with p-values in the order of 10^{-16}. Furthermore, the 95% confidence intervals of the two parameters do not overlap (Table 1, box B). The transition probabilities (Table 1, box C) denote low probability of regime switching ($\pi_{NN} = 0.986$ and $\pi_{LL} = 0.987$), with corresponding persistence indexes (6) given by $\delta^{(N)} = 71.43$ and $\delta^{(L)} = 76.92$, meaning that each regime lasts, on average, between 7 and 8 s. The unconditional probabilities (5) result $\pi_N = 0.481$ and $\pi_L = 0.519$. Results for CS2 and CS3, as shown in columns 2 and 3 of Table 1, are very similar to those of CS1.

The filtered and smoothed probabilities $\xi^{(r)}_{\tilde{t}|\tilde{t}}$ and $\xi^{(r)}_{\tilde{t}}$, defined respectively in (3) and (4), are often close to 0 or 1, suggesting a good model fit.

3.2 Step 2: analysis of the regimes' dynamics

3.2.1 Method

Let $P_{\tilde{t}}$ be the categorical variable denoting the game phase, with categories $p = O, D, Tr$ (offense, defense and transition, respectively), described in SubSection 2.2. In step 2 we investigate the connections between the regimes and $P_{\tilde{t}}$ using some descriptive statistics (see SubSection 3.2.2) and a graphical tool that will be described in the following. As mentioned above, we expect a prevalence of regime $r = N$ in defense phases and a corresponding prevalence of $r = L$ in offense phases. This conjecture can be easily verified by means of contingency tables, but we are not interested in this almost obvious result. We aim to identify departures from this evidence by plotting some kernel functions $\Phi^{(r)}_p(t)$ and comparing their pattern to the presence of a specific player or a lineup on the court. These kernel functions are defined as

$$\Phi^{(r)}_p(t) = \underset{\tilde{t}:P_{\tilde{t}}=p}{\phi}\left(\xi^{(r)}_{\tilde{t}}\right), \tag{7}$$

where $\phi(\cdot)$ is a Nadaraya-Watson kernel regression (Nadaraya 1964; Watson 1964), estimated with a Gaussian kernel and a bandwidth allowing to take into account a proper time lag (e.g. the average duration of a game phase), computed using only the times $\tilde{t}$ when the selected game phase p was occurring. By construction, $\Phi^{(r)}_p(t)$ is a function of time t, $t \in \mathbb{R} \cap [0, \tilde{T}]$, so its values can be computed also for time instants when observations are not available. This makes it possible[3] to plot the functions $\Phi^{(r)}_O(t)$ and $\Phi^{(r)}_D(t)$ in the same frame. For a given regime r, the plot of $\Phi^{(r)}_O(t)$ and $\Phi^{(r)}_D(t)$ singles out game

3 In step 3, it will also allows to perform a realignment of times to those of the variables denoting the scored points process ($\bar{t} = 1, 2, .., \bar{T}$, see SubSection 2.2).

tranches with different ways of playing with respect to the surface area. At this stage, a comparison analysis can be performed in order to inspect the relations among regimes and some specific player or a whole lineup, by highlighting, using visual tools, the portions of time when the player or the lineup was on the court.

3.2.2 Data analysis

To study the pattern of the estimated regimes conditionally to the categorical variable $P_{\tilde{t}}$, we first compute some descriptive statistics for the three case studies CS1, CS2 and CS3. We assign each observation at time $\tilde{t}$ to regime r if $\xi_{\tilde{t}}^{(r)} > 0.5$. In doing so, we generate the dichotomous variable $\hat{R}_{\tilde{t}}$ that is the estimate of the latent state $R_{\tilde{t}}$.

In CS1, we have $\hat{R}_{\tilde{t}} = N$ for 52.1% of observations (and, of course, $\hat{R}_{\tilde{t}} = L$ for the remaining 47.9%). In addition, we find that 82.4% of observations in defensive game phases correspond to regime $r = N$, while 77.3% of observations in offensive game phases correspond to regime $r = L$.

In CS2, $\hat{R}_{\tilde{t}} = N$ for 56.8% of observations, 83.0% of observations in defensive game phases correspond to regime $r = N$, while 63.7% of observations in offensive game phases correspond to regime $r = L$.

In CS3, $\hat{R}_{\tilde{t}} = N$ for 54.2% of observations, 79.1% of observations in defensive game phases correspond to regime $r = N$, while 66.2% of observations in offensive game phases correspond to regime $r = L$.

We measure the association between $\hat{R}_{\tilde{t}}$ and $P_{\tilde{t}}$ by means of the normalized association index

$$C = \sqrt{\frac{X^2}{n(k-1)}}$$

where X^2 is the usual Pearson index, n is the number of observations and k the minimum between the number of rows and columns in the bivariate table.

The existence of a fairly strong association ($C =$ 56.22% for CS1, $C = 45.47\%$ for CS2 and $C = 43.95\%$ for CS3) between regimes and game phases is expected from both (coaching) technical considerations and the above cited summary statistics of $A_{\tilde{t}}|P_{\tilde{t}} = D$ and $A_{\tilde{t}}|P_{\tilde{t}} = O$.

More interestingly, we may extend the focus to players. Tables 2–4 report the contingency tables between estimated regimes and players on the court, separately for defensive and offensive phases, for CS1, CS2 and CS3. Tables 2–4 consider only players who played for at least 5 min.

Table 2: Frequency distributions of $\hat{R}_{\tilde{t}}$ conditional to $P_{\tilde{t}}$ and player, for offensive (A) and defensive (B) phases. Case study CS1.

