


Court Surfaces: A Compositional Approach to Tennis Analytics

Pepus Daunis-i-Estadella 

University of Girona
Catalonia, Spain

Ernest Baiget 

National Institute of Physical Education of
Catalonia (INEFC-UB), Spain

Martí Casals 

National Institute of Physical Education of
Catalonia (INEFC-UB), Spain

Abstract

Compositional data analysis (CoDA) has been extensively applied in fields such as microbiomics, geosciences, and health sciences, yet its potential in sports analytics remains largely untapped. This study applies CoDA to analyze the proportional distribution of matches played on different tennis court surfaces —clay, hard, and grass— and its influence on player rankings and performance. Data from 1,171 Association of Tennis Professionals (ATP) players ranked in the top 100 were examined, revealing significant differences in surface composition across ranking categories ($p < 0.001$). High-ranked players dedicated 57.8% of their total minutes to hard courts, compared to 32.6% on clay and 9.6% on grass, while lower-ranked players showed a lower proportion of tournaments on hard and a higher proportion on grass. Temporal trends indicated a shift from clay dominance in earlier decades to a predominance of hard courts, with players from the 2010s dedicating 59.8% of their tournament play to hard surfaces. These findings underscore the potential of CoDA to uncover nuanced relationships in sports data, providing actionable insights for optimizing strategies in tennis performance.

Keywords: compositional data analysis, tennis analytics, court surfaces, player ranking, sports statistics.

1. Introduction

Sports analytics has emerged as a rapidly growing field, gaining heightened attention due to its impact on decision-making in sports, particularly popularized by the movie *Moneyball* and the statistical insights from several academic works [Albert and Koning \(2007\)](#); [Dominicy and Ley \(2023\)](#). The establishment of the American Statistical Association's (ASA) Section on Statistics in Sports in 1992 (<https://community.amstat.org/sis/home>) marked a key moment, responding to the growing demand for statistical methodologies within sports.

Over time, this field has continued to evolve, with more specialized journals, conferences and

the establishment of various sports analytics and statistics departments at universities worldwide, including Brigham Young (<https://science.byu.edu/research/statistics/sports-analytics>), Brescia (<https://bodai.unibs.it/bdsports/>), Simon Fraser Univ. (<https://www.sfu.ca/sports-analytics-group.html>), Harvard (<https://sportsanalytics.stat.harvard.edu/>), Carnegie Mellon Univ. (<https://www.stat.cmu.edu/cmsac/>), and Victoria (<https://www.vu.edu.au/institute-for-health-sport-ihes/research-areas-in-ihes/sport-performance-business>). The rise of this field in industry was followed by an increased acceptance in academia, as Syracuse University established the first bachelor's degree program for sports analytics in the United States in 2016.

In recent years, sports analytics -also referred to as sports statistics (Alamar (2013))— has expanded into diverse subfields. A prominent example is sabermetrics, the "search for objective knowledge about baseball" (Albert (2010)). Sabermetrics, named after the Society for American Baseball Research (SABR), revolutionized baseball analysis and set a precedent for data-driven approaches across other sports. Another critical domain is sports biostatistics, which applies statistical methods to improve athlete health, performance, and injury prevention (Casals and Finch (2018); Sainani, Borg, Caldwell, Butson, Tenan, Vickers, Vigotsky, Warmenhoven, Nguyen, Lohse, Knight, and Bargary (née Coffey) (2020)). These advancements highlight the increasing importance of evidence-based methodologies for optimizing strategies and ensuring athlete well-being. The Handbook of Statistical Methods and Analyses in Sports (Albert, Glickman, Swartz, and Koning (2016)) provides an up-to-date overview of research in this area, covering applications of statistical techniques to sports such as baseball, football, and basketball. Recent work by Baumer, Matthews, and Nguyen (2023) further outlines advanced methods like Bradley-Terry and Elo models for evaluating team and player performance, hierarchical models for complex data, and clustering techniques to identify patterns in player or team characteristics. These methodologies are essential for making informed decisions about sports analytics in a wide range of sports contexts.

Tennis is a complex sport, and performance is influenced by various closely interrelated factors such as strategy, physical condition, mentality, or technique. Consequently, a player's tennis performance is a multifaceted construct determined by a wide range of skills. Within this context, research on tennis has been conducted in each of the aforementioned areas (Pluim, Jansen, Williamson, Berry, Camporesi, Fagher, Heron, Janse van Rensburg, Perez, Murray, O'Connor, de Oliveira, Reid, Reijen, Saueressig, Schoonmade, Thornton, Webborn, and Ardern (2023); Crespo, Martínez-Gallego, and Filipcic (2024)). Despite the vast amount of data available in tennis, the sport has not yet experienced its full "Moneyball" moment (Kovalchik (2021)). The potential for an analytics revolution in tennis has been hindered by barriers between the data and the key stakeholders -players, coaches, and teams- who would benefit most from its analysis. While undoing these business-driven obstacles may be challenging, advances in data science and sports broadcast video analysis could play a pivotal role in ushering tennis into a new era of statistical thinking. In recent years, techniques like machine learning and other statistical methods have been applied to address some of the sport's most significant challenges, offering new ways to analyze performance data and inform strategic decisions (Leitner, Zeileis, and Hornik (2009); Giles, Peeling, Kovalchik, and Reid (2021); Gao and Sun (2024)).

Although tennis has lagged behind other sports in fully embracing advanced analytics, recent studies have begun applying machine learning and statistical techniques to address key performance-related questions. For example, survival models have been used to predict injuries and their impact on match outcomes (Whiteside, Cant, Connolly, and Reid (2017)), while spatio-temporal hierarchical models have been employed to analyze player movement and tactical decision-making (Zhou, Zong, Cao, Ruano, Chen, and Cui (2023)). Other studies have explored the role of court surfaces on player performance using machine learning algorithms (Whiteside *et al.* (2017)) and clustering techniques to identify patterns (Sampaio, Oliveira, Marinho, Neiva, and Morais (2024)). In addition, fatigue prediction models and neural networks for match outcomes have also emerged, contributing to our understanding of

player behavior in different game contexts (Giles *et al.* (2021)). These applications demonstrate the growing influence of advanced analytics in tennis. Most analyses of sports data have traditionally relied on statistical methods that focus on absolute values in the data. While these approaches offer advantages in how data are modeled and interpreted, they share an important limitation - they are not well-suited for data that convey relative information, such as time-use data or proportions. Compositional Data Analysis (CoDA) (Aitchison (1986)) takes a different approach by addressing the relative nature of data components, ensuring that the inherent relationships between parts are preserved and interpreted correctly. This makes CoDA based on log-ratio analysis (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2015)) particularly advantageous for analyzing data where the composition of elements is critical, such as the time spent in matches across different tennis court surfaces.