	A Offensive phases		B Defensive phases	
Estimated regime $\hat{R}_{\tilde{t}}$	**Bench**	**Court**	**Bench**	**Court**
Player 1 ($p1$)				
N	0.303	0.207	0.884	0.809
L	0.697	0.793	0.116	0.191
Player 2 ($p2$)				
N	0.194	0.285	0.817	0.833
L	0.806	0.715	0.183	0.167
Player 3 ($p3$)				
N	0.273	0.220	0.799	0.832
L	0.727	0.780	0.201	0.168
Player 4 ($p4$)				
N	0.291	0.190	0.847	0.804
L	0.709	0.810	0.153	0.196
Player 5 ($p5$)				
N	0.195	0.287	0.801	0.852
L	0.805	0.713	0.199	0.148
Player 6 ($p6$)				
N	0.283	0.208	0.844	0.814
L	0.717	0.792	0.155	0.186
Player 7 ($p7$)				
N	0.200	0.292	0.818	0.837
L	0.800	0.708	0.182	0.163
Player 8 ($p8$)				
N	0.199	0.260	0.793	0.847
L	0.801	0.740	0.207	0.153
Player 10 ($p10$)				
N	0.251	0.192	0.819	0.833
L	0.749	0.808	0.181	0.167

As suggested in SubSection 3.2.1, the dependence among regimes and lineups/players on the court can also be inspected by means of a graphical analysis. For the 2 most frequent lineups and the 9 selected players of CS1, Figures 1 and 2 show the functions $\Phi_D^{(L)}(t)$ (blue) and $\Phi_O^{(L)}(t)$ (red), defined in formula (7). Grey areas identify the moments where the lineup or the player was on the court. For the sake of brevity, we do not report the corresponding graphs plotted for CS2 and CS3.

3.3 Step 3: relationship between the regimes and the points scored

3.3.1 Methods

Let $SP_{\bar{t}}^{\text{team}}$ and $SP_{\bar{t}}^{\text{opp}}$ be the points scored $(0, 1, 2, 3, \cdots)$ by the team and the opponent, respectively, in the time interval $(\bar{t} - 1, \bar{t}]$, where, in general, the times $\bar{t} = 1, 2, \cdots, \bar{T}$

Table 3: Frequency distributions of $\hat{R}_{\tilde{t}}$ conditional to $P_{\tilde{t}}$ and player, for offensive (A) and defensive (B) phases.

	A Offensive phases		B Defensive phases	
Estimated regime $\hat{R}_{\tilde{t}}$	**Bench**	**Court**	**Bench**	**Court**
Player 1 ($p1$)				
N	0.438	0.360	0.917	0.824
L	0.562	0.640	0.083	0.176
Player 2 ($p2$)				
N	0.328	0.379	0.854	0.819
L	0.672	0.621	0.146	0.181
Player 4 ($p4$)				
N	0.414	0.312	0.810	0.856
L	0.586	0.688	0.190	0.144
Player 5 ($p5$)				
N	0.283	0.373	0.865	0.825
L	0.717	0.627	0.135	0.175
Player 6 ($p6$)				
N	0.434	0.339	0.804	0.838
L	0.566	0.661	0.196	0.162
Player 7 ($p7$)				
N	0.360	0.383	0.830	0.830
L	0.640	0.617	0.170	0.170
Player 8 ($p8$)				
N	0.313	0.416	0.849	0.813
L	0.687	0.584	0.151	0.187
Player 9 ($p9$)				
N	0.375	0.289	0.822	0.882
L	0.625	0.711	0.178	0.118
Player 10 ($p10$)				
N	0.351	0.378	0.824	0.837
L	0.649	0.622	0.176	0.163

Case study CS2.

Table 4: Frequency distributions of $\hat{R}_{\tilde{t}}$ conditional to $P_{\tilde{t}}$ and player, for offensive (A) and defensive (B) phases.

	A Offensive phases		B Defensive phases	
Estimated regime $\hat{R}_{\tilde{t}}$	**Bench**	**Court**	**Bench**	**Court**
Player 1 ($p1$)				
N	0.378	0.312	0.806	0.782
L	0.622	0.688	0.194	0.218
Player 2 ($p2$)				
N	0.311	0.366	0.820	0.760
L	0.689	0.634	0.180	0.240
Player 3 ($p3$)				
N	0.311	0.389	0.770	0.838
L	0.689	0.611	0.230	0.162
Player 5 ($p5$)				
N	0.406	0.312	0.801	0.788
L	0.594	0.688	0.199	0.212
Player 6 ($p6$)				
N	0.341	0.334	0.779	0.807
L	0.659	0.666	0.221	0.194
Player 7 ($p7$)				
N	0.315	0.386	0.779	0.812
L	0.685	0.614	0.221	0.188
Player 8 ($p8$)				
N	0.379	0.313	0.786	0.793
L	0.621	0.687	0.214	0.207
Player 9 ($p9$)				
N	0.334	0.340	0.825	0.764
L	0.666	0.660	0.175	0.236
Player 10 ($p10$)				
N	0.339	0.337	0.769	0.801
L	0.661	0.663	0.231	0.199

Case study CS3.