CoDA has proven to be highly effective across a wide range of scientific fields (Navarro Lopez, Gonzalez-Morcillo, Forteza, and Linares-Mustarós (2021)) such as microbiome (Gloor, Macklaim, Pawlowsky-Glahn, and Egozcue (2017)), geosciences (Martín-Fernández, Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado (2017)), and health (Dumuid, Martín-Fernández, El-lul, Kenett, Wake, Simm, Baur, and Olds (2020)). However, despite its potential, CoDA remains underutilized in sports analytics. Most existing research has focused on physical activity (Chastin, Palarea-Albaladejo, Dontje, and Skelton (2015); Dumuid, Wake, Clifford, Burgner, Carlin, Mensah, Frayssé, Lycett, Baur, and Olds (2019); Verswijveren, Lamb, Martín-Fernández, Winkler, Leech, Timperio, Salmon, Daly, Cerin, Telford, Telford, Olive, and Ridgers (2021)), but to our knowledge, CoDA has yet to be applied to sports sciences or sports analytics. This presents a clear gap and an opportunity to apply CoDA to address key questions in this field, such as how the composition of court surfaces played on affects player performance outcomes, including best ranking category, backhand type, or decade of debut, in order to capture the evolution of the tennis ball and the trend of surface homogenization. By leveraging CoDA, we can develop new insights to understanding how relative data impacts sports outcomes. CoDA are nonnegative data carrying relative, rather than absolute, information (Aitchison and Ng (2005)). These are often data with a constant-sum constraint on the sample values, for example proportions or percentages that add up to a sum of 1 or 100%, respectively.

In sports analytics, compositional structures naturally emerge in several contexts. A common example in football is percentage of ball possession, where the time a team controls the ball is divided into defensive, midfield, and attacking phases. An increase in possession in one area necessarily reduces the proportion in the others, making direct correlations between possession metrics misleading. For instance, consider two teams where the proportions of defensive, midfield, and attacking phases are analyzed. In team A, defensive constitutes 50%, midfield 40%, and attack 10%, while in team B, defensive is 60%, midfield 30%, and attack 10%. If we plot the relationship between the proportions of defensive and midfield phases, we might observe a strong negative correlation. However, this correlation is spurious, as it arises purely from the fact that increasing the proportion of one category (e.g., defensive) necessarily decreases the proportions of the other categories (e.g., midfield and attack), regardless of any genuine sport relationship between these categories. Similarly, in basketball, the distribution of shot attempts among one-point (free throws), two-point, and three-point shots forms a composition. The relative emphasis on one type of shot affects the proportions of the others, making CoDA a valuable tool for analyzing offensive strategies.

This constraint inherently creates interdependence among the proportions, leading to correlations that do not reflect genuine sport relationships but rather mathematical artifacts. The problem of spurious correlations that results from constant-sum constraint has been known for over a century (Pearson (1896)), and even compositional data based on random counts exhibit important correlations due to this closure (Aitchison (1982)). Ratios between components of a composition are important since they are unaffected by the particular set of components chosen. Logarithms of ratios (logratios) are the fundamental transformation in the ratio approach to compositional data analysis, all data thus need to be strictly positive,

so that zero values present a major problem. There are several types of logratio transformations that rely on geometric means to combine parts, the additive logratio (alr): [Aitchison \(1982\)](#); the centered logratio (clr): [Aitchison \(1986\)](#); and the isometric logratio (ilr): [Egozcue, Pawlowsky-Glahn, Figueras, and Vidal \(2003\)](#). Logratio transformations are fundamental tools in CoDA, enabling the application of standard statistical techniques. These transformations map compositional data from a constrained space into an unconstrained Euclidean space. By doing so, they remove the problematic effects of the constant-sum constraint and spurious correlations. Among these, the ilr transformation is particularly notable for its orthogonality and ability to provide interpretable results in terms of balances between parts of the composition.

One of the most critical factors in tennis analytics is the court surface. Different surfaces—clay, hard, and grass courts—pose unique challenges due to the biomechanical differences and friction properties of each one ([Miller \(2006\)](#)). It has been widely shown how the playing surface has an important influence on the playing style, the physical and physiological demands of a competition match, and could exert an influence on the retirement incidence in professional tennis ([Fernandez-Fernandez, Kinner, and Ferrauti \(2010\)](#); [Oliver, Baiget, Cortés, Martínez, Crespo, and Casals \(2024\)](#)). This makes surface type a key confounding factor in any research question related to tennis analytics. While many studies have investigated surface-related factors, important remain unexplored, especially regarding how the relative composition of surfaces played on varies across players with different ranking levels and performance characteristics. Furthermore, although, it is well known how the biometric variables such as weight, height, and age or player characteristics (e.g., dominant hand or backhand style) have an impact on his performance ([Baiget, Corbi, and López \(2022\)](#); [Wong, Keung, Lau, Ng, Chung, and Chow \(2014\)](#)), it has not been determined how these parameters could be related to surface composition, influencing player outcomes in ways that are not yet fully understood.

For this reason, the objective of this study is to examine how the proportional distribution of matches played on different court surfaces—clay, hard, and grass— influences ranking categories and performance characteristics in professional tennis, using CoDA. By addressing the relative nature of match distribution, often overlooked by traditional statistical methods, CoDA allows us to identify surface-specific playing patterns and their association with player outcomes. The study aims to provide a nuanced understanding of how court surface composition relates to ranking and performance, offering valuable insights for players, coaches, and tennis analysts.

1.1. Study design

A retrospective cohort study was performed to analyze the relationship between the distribution of matches played on different court surfaces and men's professional tennis performance.

1.2. Participants

The sample comprises 1,174 male professional tennis players who have reached the Association of Tennis Professionals (ATP) top 100 singles ranking at least once during their careers. This includes both active and former players, ensuring a comprehensive representation of professional-level performance across different eras. Players with incomplete data or missing records for surface-specific participation were excluded from the analysis to maintain the integrity of the dataset.

Data

Data were collected retrospectively for all players who have been ranked in the ATP Top 100 at least once, with a cutoff date of September 12, 2022. The dataset was constructed by scraping information from the official ATP Tour website. The data is available on a publicly

accessible GitHub repository https://github.com/marticasals/CoDA_tennis.

Variables include the number of matches played on different surfaces (clay, hard, and grass court), the tournament categories (Grand Slams, Masters, ATP 250 or ATP 500), and the highest ATP ranking achieved by each player. From this variable, a new variable called ranking category was created, classifying players into three ranking groups: *High* for rankings 1–10, *Medium* for rankings 11–50, and *Low* for rankings 51–100. Additionally, anthropometric variables such as weight, height, and age were recorded, along with player characteristics such as dominant hand and backhand type (one-handed or two-handed). Performance-related variables were included following a hierarchical framework to account for varying levels of precision and control. These variables comprise the total number of matches played (a more controllable factor for players), the total number of games played (an intermediate metric influenced by match outcomes), and the total minutes played (a precise but uncertain variable determined by the length and dynamics of each game). This hierarchy allows for a nuanced analysis of player load and its potential impact on performance.

Statistical analysis

To investigate differences between groups, we performed a compositional data analysis with CoDA-based descriptives: compositional means were calculated by normalizing the geometric means to add up to 100% and the total variance as a measure of the overall dispersion.

Ternary composition analysis was used to visualize the relationships between playing surfaces and our variable of interest. The frequency of one variable is maximum at the vertex. As we move downward, the percentage of that part decreases and becomes zero on the opposite line.

Multivariate analysis of variance (MANOVA) was applied to the ilr-transformed data to determine the effects of our independent variables (surface composition) on g groups of parts, using the corresponding p-value as a metric to evaluate statistical significance, in this case to examine whether the distribution of tournaments on different surfaces influences player rankings or performances.