are different from the observation times $\tilde{t} = 1, 2, \cdots, \tilde{T}$ of the process $\{A_{\tilde{t}}\}$. In fact, in order to obtain meaningful values for the variables $SP_{\tilde{t}}^{\text{team}}$ and $SP_{\tilde{t}}^{\text{opp}}$, the interval between two consecutive times $\bar{t} - 1$ and $\bar{t}$ has to be in the order of at least 1 min. So, we here need to solve the problem of time misalignment. For a given regime r, we denote with $\mathbf{Y}_{\bar{t}}^{\text{team},r}$ and $\mathbf{Y}_{\bar{t}}^{\text{opp},r}$ the vectors

$$\mathbf{Y}_{\bar{t}}^{\text{team},r} = \begin{bmatrix} SP_{\bar{t}}^{\text{team}} \\ \nabla\Phi_O^{(r)}(\bar{t}) \end{bmatrix} \quad \text{and} \quad \mathbf{Y}_{\bar{t}}^{\text{opp},r} = \begin{bmatrix} SP_{\bar{t}}^{\text{opp}} \\ \nabla\Phi_D^{(r)}(\bar{t}) \end{bmatrix}, \quad (8)$$

where ∇ denotes the first-difference operator, $\nabla\Phi_\star^{(r)}(\bar{t}) = \Phi_\star^{(r)}(\bar{t}) - \Phi_\star^{(r)}(\bar{t}-1)$, necessary to filter out the one-lag dependence induced by construction by the Markov property assumed in step 1 to regulate the regime switching. Note that the realignment of the times related to the regime dynamics ($\tilde{t}$) and those related to the recording of scored

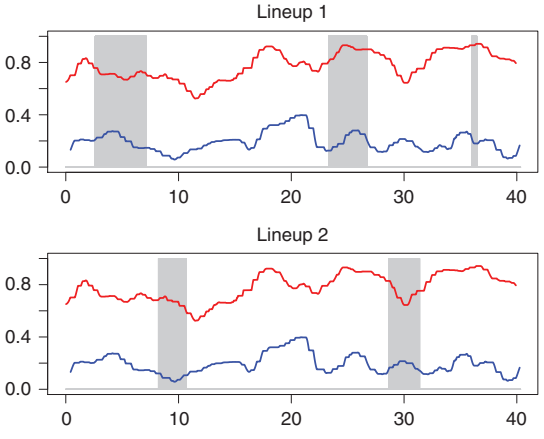

Figure 1: Lineups on the court (grey) against Nadaraya-Watson kernel functions.
The charts display, in y-axis, functions $\Phi_D^{(L)}(t)$ (in blue) and $\Phi_O^{(L)}(t)$ (in red), in x-axis, Time (in min).

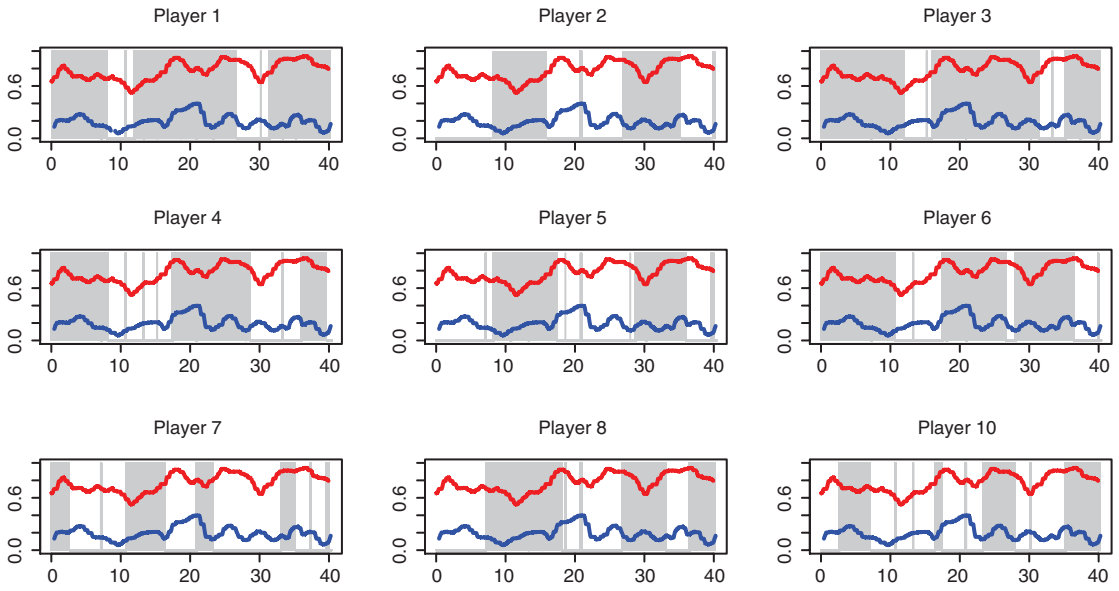

Figure 2: Players on the court (grey) against Nadaraya-Watson kernel functions.
The charts display, in y-axis, functions $\Phi_D^{(t)}(t)$ (in blue) and $\Phi_O^{(t)}(t)$ (in red), in-axis, Time (in min).

points ($\bar{t}$) is solved by using the functions $\Phi_p^{(r)}(t)$ defined in (7), that can be computed for all $t \in \mathbb{R} \cap [0, \tilde{T}]$ and, in this case, are evaluated at the times $\bar{t} = 1, 2, \cdots, \bar{T}$. In order to investigate the relationship between regimes and scored points, we assume a VAR model (Sims 1980) for the processes $\{\mathbf{Y}_{\bar{t}}^{\text{team},r}\}$ and $\{\mathbf{Y}_{\bar{t}}^{\text{opp},r}\}$:

$$\mathbf{Y}_{\bar{t}}^{\text{team},r} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1' \mathbf{Y}_{\bar{t}-1}^{\text{team},r} + \cdots + \boldsymbol{\eta}_q' \mathbf{Y}_{\bar{t}-q}^{\text{team},r} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{team},r} \quad (9)$$

and

$$\mathbf{Y}_{\bar{t}}^{\text{opp},r} = \boldsymbol{\omega}_0 + \boldsymbol{\omega}_1' \mathbf{Y}_{\bar{t}-1}^{\text{opp},r} + \cdots + \boldsymbol{\omega}_s' \mathbf{Y}_{\bar{t}-s}^{\text{opp},r} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{opp},r}, \quad (10)$$

where $\boldsymbol{\eta}_0$ and $\boldsymbol{\omega}_0$ are 2×1 vectors of the intercepts to be estimated, $\boldsymbol{\eta}_1' \ldots \boldsymbol{\eta}_q'$ and $\boldsymbol{\omega}_1' \ldots \boldsymbol{\omega}_q'$ are 2×2 matrices of the coefficients to be estimated, while $\boldsymbol{\varepsilon}_{\bar{t}}^{\text{team},r}$ and $\boldsymbol{\varepsilon}_{\bar{t}}^{\text{opp},r}$ are the usual innovation processes.[4]

3.3.2 Data analysis

As described in SubSection 3.3.1, we estimate the two VAR models of equations (9) and (10).

From here onwards, we will consider time series observed at the end of each minute of the game, i.e. at times $\bar{t} = 1, 2, \cdots, 40$.

To choose the orders of the VAR models, we resort to the Bayesian Information Criterion (BIC), that suggests to set $q = 1$ and $s = 1$ in both CS1 and CS3. For CS2, instead, we should opt for $q = 3$ and $s = 2$. However, the differences between the BIC values corresponding to the two models of order 1 and the optimal ones are less than 0.13 and 0.45, respectively. According to Kass and Raftery (1995), this denotes a negligible evidence against the models with higher BIC values. For this reason, in order to ensure comparability of results, we set $q = 1$ and $s = 1$ also for CS2.

So, we fit models (9) and (10) in the form

$$\mathbf{Y}_{\bar{t}}^{\text{team},L} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1' \mathbf{Y}_{\bar{t}-1}^{\text{team},L} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{team},L} \quad (11)$$

and

$$\mathbf{Y}_{\bar{t}}^{\text{opp},N} = \boldsymbol{\omega}_0 + \boldsymbol{\omega}_1' \mathbf{Y}_{\bar{t}-1}^{\text{opp},N} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{opp},N}, \quad (12)$$

where, as defined in (8), the vectors $\mathbf{Y}_{\bar{t}}^{\text{team},L}$ and $\mathbf{Y}_{\bar{t}}^{\text{opp},N}$ contain the scored points and the first differences of the corresponding functions Φ. From a methodological point of view, this choice is justified by the necessity to filter out the one-lag dependence induced by the Markov property in the MSM, as mentioned before. In practice, this necessity is confirmed by the Augmented Dickey Fuller (ADF) test for unit roots, that results, for CS1, -1.468 and

4 The use of count data (the scored points) within a VAR model could be criticized from a formal point of view. Brandt and Sandler (2012) propose a bayesian Poisson VAR model (see the same reference for a literature review on this topic) to treat count data in a multivariate autoregressive context. Nevertheless, we point out that this is a case of mixed data: count data and numerical variables. In addition, in the theory behind VAR model there are not strict requirements about the nature of the variables to be used.

-0.558 for $\Phi_O^{(L)}(t)$ and $\Phi_D^{(N)}(t)$, respectively (the threshold for $\alpha = 0.05$ is -1.95, so the null hypothesis of the presence of a unit root cannot be rejected). Analogous results hold for CS2 (-1.936 and -1.770 for $\Phi_O^{(L)}(t)$ and $\Phi_D^{(N)}(t)$, respectively) and CS3 (-1.624 and -1.662 for $\Phi_O^{(L)}(t)$ and $\Phi_D^{(N)}(t)$, respectively).

Estimation results of the three case studies are summarized in Table 5 and 6 for models (11) and (12) respectively.

Table 5: VAR model $\mathbf{Y}_{\bar{t}}^{\text{team},L} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1 \mathbf{Y}_{\bar{t}-1}^{\text{team},L} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{team},L}$.

	CS1 Coef (S.e.)	CS2 Coef (S.e.)	CS3 Coef (S.e.)
Results for equation $SP_{\bar{t}}^{\text{team}}$:			
$SP_{\bar{t}-1}^{\text{team}}$	−0.032	−0.281	0.043
	(0.159)	(0.157)	(0.169)
$\nabla\Phi_O^{(L)}(\bar{t}-1)$	7.951	4.323	7.394
	(4.570)	(3.230)	(3.649)
Intercept	2.194***	2.252***	1.649***
	(0.433)	(0.416)	(0.402)
Results for equation $\nabla\Phi_O^{(L)}(\bar{t})$:			
$SP_{\bar{t}-1}^{\text{team}}$	−0.005	0.004	−0.008
	(0.006)	(0.008)	(0.008)
$\nabla\Phi_O^{(r)}(\bar{t}-1)$	0.258	0.235	0.073
	(0.162)	(0.164)	(0.177)
Intercept	0.012	−0.012	0.019
	(0.015)	(0.021)	(0.020)

Signif. codes: $\cdot p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table 6: VAR model $\mathbf{Y}_{\bar{t}}^{\text{opp},N} = \boldsymbol{\omega}_0 + \boldsymbol{\omega}_1 \mathbf{Y}_{\bar{t}-1}^{\text{opp},N} + \boldsymbol{\varepsilon}_{\bar{t}}^{\text{opp},N}$.