According to the principle of working on coordinates (Mateu-Figueras and Egozcue (2013)), the MANOVA model $H_0 : \mu_1 = \dots = \mu_g$, where μ is the mean vector (clay, hard, grass), and their classical assumptions are equivalent to assuming logratio normality and homoscedasticity for the ilr-coordinates. In addition, the original null hypothesis $H_0 : \mu_1 = \dots = \mu_g$ is equivalent to the null hypothesis $H_0 : \text{ilr}(\mu_1) = \dots = \text{ilr}(\mu_g)$. Therefore, the statistics of contrast based on the sum of square matrices of data distributed in g groups, the matrices sum of squares total (T), between-groups (B), and within-groups sums of squares matrices (W) will be calculated on ilr-coordinates. The most common contrast statistics are the following: Wilks' λ ($\det(W)/\det(T)$), Pillai's trace ($\text{trace}(BT^{-1})$), Lawley-Hotelling trace ($\text{trace}(W^{-1}B)$), and Roy's largest root of matrix ($W^{-1}B$). Importantly, the compositional MANOVA contrast is invariant under a change of log-ratio basis because the four statistics are invariant functions of the eigenvalues of matrix $W^{-1}B$.

To understand and complete the results of the multivariate tests, separate t-tests, one for each ilr log-ratio, were performed to evaluate the contribution of each ilr to any possible difference between the independent variables (Martín-Fernández, Daunis-i Estadella, and Mateu-Figueras (2015)). Post-hoc analysis is performed with the pairwise t-tests, using a multiple comparison correction. The differences are presented with compact letter display (CLD), labelling different groups with different letters. Each category that shares a compositional mean that is not statistically different from another category shares the same letter.

When we have two or more groups in our data set, we can use the geometric-mean bar plot to graphically compare the centers. For each group, we initially compute the ratio between the overall geometric mean and the geometric mean of the group. Subsequently, each part is visualized in a bar plot with a logarithmic scale. When the group's center aligns with the overall center, each component's ratio equals 1, resulting in a zero in log-scale. Conversely, if a group's center differs from the overall center, the ratio deviates from one, yielding a positive or

negative logarithm. Therefore, large bars (positive or negative) indicate substantial disparities in means. It is important to note that it is not possible to log-transform zeros or calculate geometric means. However, in our dataset, there were no zeros present. The top 100 players represent the highest level of world tennis, and to reach this level of competition, tennis players compete on all surfaces at least once in their careers. For instance, the top 100 players can directly enter the main draws of the four Grand Slams, ensuring that they play on clay, hard court, and grass.

All data analysis and statistical modelling was performed using Codapack Software v.2.03.01 (Comas-Cufí and Thió-Henestrosa (2011)) and the R system for statistical computing v4.1.1 (R Core Team (2021)), package *ggtern* (Hamilton and Ferry (2018)). Statistical significance was generally concluded at the usual 5% significance level ($p < 0.05$).

2. Results

A total of 1,171 players from the historical ATP top 100 singles rankings were initially included in the study. Players who had not competed on all three court surfaces ($n=43$) and the only ambidextrous player ($n=1$) were excluded from the analysis. As a result, 1,127 players were included for analyses involving the variable *number of tournaments*, and 1,126 players for the variable *number of games*. The total number of players was further reduced in the analysis of *minutes played* due to incomplete data on match durations on all surfaces, resulting in a final sample of 705 players.

The three ternary diagrams (Figure 1) visualize the distribution of the number of tournaments (1), the number of games (2), and the total number of minutes played (3) on three different surfaces of tennis courts: clay, hard, and grass. The data points within each ternary plot indicate different proportions of play on these surfaces.

In all three diagrams, there is a higher concentration of points toward the clay and hard court edges, suggesting that players tend to play more tournaments, games, and minutes on these surfaces rather than on grass. (Figures 1 (a) and (b)) show the distribution of tournaments and games played on each surface. Most tournaments appear to have a higher proportion on clay and hard courts. In (Figure 1 (c)) the distribution appears to follow the same pattern, but with a smaller proportion of playing time on grass, likely because this court is faster.

The three geometric centers of (clay, hard, grass) are: tournaments=(40.4%, 47.1%, 12.5%), games=(39.3%, 46.6%, 14.1%), and minutes (34.4%, 54.4%, 11.1%). Showing the decreased proportion of minutes played on clay and grass courts towards hard courts. The total variances are equal to 1.460, 1.102, and 1.287, respectively. The number of games played is the composition of the courts with higher variability.

Given the similar pattern that the three graphs have, we will only show from now the ternary plot for tournaments, for the sake of simplicity.

2.1. Ranking categories

Overall, clay and hard courts predominate in terms of the number of tournaments played across all ranking categories (Figure 2). Compared players in the top-ranking category (categorized as *High*) showed a higher frequency of participation in tournaments on hard surfaces, accounting for 50.9% of all tournaments, compared to 38.9% on clay and 10.2% on grass, with low variability (Table 1). Players in lower-ranked categories (Medium and Low) exhibited a lower proportion of tournaments on hard and a higher proportion on grass, as shown in Table 1. The smallest compositional variances are always associated with the high-ranking group.

The distribution of games played and minutes spent on each surface followed a similar pattern across player categories. The top-ranking category players accumulated 57.8% of their total minutes on hard courts (see Table 1), compared to 32.6% on clay and 9.6% on grass.

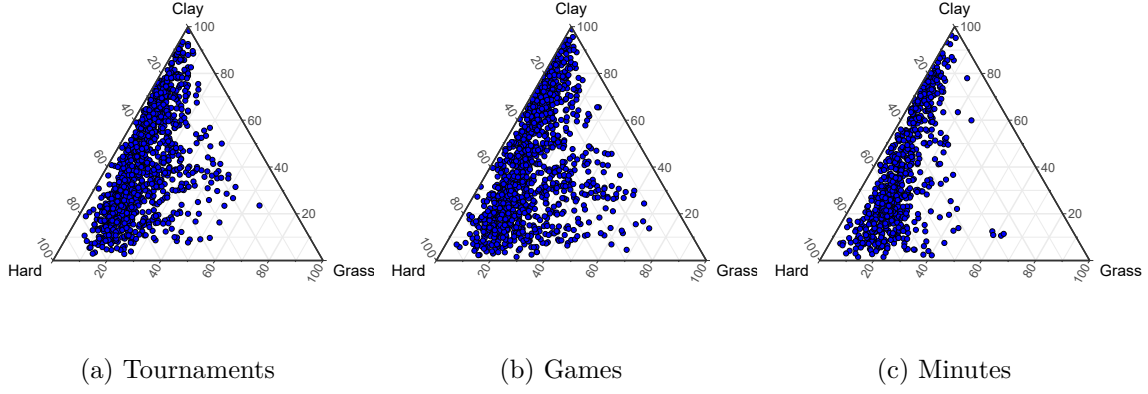


Figure 1: Ternary diagrams of number of tournaments (a), games (b), and minutes played (c) on different court surfaces

Table 1: Distribution of Tournaments, Games, and Minutes Played Across Different Court Surfaces by rank categories: Compositional Center, Total Variance, and Sample Size. Compact letter display (CLD) indicates statistical significance groupings.