	CS1 Coef (S.e.)	CS2 Coef (S.e.)	CS3 Coef (S.e.)
Results for equation $SP_{\bar{t}}^{\text{opp}}$:			
$SP_{\bar{t}-1}^{\text{opp}}$	−0.119	0.054	0.029
	(0.170)	(0.159)	(0.162)
$\nabla\Phi_D^{(N)}(\bar{t}-1)$	1.610	7.934	−3.590
	(4.322)	(5.285)	(3.688)
Intercept	2.436***	1.740***	1.586***
	(0.448)	(0.400)	(0.357)
Results for equation $\nabla\Phi_D^{(N)}(\bar{t})$:			
$SP_{\bar{t}-1}^{\text{opp}}$	−0.016*	−0.012*	−0.011
	(0.006)	(0.005)	(0.007)
$\nabla\Phi_D^{(N)}(\bar{t}-1)$	0.247	0.180	−0.038
	(0.156)	(0.169)	(0.161)
Intercept	0.012	−0.012	0.019
	(0.016)	(0.013)	(0.016)

Signif. codes: $\cdot p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

4 Discussion

The idea that the surface area switches from narrow to large when moving from defense to offense is supported both by team sports coaching experience and statistical evidence. In fact, the median values of the surface areas are found to be appreciably different in defensive and offensive phases in all the analysed case studies (in the range of m^2 22–25 and 44–52, for defense and offense, respectively). The three-step procedure proposed in this paper is designed to detect departures from this basic evidence and to relate these departures to some game variables, such as the players on the court or the points scored by the team and the opponent.

In the first step we described the dynamics of narrow and large surface areas in terms of a stochastic process switching between two regimes according to a latent Markov process. In all the tree case studies, a significant presence of two regimes has been detected, characterized by average surface areas in the range of about m^2 21.7–24.7 and 55.7–63.3 for regime $r = N$ (narrow) and $r = L$ (large), respectively (Table 1). In all the cases, the regimes have a fast switching: they last on average between 7 and 8 s, which suggests that the association of the narrow (large) regime to defense (offense) phases cannot be considered as a perfect matching and stimulates a deeper investigation on the occurrence of regime N and L during offense and defense phases, respectively.

In the second step, we carried out a deeper investigation on the dynamics of the two regimes detected in the first step. To start, we inspected contingency tables of regimes and game phases (offense or defense) and of regimes and players, separately for offense and defense. In all the case studies, the association between regimes and game phase is quite strong (normalized association index between 43.95% and 56.22%), but not huge. This confirms both the initial evidence of an association between narrow (large) surface areas and defense (offense) and, at the same time, the necessity of understanding the dynamics and the effects of departures from it, as the matching is not perfect. With reference to players (Tables 2–4), interesting remarks can be drawn with respect to some selected ones. For example, in CS1, during offensive phases (box A), when player $p1$ is on the bench, the relative frequency of regime N is 0.303. When the same player is on the court, the same relative frequency decreases to 0.207.

This means that, on average, the team tends to play more spread in offense when player $p1$ is on the court. The opposite holds with player $p2$: on average the team tends to narrow its offensive surface area when he is on the court, with the relative frequency of regime N moving

from 0.194 to 0.285 when he is on the bench or on the court, respectively.

Similar remarks can be made with reference to other players and/or defensive phases (box B). This allows to describe the effect of each player on the game played by the team. We are not allowed to know whether a specific effect due to the presence of a player positively or negatively affects the team performance, but this can anyway be useful for the coach, who can (for example) measure the extent to which his guidelines with reference to surface areas are complied with.

After the inspection of contingency tables, we defined the smoothed functions $\Phi_D^{(L)}(t)$ and $\Phi_O^{(L)}(t)$, that allow a useful graphical inspection of the regimes patterns during offense and defense. Also this tool can help coaches in analysing the game played by the team. In fact, fluctuations in the functions $\Phi_D^{(L)}(t)$ and $\Phi_O^{(L)}(t)$ shed light on game portions with different features from the point of view of the surface area. As an example of their possible use, consider the two functions computed for CS1 and plotted in Figures 1 and 2, with superimposed grey areas corresponding to the presence on the court of a specific lineup or player. $\Phi_D^{(L)}(t)$ is represented with a blue line and $\Phi_O^{(L)}(t)$ with a red one. We may notice that the offensive play until the first half of the second quarter (around minute 15) seems to be appreciably different from what follows, as the kernel smoothed probabilities $\Phi_O^{(L)}(t)$ of regime L are considerably lower in the first 15 min than in the rest of the match. In addition, something happened in the offensive play around minute 30, as the function $\Phi_O^{(L)}(t)$ clearly falls. Similar remarks can be done with reference to defensive phases: a peak is evident in the function $\Phi_D^{(L)}(t)$ in the middle of the match, around minute 20. These observed fluctuations can be related to the presence of a specific lineup on the court (Figure 1) or a certain player (Figure 2).

As regards the lineups, for example, we notice that the fall of $\Phi_O^{(L)}(t)$ around minute 30 occurred when the lineup 2 was on the court, and the same lineup is also related to another game period with low values of $\Phi_O^{(L)}(t)$, at the end of the first quarter, and periods with increased values of $\Phi_D^{(L)}(t)$. Analogous remarks can be done with reference to single players. Again, at this step, we are not able to assess whether the observed fluctuations have a positive or negative impact on the overall team performance, that was the aim of step 3. Anyway the proposed charts allow a deeper knowledge of the surface area dynamics during the match and with reference to the lineups and players. In addition further relevant game variables (e.g. the implementation of playbooks, the coach's judgments on the technical performance during the match, ...) could be added to the graphs and compared to the observed fluctuations of functions $\Phi_D^{(L)}(t)$ and $\Phi_O^{(L)}(t)$.

Table 7: Number of attempted and made shots.

	2-point	3-point	Free throws
Shots of the team (made/attempted)			
CS1	14/31	13/23	17/19
CS2	12/28	11/34	11/11
CS3	15/34	8/21	15/20
Shots of the opponent (made/attempted)			
CS1	24/37	7/23	18/22
CS2	27/46	3/16	13/14
CS3	16/37	11/22	3/7

Finally, the third step aimed at finding evidence about the effect of regimes' dynamics on the points scored by the team and the opponent. While the evidence found in the first two steps was basically common to all the three analysed case studies, the features that emerged at this point seem to be genuinely match-specific. On some level, this can be justified in view of the different tactics decided by coaches or the different ways of playing determined by the interaction of the two specific teams involved in the match. One important information that will allow to give some insights for a possible interpretation of these match-specific results is given by the number of attempted and made shots (Table 7) in the three case studies.

Besides intercepts, the parameters of both models (11) and (12) tend to have low significance (Tables 5 and 6, respectively). With regard to model (11), we found a (weakly significant, $\alpha = 0.1$) evidence for an effect of variable $\nabla\Phi_O^{(L)}(\bar{t}-1)$ on $SP_{\bar{t}}^{\text{team}}$ in CS1 and CS3. The coefficients' estimates (7.951 and 7.394) suggest that an increased probability of $r = L$ in offense has a high positive effect on the points made by the team. In CS2, such influence of the surface area on the scored points was not detected. Instead, we found a negative correlation between the points scored during 1 min on the points scored in the following one, which highlights a certain variability in the game under this point of view. Looking at Table 7, we notice that the team of CS2 has resorted to 3-point shots more often that the other two (and with a lower scoring percentage). This could explain the variability of the scored points in successive times, as points gained with 3-point shots are more volatile. This variability may have masked the possible relationship between surface area and scored points, or the absence of this relationship may be due right to the different tactic, more based on 3-point shots.

In model (12) we found a significant ($\alpha = 0.05$) effect of variable $SP_{\bar{t}-1}^{\text{opp}}$ on $\nabla\Phi_D^{(N)}(\bar{t})$ for CS1 and CS2. The

coefficients' estimates (-0.016 and -0.012) suggest a weak negative effect of the points made by the opponent on the variation of probability to be in regime $r = N$ in defense. So, when $SP^{\text{opp}}_{\bar{t}-1}$ is high, the probability to be in regime $r = N$ at time $\bar{t}$ decreases: a more efficient game of the opponent (in terms of scored points) seems to force the team to keep the defense spread. Such relationship was not detected in CS3. Looking at Table 7 we observe a substantial difference between cases CS1/CS2 and case CS3 from the point of view of the points scored by the opponent: the teams of CS1 and CS2 have faced opponents with high scoring percentages in 2-point shots (65% and 59%) and low scoring percentages in 3-point shots (30% and 19%), while the opposite has happened to the team of CS3 (43% in 2-point shots and 50% in 3-point shots). This may explain the reason for a different tactical reaction to the points scored by the opponent.

5 Concluding remarks

The analysis of players' trajectories in team sports, which nowadays is made possible by the increasing Information Technology potentialities, could bring important information to analysts, experts and coaches, offering both interpretable analysis tools and concrete suggestions about the way to improve performance. We contribute to this growing area of research by adopting a statistical modelling approach to study relationships among surface areas and team performance in basketball. To the best of our knowledge, Markov Switching models are used here for the first time in basketball studies.

Using sensor data concerned with players' movements on the court, we analysed the surface areas (measured by the area of convex hulls), since the area covered by the team on the court is often considered, by experts, an important perspective on the game. We proposed a three-step procedure. In the first step we detected structural changes in the space among players by means of a Markov Switching model. After a deep analysis of these structural changes with respect to game phases, players and lineups on the court (step 2), in step 3 we studied, by means of Vector Autoregressive models, the causal relation between surface area and team performance.

The main findings of the presented case study can be summarized as follows: we found (i) robust evidences of the presence of structural changes, by estimating two well-separated regimes that relate, respectively, to large and narrow convex hull areas; (ii) some players and some combination of players on the court that show different probabilities to be in the two regimes; (iii) some match-specific causal relationships, which have to be interpreted with reference to the specific characteristics of each match.

These results could be used by basketball coaches and experts as they relate with tactics, specifically with the choice of game strategies, players and lineups. Further developments could be carried out following three directions: (i) to jointly analyse the trajectories of the players of both the team and the opponent; (ii) to introduce the ball trajectory and measure its role on the relationship between surface area and performance; (iii) to include the dimension of game schemes and coaches' tactics in the context of a three-dimensional analysis covering game strategy, players' movements and team performance.

Acknowledgments: The authors are grateful to the anonymous reviewers for their valuable comments, which greatly improved the paper, and also thank MYagonism (`www.myagonism.com`) for having supplied the data. A special thanks goes to Raffaele Imbrogno ("Foro Italico" University, Roma IV), Paolo Raineri (MYagonism) for fruitful discussions, and to Tullio Facchinetti (University of Pavia) for the help with data manipulation. We also thank Giuseppe Arbia (Catholic University of the Sacred Heart) and Marcello Chiodi (University of Palermo) for useful comments at the SIS 2017 conference. Research carried out in collaboration with the Big & Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title "Big Data Analytics in Sports", bodai.unibs.it/bdsports), granted by Fondazione Cariplo and Regione Lombardia.

References

Annis, D. H. 2006. "Optimal End-Game Strategy in Basketball." *Journal of Quantitative Analysis in Sports* 2(2):1.