	Number of tournaments			Number of games			Minutes played		
Surface	Ranking			Ranking			Ranking		
Clay	High	Medium	Low	High	Medium	Low	High	Medium	Low
Hard	0.389	0.381	0.432	0.377	0.370	0.422	0.326	0.328	0.372
Grass	0.509	0.497	0.430	0.504	0.492	0.428	0.578	0.562	0.508
	0.102	0.122	0.138	0.120	0.139	0.150	0.096	0.111	0.120
Variance	0.695	1.022	1.291	0.992	1.424	1.636	0.847	1.124	1.659
N	176	475	476	176	475	475	129	308	268
CLD	a	b	c	a	a	b	a	b	c

MANOVA test $\mu(\text{clay}, \text{hard}, \text{grass})_{\text{High}} = \mu(\text{clay}, \text{hard}, \text{grass})_{\text{Med.}} = \mu(\text{clay}, \text{hard}, \text{grass})_{\text{Low}}$ indicated significant differences across player categories (High, Medium, Low) ($p < 0.001$) in both tournaments Wilks' $\lambda = 0.942$, $F_{4,2246}(17.064) < 0.001$, number of games Wilks' $\lambda = 0.968$, $F_{4,2244}(9.324) < 0.001$, and minutes played Wilks' $\lambda = 0.953$, $F_{4,1402}(8.609) < 0.001$. Specifically, the high-ranking category players were found to participate in a significantly higher number of tournaments on hard surfaces (Figure 3), and the opposite, the low-ranking category players in a lower number of hard surfaces. Similar patterns are obtained for number of games and minutes played.

The pairwise t-tests showed that ranking categories differed significantly except between High and Medium for number of games, using a multiple comparison correction. Table 1 presents these differences with compact letter display (CLD), labelling different groups with different letters. Each ranking that shares a compositional mean that is not statistically different from another decade shares the same letter.

2.2. Backhand style

Regarding backhand style, players using a two-handed backhand spent more time playing on hard surfaces (57.4%) compared to one-handed backhand players (51.4%), meanwhile unknown backhand players had a slight preference for clay courts (Figure 4).

Specifically, two-handed and one-handed backhand players were found to participate in a significantly higher number of tournaments on hard surfaces (see Table 2), and the opposite unknown backhand players in a lower number of hard surfaces. Similar patterns are obtained for the number of games and minutes played.

MANOVA analysis provides significant differences related to the backhand style and surface types in both tournaments played Wilks' $\lambda = 0.835$, $F_{4,2246}(52.997) < 0.001$, number of

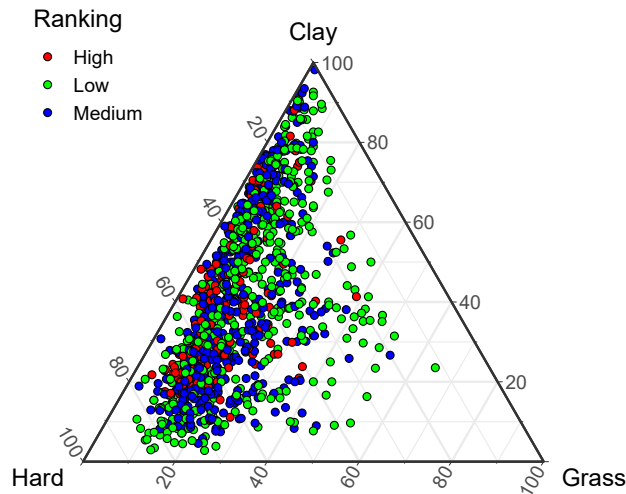


Figure 2: Ternary diagram of number of tournaments played on different surfaces by ranking

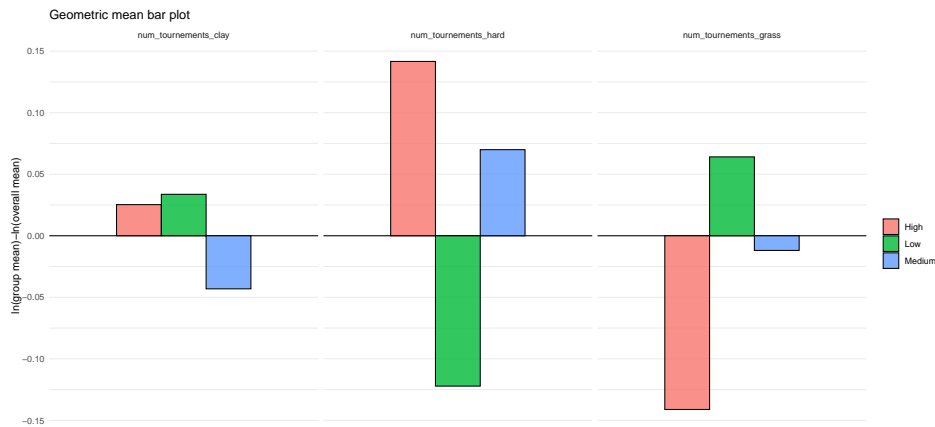


Figure 3: Geometric-mean bar plot comparing the centers of surfaces across ranking categories for tournaments

games Wilks' $\lambda = 0.859$, $F_{4,2244}(44.252) < 0.001$, and minutes played Wilks' $\lambda = 0.981$, $F_{4,1402}(3.451) = 0.008$.

But the pairwise comparisons (see Table 2 CLD) highlight differences. Only the minutes played do not show significant differences between Two backhand style and all the other ones. The geometric-mean bar plot comparing the centers of surfaces and backhand styles for tournaments (Figure 5) shows us the opposition between the two-handed and one-handed backhand and the unknown backhand. This has a slight prevalence on the surfaces of clay and grass.

In addition, there is also a light interaction in tournaments played ($p = 0.020$), number of games ($p = 0.029$), and minutes spent ($p = 0.001$) between back-hand style and ranking category. This is the only interaction among all analyses.

2.3. Decade of debut

The analysis of the decade of debut reveals notable trends in the proportional distribution of court surfaces over time (see Table 3). Over the six decades analyzed, the proportion of matches played on clay surfaces has progressively decreased, while the proportion on hard courts has shown a consistent increase. Grass surfaces experienced a significant decline from the 1960s to the 1980s, followed by a slight rebound in subsequent decades. Notably, a trend break is observed for grass courts in the 1990s. Furthermore, variability in surface preferences

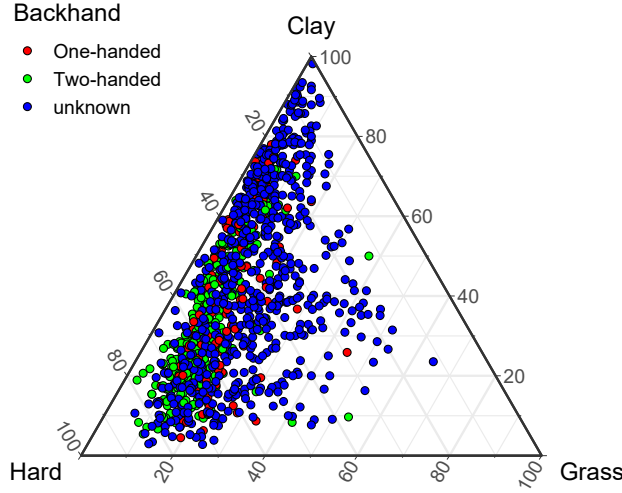


Figure 4: Ternary diagram of number of tournaments played by backhand style

Table 2: Distribution of Tournaments, Games, and Minutes Played Across Different Court Surfaces by backhand style: Compositional Center, Total Variance, and Sample Size. Compact letter display (CLD) indicates statistical significance groupings.

	Number of tournaments			Number of games			Minutes played		
Surface	Backhand style			Backhand style			Backhand style		
	Two	One	Unknown	Two	One	Unknown	Two	One	Unknown
Clay	0.318	0.377	0.452	0.312	0.374	0.436	0.311	0.368	0.375
Hard	0.574	0.514	0.412	0.570	0.508	0.408	0.575	0.529	0.514
Grass	0.108	0.109	0.136	0.118	0.117	0.156	0.114	0.103	0.111
Variance	0.578	0.818	1.324	0.754	1.177	1.765	0.812	1.286	1.825
N	328	133	666	328	133	665	320	117	268
CLD	a	b	c	a	b	c	ab	a	b

among players has been decreasing over time, suggesting a shift toward more uniform surface distributions in professional tennis.

These pattern change can be visualized in (Figure 6), moving from clay-grass towards hard-clay, first, and finally to hard.

MANOVA analysis indicated significant differences across decades and court types ($p < 0.001$) in both tournaments Wilks' $\lambda = 0.631$, $F_{10,2240}(58.057) < 0.001$, number of games Wilks' $\lambda = 0.643$, $F_{10,2238}(55.276) < 0.001$, and minutes played Wilks' $\lambda = 0.934$, $F_{10,1396}(4.863) < 0.001$. But the pairwise comparisons (see CLD Table 3) and the geometric-mean bar plot (Figures 7 and 8) highlight specific patterns and differences.

Specifically, tournaments and the number of games show the same pairwise relationships: the 1970s, 1980s, and 1990s do not show significant differences. The 2000s and 2010s differ from each other and from all other decades. The 1960s do not differ from the 1990s.

The pattern of pairwise relationships of the composition of minutes played on different surfaces is different. The pattern of the 1960s and 1970s is very different. 1960s clay predominates and 1970s increases grass. The last 3 decades go up grass and hard surfaces and go down clay. The 1970s-1980s-2000s-2010s have no significant differences. The 1990s, 2000s and 1970s are similar between them, and finally the 1960s presents significant differences (Table 3).

3. Discussion

This study analyzed the distribution of tournaments, games, and minutes played on different

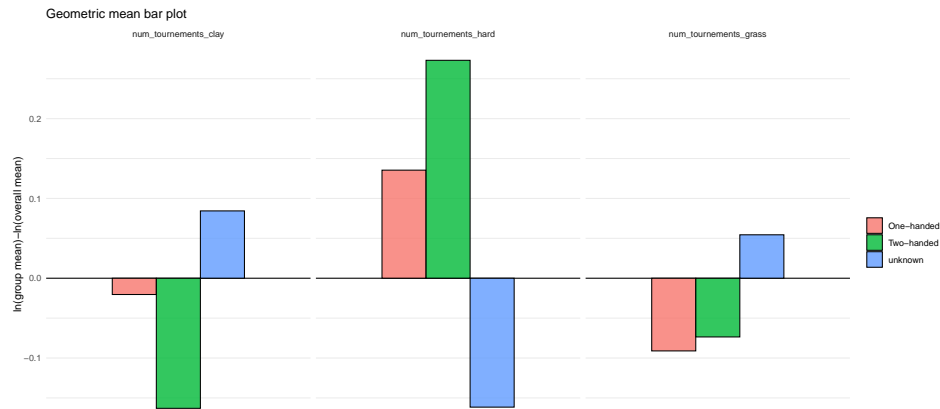


Figure 5: Geometric-mean bar plot comparing the centers of surfaces and backhand style

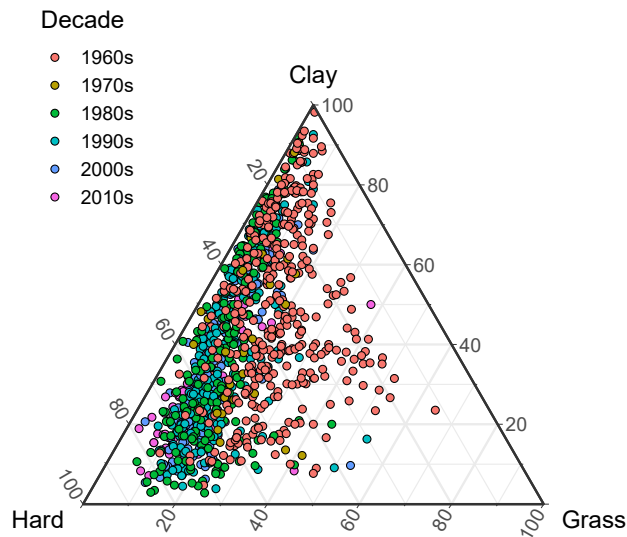


Figure 6: Ternary diagram of number of tournaments played by decades and their centers

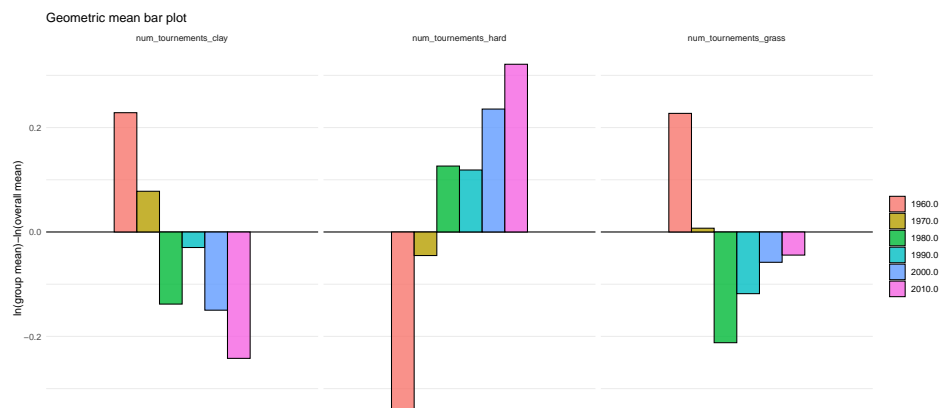


Figure 7: Geometric-mean bar plot comparing the centers of surfaces of tournaments by decades

Table 3: Distribution of Tournaments, Games, and Minutes Played across different court surfaces by decade: Compositional Center, Total Variance, and Sample Size. Compact letter display (CLD) indicates statistical significance groupings.

Surface	Number of tournaments					
	Decade					
	2010s	2000s	1990s	1980s	1970s	1960s
Clay	0.292	0.328	0.379	0.356	0.431	0.509
Hard	0.598	0.561	0.513	0.541	0.444	0.334
Grass	0.110	0.111	0.108	0.103	0.124	0.157
Variance	0.505	0.680	0.887	1.344	1.132	1.126
N	103	190	230	199	32	373
CLD	c	d	ab	a	a	b

Surface	Number of games					
	Decade					
	2010s	2000s	1990s	1980s	1970s	1960s
Clay	0.290	0.323	0.374	0.336	0.406	0.490
Hard	0.590	0.559	0.511	0.552	0.442	0.321
Grass	0.121	0.118	0.115	0.112	0.152	0.189
Variance	0.651	0.882	1.230	1.795	1.498	1.505
N	103	190	230	199	32	372
CLD	c	d	ab	a	a	b

Surface	Minutes played					
	Decade					
	2010s	2000s	1990s	1980s	1970s	1960s
Clay	0.286	0.330	0.373	0.345	0.207	0.562
Hard	0.594	0.555	0.515	0.557	0.627	0.320
Grass	0.120	0.115	0.111	0.098	0.166	0.118
Variance	0.714	0.956	1.343	1.843	1.173	1.023
N	101	189	229	162	8	16
CLD	a	ac	c	a	ac	b

surfaces through a CoDA approach, using a specific sample of professional tennis players who have achieved an ATP Top 100 ranking at least once in their careers. The analysis considered ranking categories, handedness, backhand style, and the decade of debut. The results provide insights into how surface preferences evolve based on player ranking, playing style, and temporal factors, revealing several important trends and interactions.