Araújo, D. and K. Davids. 2016. "Team Synergies in Sport: Theory and Measures." *Frontiers in Psychology* 7:1449.

Araújo, D., K. Davids, and R. Hristovski. 2006. "The Ecological Dynamics of Decision Making in Sport." *Psychology of Sport and Exercise* 7(6):653–676.

Araújo, D., K. W. Davids, J. Y. Chow, P. Passos, and M. Raab. 2009. "The Development of Decision Making Skill in Sport: An Ecological Dynamics Perspective." in *Perspectives on Cognition and Action in Sport*. Suffolk, USA: Nova Science Publishers, Inc., pp. 157–169.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics* 41(1):164–171.

Brandt, P. T. and T. Sandler. 2012. "A Bayesian Poisson Vector Autoregression Model." *Political Analysis* 20(3):292–315.

Brown, M. and J. Sokol. 2010. "An Improved LRMC Method for NCAA Basketball Prediction." *Journal of Quantitative Analysis in Sports* 6(3):1–23.

Cooper, W. W., J. L. Ruiz, and I. Sirvent. 2009. "Selecting Non-Zero Weights to Evaluate Effectiveness of Basketball Players with DEA." *European Journal of Operational Research* 195(2):563–574.

Crocker, P. R. and T. R. Graham. 1995. "Coping by Competitive Athletes with Performance Stress: Gender Differences and Relationships with Affect." *The Sport Psychologist* 9(3):325–338.

Deshpande, S. K. and S. T. Jensen. 2016. "Estimating an NBA Player's Impact on his Team's Chances of Winning." *Journal of Quantitative Analysis in Sports* 12(2):51–72.

Duarte, R., D. Araújo, V. Correia, K. Davids, P. Marques, and M. J. Richardson. 2013. "Competing Together: Assessing the Dynamics of Team–Team and Player–Team Synchrony in Professional Association Football." *Human Movement Science* 32(4):555–566.

Fearnhead, P. and B. M. Taylor. 2011. "On Estimating the Ability of NBA Players." *Journal of Quantitative Analysis in Sports* 7(3):1–18.

Fewell, J. H., D. Armbruster, J. Ingraham, A. Petersen, and J. S. Waters. 2012. "Basketball Teams as Strategic Networks." *PLoS One* 7(11): e47445.

Fonseca, S., J. Milho, B. Travassos, and D. Araújo. 2012. "Spatial Dynamics of Team Sports Exposed by Voronoi Diagrams." *Human Movement Science* 31(6):1652–1659.

Frencken, W., K. Lemmink, N. Delleman, and C. Visscher. 2011. "Oscillations of Centroid Position and Surface Area of Soccer Teams in Small-Sided Games." *European Journal of Sport Science* 11(4):215–223.

Goldfarb, D. 2014. 1'An Application of Topological Data Analysis to Hockey Analytics." *arXiv preprint arXiv:1409.7635*.

Greihaine, J.-F., P. Godbout, and Z. Zerai. 2011. "How the "Rapport de Forces" Evolves in a Soccer Match: The Dynamics of Collective Decisions in a Complex System." *Revista de Psicología del Deporte* 20(2):747–764.

Gudmundsson, J. and M. Horton. 2016. "Spatio-Temporal Analysis of Team Sports–A Survey." *arXiv preprint arXiv:1602.06994*.

Gupta, A. A. 2015. "A New Approach to Bracket Prediction in the NCAA Men's Basketball Tournament Based on a Dual-Proportion Likelihood." *Journal of Quantitative Analysis in Sports* 11(1):53–67.

Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica: Journal of the Econometric Society* 57(2):357–384.

Hamilton, J. D. 2010. "Regime Switching Models." Pp. 202–209 in *Macroeconometrics and Time Series Analysis*. London: Palgrave Macmillan.

Kass, R. E. and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90(430):773–795.

Kim, C.-J. 1994. "Dynamic Linear Models with Markov-Switching." *Journal of Econometrics* 60(1–2):1–22.

Kowshik, G., Y.-H. Chang, and R. Maheswaran. 2012. *Visualization of Event-Based Motion-Tracking Sports Data*. Technical report, Technical report, University of Southern California.

Kubatko, J., D. Oliver, K. Pelton, and D. T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports* 3(3):1–22.

Lamas, L., D. D. R. Junior, F. Santana, E. Rostaiser, L. Negretti, and C. Ugrinowitsch. 2011. "Space Creation Dynamics in Basketball Offence: Validation and Evaluation of Elite Teams." *International Journal of Performance Analysis in Sport* 11(1):71–84.

Lindgren, G. 1978. "Markov Regime Models for Mixed Distributions and Switching Regressions." *Scandinavian Journal of Statistics* 5(2):81–91.

Loeffelholz, B., E. Bednar, and K. W. Bauer. 2009. "Predicting NBA Games using Neural Networks." *Journal of Quantitative Analysis in Sports* 5(1):1–15.

Lopez, M. J. and G. J. Matthews. 2015. "Building an NCAA Men's Basketball Predictive Model and Quantifying its Success." *Journal of Quantitative Analysis in Sports* 11(1):5–12.

Manner, H. 2016. "Modeling and Forecasting the Outcomes of NBA Basketball Games." *Journal of Quantitative Analysis in Sports* 12(1):31–41.

Metulini, R. 2016. "Spatio-Temporal Movements in Team Sports: A Visualization Approach Using Motion Charts." *arXiv preprint arXiv:1611.09158*.

Metulini, R. 2017. "Filtering Procedures for Sensor Data in Basketball." *Statistica & Applicazioni* 15(2).