3.1. Playing surface and ranking categories

Professional tennis tournaments are played on various surfaces, each of which influences the match load differently. In this context, this study evaluated load through two perspectives: the number of tournaments played, which reflects a “self-regulated” load as players decide which tournaments to attend; and other load factors such as match duration, measured by the number of games and minutes played, which are inherently “external” and not initially determined by the players themselves. Clay courts have been associated with longer playing times, higher stroke counts, and greater distances covered compared to hard courts (Girard, Jean-Paul, and Millet (2010)). For this reason, tennis players adapt their match strategies and training routines depending on the surface (Fabre, Martin, Gondin, and Grelot (2012), Murias, Lanatta, Arcuri, and Laíño (2007)).

That clay and hard courts predominate in terms of the number of tournaments played is expected, given the limited availability of grass tournaments in the annual calendar, which are typically restricted to around four weeks. Moreover, professional tennis players often avoid changing court surfaces between tournaments, as such changes can alter movement patterns and playing styles (Néri-Fuchs, Sedeaud, Larochelambert, Marc, Toussaint, and Brocherie

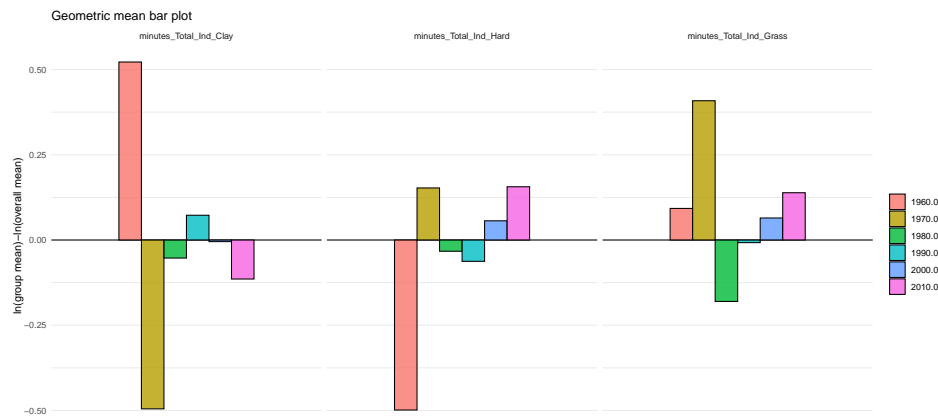


Figure 8: Geometric-mean bar plot comparing the centers of surfaces of minutes played by decades

(2023)).

The results of this study indicate a higher frequency of participation in tournaments on hard courts among players in the top-ranking category, who allocated 57.78% of their total minutes to hard courts, compared to 32.12% on clay and 10.10% on grass. This distribution aligns with the higher proportion of prestigious tournaments (e.g., Master 1000 and Open 500) held on hard courts over the last 10-20 years. Consequently, top players, who qualify for these events, tend to play more frequently on hard courts, while lower-ranked players exhibit a more balanced distribution across surfaces due to limited access to these tournaments.

3.2. Backhand style and playing surface

Professional players using a two-handed backhand play more tournaments, games, and minutes on hard surfaces compared to those with a one-handed backhand. However, this performance parameter is influenced by temporal trends. In recent decades, the majority of players have adopted the two-handed backhand, whereas 30 years ago the one-handed backhand was more prevalent (Genevois, Reid, Rogowski, and Crespo (2015)). This shift coincides with an increase in the number of tournaments played on hard surfaces compared to grass and clay. The results of this study do not show a direct effect of playing surface on backhand style; instead, the observed differences likely reflect broader trends in the evolution of playing styles and surface availability.

For instance, in 2013, 78% of ATP players and 96% of WTA players used a two-handed backhand (Eng and Hagler (2014)), and by the end of 2024, 9 of the top 10 players will use this style, according to the official ATP Tour singles ranking website. By contrast, at the end of 1990, only 2 of the top 10 players employed a two-handed backhand. This change can be attributed, in part, to the increasing physical demands of modern tennis, which favor the greater stability and power provided by the two-handed backhand. Hard surfaces, which now dominate the professional calendar, further amplify these demands, requiring players to adapt their techniques to the faster pace and reduced reaction times. This evolution highlights how shifts in tournament conditions and playing styles are interconnected, shaping player performance and strategies over time.

3.3. Decade of debut and playing surface

Analysis of the decade of debut revealed that players who started their careers in the 2000s and 2010s predominantly participated on hard surfaces, while players from earlier decades (1990s, 1980s, and 1970s) exhibited a more balanced distribution between hard and clay courts. The shift in surface preference over time can likely be attributed to the changing landscape of

professional tennis, where many players in the past did not train on grass courts due to the limited number of grass tournaments. However, in recent years, players have become more versatile, adapting to all surfaces with the increasing availability of tournaments on hard courts and a more globalized competitive circuit.

In the past, players were often considered "specialists" on specific surfaces, with many focusing primarily on either grass or clay. Nowadays, the trend is for players to be proficient on most surfaces, with the ATP's efforts to standardize surface speeds contributing to this shift. The ATP has worked to homogenize the speed of surfaces by assigning slower balls (type 1) to fast surfaces like grass, while faster balls (type 3) are used on slower surfaces such as clay, with medium-speed balls (type 2) designated for moderate surfaces [ITF \(2020\)](#).

It is also important to note that, historically, many top players were specialists in particular playing styles associated with specific surfaces. For example, grass and hard court players such as Boris Becker, Stefan Edberg, and Pete Sampras were known for their effective volleying, while players like Mats Wilander, Sergi Bruguera, Jim Courier, and Michael Chang excelled on clay with their strong, consistent groundstrokes and patient, physical style of play. In contrast, modern clay court players now display a more aggressive playing style, hitting faster shots from all areas of the court ([Jaramillo \(2012\)](#)). Today, many clay court specialists are versatile, performing well on both clay and hard surfaces.

The most notable trend over recent decades is the decreasing number of clay court tournaments played by top 100 players, alongside an increase in participation on hard courts. In contrast, grass court participation has remained relatively stable. This shift reflects the broader trend of surface homogenization, which emphasizes the increasing adaptability and versatility required from professional players.

3.4. CoDA and tennis analytics

CoDA is particularly well-suited for sports analytics, and its potential for addressing the relative nature of data should be further explored in this field. This is especially crucial when analyzing factors such as surface distribution in tennis. Traditional statistical methods often treat data as independent variables, whereas CoDA accounts for the interconnectedness of data components, making it an ideal tool for examining the proportions of time spent on different surfaces. This approach allows for a more accurate representation of player behavior and performance patterns, offering valuable insights for coaches, analysts, and players themselves. Given its capacity to handle compositional data, CoDA presents a promising avenue for resolving complex problems in sports analytics, including those related to surface preferences and player performance in tennis. Exploring its application in tennis analytics could provide significant advancements in understanding and optimizing player strategies and training.

3.5. Conclusions

In conclusion, surface preferences, along with player characteristics such as ranking, hand dominance, and backhand style, play a crucial role in tournament participation and match duration. Top 100 ATP players consistently favor hard courts, while lower-ranked players display more variety in their surface preferences. This distribution also reflects the evolution of the ATP circuit, where versatility has become increasingly important for remaining competitive across different surfaces.

Acknowledgments

This research was funded by the Agency for Management of University and Research Grants (Generalitat de Catalunya - Spain), grants number 2021SGR01197 and 2021SGR01421.

Additionally, this work has been supported by the Ministerio de Ciencia, Innovación y Uni-

versidades (Spain), grants number PID2021-123833OB-I00 and PID2019-104830RB-I00.

Acknowledgements are due to Joan Martínez for his work on data.

References

- Aitchison J (1982). “The Statistical Analysis of Compositional Data.” *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**, 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x.
- Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). ISBN 0-412-28060-4. 416 p.
- Aitchison J, Ng K (2005). “The Role of Perturbation in Compositional Data Analysis.” *Statistical Modelling*, **5**. doi:10.1191/1471082X05st091oa.
- Alamar BC (2013). *Sports Analytics: A Guide for Coaches, Managers and Other Decision Makers*. Columbia University Press, New York. ISBN 978-0-231-16292-0. doi:10.1016/j.smr.2013.06.005.
- Albert J (2010). “Sabermetrics: The Past, the Present, and the Future.” In JA Gallian (ed.), *Mathematics and Sports*. The Mathematical Association of America, Washington, D.C. ISBN 9780883853498. doi:10.5948/UP09781614442004.002.
- Albert J, Glickman ME, Swartz TB, Koning RH (eds.) (2016). *Handbook of Statistical Methods and Analyses in Sports, 1st edition*. Chapman and Hall/CRC. doi:doi.org/10.1201/9781315166070.
- Albert J, Koning RH (eds.) (2007). *Statistical Thinking in Sports (1st ed.)*. Chapman and Hall/CRC. doi:doi.org/10.1201/9781584888697.
- Baiget E, Corbi F, López J (2022). “Influence of Anthropometric, Ball Impact and Landing Location Parameters on Serve Velocity in Elite Tennis Competition.” *Biology of Sport*, **40**, 273–281. doi:10.5114/biolsport.2023.112095.
- Baumer B, Matthews G, Nguyen Q (2023). “Big Ideas in Sports Analytics and Statistical Tools for Their Investigation.” *WIREs Computational Statistics*, **15**. doi:10.1002/wics.1612.
- Casals M, Finch C (2018). “Sports Biostatistician: A Critical Member of All Sports Science and Medicine Teams for Injury Prevention.” *British Journal of Sports Medicine*, **52**, 1457–1461. doi:10.1136/bjsports-2016-042211rep.
- Chastin S, Palarea-Albaladejo J, Dontje M, Skelton D (2015). “Combined Effects of Time Spent in Physical Activity, Sedentary Behaviors and Sleep on Obesity and Cardio-Metabolic Health Markers: A Novel Compositional Data Analysis Approach.” *PLoS ONE*, **10**, 1–37. doi:10.1371/journal.pone.0139984.
- Comas-Cufí M, Thió-Henestrosa S (2011). “CoDaPack 2.0: A Stand-alone, Multi-platform Compositional Software.” In JJ Egozcue, R Tolosana-Delgado, MI Ortego (eds.), *CoDa-Work’11: 4th International Workshop on Compositional Data Analysis*. Sant Feliu de Guíxols. ISBN 978-84-87867-76-7.
- Crespo M, Martínez-Gallego R, Filipcic A (2024). “Determining the Tactical and Technical Level of Competitive Tennis Players Using a Competency Model: A Systematic Review.” *Frontiers in Sports and Active Living*, **6**. doi:10.3389/fspor.2024.1406846.

- Dominicy Y, Ley C (2023). *Statistics Meets Sports: What We Can Learn from Sports Data*. Cambridge Scholars Publishing. ISBN 9781527592742. URL <https://books.google.es/books?id=JVKqEAAQBAJ>.
- Dumuid D, Martín-Fernández J, Ellul S, Kenett R, Wake M, Simm P, Baur L, Olds T (2020). “Analysing Body Composition as Compositional Data: An Exploration of the Relationship between Body Composition, Body Mass and Bone Strength.” *Statistical Methods in Medical Research*, **30**, 962280220955221. doi:10.1177/0962280220955221.
- Dumuid D, Wake M, Clifford S, Burgner D, Carlin J, Mensah F, Frayssé F, Lycett K, Baur L, Olds T (2019). “The Association of the Body Composition of Children with 24-Hour Activity Composition.” *The Journal of Pediatrics*, **208**. doi:10.1016/j.jpeds.2018.12.030.
- Egozcue JJ, Pawlowsky-Glahn V, Figueras G, Vidal C (2003). “Isometric Logratio Transformations for Compositional Data Analysis.” *Mathematical Geology*, **35**, 279–300. doi:10.1023/A:1023818214614.
- Eng D, Hagler D (2014). “A Novel Analysis of Grip Variations on the Two-handed Backhand.” *Coaching & Sport Science Review*.
- Fabre JB, Martin V, Gondin J, Grelot L (2012). “Effect of Playing Surface Properties on Neuromuscular Fatigue in Tennis.” *Medicine and Science in Sports and Exercise*, **44**, 2182–9. doi:10.1249/MSS.0b013e3182618cf9.
- Fernandez-Fernandez J, Kinner V, Ferrauti A (2010). “The Physiological Demands of Hitting and Running in Tennis on Different Surfaces.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association*, **24**, 3255–64. doi:10.1519/JSC.0b013e3181e8745f.
- Gao Z, Sun W (2024). “Insights into the Tennis Court through Machine Learning: Analysis and Evaluation of the Wimbledon Men’s Singles Final.” *Highlights in Science, Engineering and Technology*, **98**, 507–514. doi:10.54097/001kgz04.
- Genevois C, Reid M, Rogowski I, Crespo M (2015). “Performance Factors Related to the Different Tennis Backhand Groundstrokes: A Review.” *Journal of Sport Sciences and Medicine*, **14**, 194–202.
- Giles B, Peeling P, Kovalchik S, Reid M (2021). “Differentiating Movement Styles in Professional Tennis: A Machine Learning and Hierarchical Clustering Approach: Identifying COD Profiles in Professional Tennis.” *European Journal of Sport Science*, **23**, 1–20. doi:10.1080/17461391.2021.2006800.
- Girard O, Jean-Paul M, Millet G (2010). “Effects of the Playing Surface on Plantar Pressures during the First Serve in Tennis.” *International Journal of Sports Physiology and Performance*, **5**, 384–93. doi:10.1123/ijsp.5.3.384.
- Gloor G, Macklaim J, Pawlowsky-Glahn V, Egozcue JJ (2017). “Microbiome Datasets Are Compositional: And This Is Not Optional.” *Frontiers in Microbiology*, **8**, 2224. doi:10.3389/fmicb.2017.02224.
- Hamilton NE, Ferry M (2018). “ggtern: Ternary Diagrams Using ggplot2.” *Journal of Statistical Software, Code Snippets*, **87**(3), 1–17. doi:10.18637/jss.v087.c03.
- ITF (2020). *Approved Tennis Balls, Classified Surfaces and Recognised Courts: A Guide to Products and Test Methods*. International Tennis Federation. Retrieved from: <https://www.itftennis.com/media/2714/2020-itf-approved-tennis-balls-classified-court-surfaces-and-recognised-courts.pdf>.