Metulini, R., M. Manisera, and P. Zuccolotto. 2017a. "Sensor Analytics in Basketball." *Proceedings of the 6th International Conference on Mathematics in Sport. ISBN 978-88-6938-058-7*.

Metulini, R., M. Manisera, and P. Zuccolotto. 2017b. "Space-Time Analysis of Movements in Basketball Using Sensor Data." *Statistics and Data Science: New Challenges, New Generations" SIS2017 proceeding. Firenze Uiversity Press. eISBN: 978-88-6453-521-0*.

Moura, F. A., L. E. B. Martins, R. D. O. Anido, R. M. L. De Barros, and S. A. Cunha. 2012. "Quantitative Analysis of Brazilian Football Players' Organisation on the Pitch." *Sports Biomechanics* 11(1):85–96.

Nadaraya, E. A. 1964. "On Estimating Regression." *Theory of Probability & Its Applications* 9(1):141–142.

Ozmen, M. U. 2012. "Foreign Player Quota, Experience and Efficiency of Basketball Players." *Journal of Quantitative Analysis in Sports* 8(1):1–18.

Page, G. L., G. W. Fellingham, and C. S. Reese. 2007. "Using Box-Scores to Determine a Position's Contribution to Winning Basketball Games." *Journal of Quantitative Analysis in Sports* 3(4):1–18.

Page, G. L., B. J. Barney, and A. T. McGuire. 2013. "Effect of Position, Usage Rate, and Per Game Minutes Played on NBA Player Production Curves." *Journal of Quantitative Analysis in Sports* 9(4):337–345.

Passos, P., D. Araújo, and A. Volossovitch. 2016. *Performance Analysis in Team Sports*. London: Routledge.

Passos, P., K. Davids, D. Araújo, N. Paz, J. Minguéns, andJ. Mendes. 2011. "Networks as a Novel Tool for Studying Team Ball Sports as Complex Social Systems." *Journal of Science and Medicine in Sport* 14(2):170–176.

Perica, A., S. Trninić, and I. Jelaska. 2011. "Introduction into the Game States Analysis System in Basketball." *Fizička kultura* 65(2):51–78.

Perše, M., M. Kristan, S. Kovačič, G. Vučkovič, and J. Perš. 2009. "A Trajectory-Based Analysis of Coordinated Team Activity in a Basketball Game." *Computer Vision and Image Understanding* 113(5):612–621.

Piette, J., S. Anand, and K. Zhang. 2010. "Scoring and Shooting Abilities of NBA Players." *Journal of Quantitative Analysis in Sports* 6(1):1–23.

Richardson, M. J., R. L. Garcia, T. D. Frank, M. Gergor, and K. L. Marsh. 2012. "Measuring Group Synchrony: A Cluster-Phase Method for Analyzing Multivariate Movement Time-Series." *Frontiers in physiology* 3(405):1–10.

Ruiz, F. J. and F. Perez-Cruz. 2015. "A Generative Model for Predicting Outcomes in College Basketball." *Journal of Quantitative Analysis in Sports* 11(1):39–52.

Sanchez-Espigares, J. A. and A. Lopez-Moreno. 2014. *MSwM: Fitting Markov Switching Models*. R package version 1.2. URL:https://CRAN.R-project.org/package=MSwM.

Sims, C. A. 1980. "Macroeconomics and reality." *Econometrica: Journal of the Econometric Society* 48(1):1–48.

Stein, M., H. Janetzko, D. Seebacher, A. Jäger, M. Nagel, J. Hölsch, S. Kosub, T. Schreck, D. A. Keim, and M. Grossniklaus. 2017. "How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects." *Data* 2(1):2.

Therón, R. and L. Casares. 2010. "Visual Analysis of Time-Motion in Basketball Games." in *International Symposium on Smart Graphics*. Berlin, Heidelberg: Springer, pp. 196–207.

Travassos, B., D. Araújo, K. Davids, P. T. Esteves, and O. Fernandes. 2012. "Improving Passing Actions in Team Sports by Developing Interpersonal Interactions between Players." *International Journal of Sports Science & Coaching* 7(4):677–688.

Travassos, B., D. Araújo, R. Duarte, and T. McGarry. 2012. "Spatiotemporal Coordination Behaviors in Futsal (Indoor Football) are Guided by Informational Game Constraints." *Human Movement Science* 31(4):932–945.

Turvey, M. and R. E. Shaw. 1995. "Toward an Ecological Physics and a Physical Psychology." *The Science of the Mind: 2001 and Beyond*, Chapter 11, pp. 144–169.

Wasserman, S. and K. Faust. 1994. *Social Network Analysis: Methods and Applications*, Vol. 8. Cambridge, United Kingdom: Cambridge University Press.

Watson, G. S. 1964. "Smooth Regression Analysis." *Sankhyā: The Indian Journal of Statistics, Series A* 26(4):359–372.

West, B. T. 2008. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007." *Journal of Quantitative Analysis in Sports* 4(2):8.

Yuan, L.-H., A. Liu, A. Yeh, A. Kaufman, A. Reece, P. Bull, A. Franks, S. Wang, D. Illushin, and L. Bornn. 2015. "A Mixture-of-Modelers Approach to Forecasting NCAA Tournament Outcomes." *Journal of Quantitative Analysis in Sports* 11(1):13–27.

Zuccolotto, P., M. Manisera, and M. Sandri. 2017. "Big Data Analytics for Modeling Scoring Probability in Basketball: The Effect of Shooting under High-Pressure Conditions." *International Journal of Sports Science & Coaching* (OnLine First).