- Jaramillo G (2012). “How to Train Aggressive Clay Court Strategy and Tactics.” *ITF Coaching & Sport Science Review*, **20**(56), 4–7. doi:10.52383/itfcoaching.v20i56.398. URL <https://itfcoachingreview.com/index.php/journal/article/view/398>.
- Kovalchik S (2021). “Why Tennis Is Still Not Ready to Play Moneyball.” *Harvard Data Science Review*. doi:10.1162/99608f92.b665c0f4.
- Leitner C, Zeileis A, Hornik K (2009). “Is Federer Stronger in a Tournament Without Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009.” *Austrian Journal of Statistics*, **38**(4), 277–286. doi:10.17713/ajs.v38i4.280.
- Martín-Fernández J, Pawlowsky-Glahn V, Egozcue JJ, Tolosona-Delgado R (2017). “Advances in Principal Balances for Compositional Data.” *Mathematical Geosciences*, **50**. doi:10.1007/s11004-017-9712-z.
- Martín-Fernández JA, Daunis-i Estadella P, Mateu-Figueras G (2015). “On the Interpretation of Differences between Groups for Compositional Data.” *SORT*, **39**, 231–252.
- Mateu-Figueras G and Pawlowsky-Glahn V, Egozcue JJ (2013). “The Normal Distribution in Some Constrained Simple Spaces.” *Statistics and Operations Research Transactions (SORT)*, **37**, 29–56.
- Miller S (2006). “Modern Tennis Rackets, Balls, and Surfaces.” *British Journal of Sports Medicine*, **40**, 401–5. doi:10.1136/bjsm.2005.023283.
- Murias J, Lanatta D, Arcuri C, Laíño F (2007). “Metabolic and Functional Responses Playing Tennis on Different Surfaces.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association*, **21**, 112–7. doi:10.1519/R-19065.1.
- Navarro Lopez C, Gonzalez-Morcillo S, Forteza C, Linares-Mustarós S (2021). “A Bibliometric Analysis of the 35th Anniversary of the Paper “The Statistical Analysis of Compositional Data” by John Aitchison (1982).” *Austrian Journal of Statistics*, **50**, 38–55. doi:10.17713/ajs.v50i2.1066.
- Néri-Fuchs JB, Sedeaud A, Larochelambert Q, Marc A, Toussaint JF, Brocherie F (2023). “Medical Withdrawals in Elite Tennis in Reference to Playing Standards, Court Surfaces and Genders.” *Journal of Science and Medicine in Sport*, **26**. doi:10.1016/j.jsams.2023.04.002.
- Oliver L, Baiget E, Cortés J, Martínez J, Crespo M, Casals M (2024). “Retirements of Professional Tennis Players in ATP and WTA Tour Events.” *European Journal of Sport Science*, **24**, 1526–1536. doi:10.1002/ejsc.12177.
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons. ISBN 9781118443064. doi:10.1002/9781119003144.
- Pearson K (1896). “On a Form of Spurious Correlation which May Arise when Indices Are Used in the Measurement of Organs.” *Proceedings of The Royal Society of London*, **60**, 489–498. doi:10.1098/rsp1.1896.0076.
- Pluim B, Jansen M, Williamson S, Berry C, Camporesi S, Fagher K, Heron N, Janse van Rensburg DC, Perez V, Murray A, O’Connor S, de Oliveira F, Reid M, Reijen M, Saueressig T, Schoonmade L, Thornton J, Webborn N, Ardern C (2023). “Physical Demands of Tennis Across the Different Court Surfaces, Performance Levels and Sexes: A Systematic Review with Meta-analysis.” *Sports Medicine (Auckland, N.Z.)*, **53**. doi:10.1007/s40279-022-01807-8.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

- Sainani K, Borg D, Caldwell A, Butson M, Tenan M, Vickers A, Vigotsky A, Warmenhoven J, Nguyen R, Lohse K, Knight E, Bargary (née Coffey) N (2020). “Call to Increase Statistical Collaboration in Sports Science, Sport and Exercise Medicine and Sports Physiotherapy.” *British Journal of Sports Medicine*, **55**, bjsports-2020. doi:10.1136/bjsports-2020-102607.
- Sampaio T, Oliveira J, Marinho D, Neiva H, Morais J (2024). “Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review.” *Applied Sciences*, **14**. doi:10.3390/app14135517.
- Verswijveren S, Lamb K, Martín-Fernández J, Winkler E, Leech R, Timperio A, Salmon J, Daly R, Cerin E, Telford R, Telford R, Olive L, Ridgers N (2021). “Using Compositional Data Analysis to Explore Accumulation of Sedentary Behavior, Physical Activity and Youth Health.” *Journal of Sport and Health Science*, **11**. doi:10.1016/j.jshs.2021.03.004.
- Whiteside D, Cant O, Connolly M, Reid M (2017). “Monitoring Hitting Load in Tennis Using Inertial Sensors and Machine Learning.” *International Journal of Sports Physiology and Performance*, **12**, 1–20. doi:10.1123/ijsp.2016-0683.
- Wong F, Keung J, Lau N, Ng D, Chung J, Chow D (2014). “Effects of Body Mass Index and Full Body Kinematics on Tennis Serve Speed.” *Journal of Human Kinetics*, **40**, 21–8. doi:10.2478/hukin-2014-0003.
- Zhou Y, Zong S, Cao R, Ruano M, Chen C, Cui Y (2023). “Using Network Science to Analyze Tennis Stroke Patterns.” *Chaos Solitons & Fractals*, **170**, 113305. doi:10.1016/j.chaos.2023.113305.

Affiliation:

Pepus Daunis-i-Estadella
Dept. Computer Science, Applied Mathematics, and Statistics
University of Girona
E-17003 Girona, Catalonia, Spain
E-mail: pepus@imae.udg.edu
URL: <http://www.udg.edu/personal/pepus-daunis-i-estadella>

Ernest Baiget
National Institute of Physical Education of Catalonia (INEFC)
University of Barcelona
E-08038 Barcelona, Catalonia, Spain
E-mail: ebaiget@gencat.cat
URL: https://inefc.gencat.cat/ca/detalls/article/baiget_vidal

Martí Casals
National Institute of Physical Education of Catalonia (INEFC)
University of Barcelona
Sport and Physical Activity Studies Centre (CEEAF), Faculty of Medicine
University of Vic-Central University of Catalonia
E-08038 Barcelona, Catalonia, Spain
E-mail: marticasals@gencat.cat
URL: https://inefc.gencat.cat/ca/inefc_barcelona/coneix_inefc/professorat/detalls/casals-toquero