



“Hellenic Open University”

“Postgraduate Course”

Postgraduate Dissertation

“The Impact of Data Analytics on Performance and Strategy in
Professional Tennis: A Business Information Systems Perspective”

“Stylianos Kourogiorgas”

Supervisor: “Athanasios Rentizelas”

Patras, Greece, “Month” “Year”

Theses / Dissertations remain the intellectual property of students (“authors/creators”), but in the context of open access policy they grant to the HOU a non-exclusive license to use the right of reproduction, customisation, public lending, presentation to an audience and digital dissemination thereof internationally, in electronic form and by any means for teaching and research purposes, for no fee and throughout the duration of intellectual property rights. Free access to the full text for studying and reading does not in any way mean that the author/creator shall allocate his/her intellectual property rights, nor shall he/she allow the reproduction, republication, copy, storage, sale, commercial use, transmission, distribution, publication, execution, downloading, uploading, translating, modifying in any way, of any part or summary of the dissertation, without the explicit prior written consent of the author/creator. Creators retain all their moral and property rights.



**“The Impact of Data Analytics on Performance and Strategy in
Professional Tennis: A Business Information Systems Perspective”**

“Stylianos Kourogiorgas”

Supervising Committee

Supervisor:

“Athanasios Rentizelas”

“Affiliation”

Co-Supervisor:

“Dimitrios Fouskakis”

“Affiliation”

Patras, Greece, “Month” “Year”

“Acknowledgments and / or Dedication”

Abstract

Professional tennis is a constantly increasing data-informed environment, where match statistics and KPIs support the decisions that were once primarily guided by experience and intuition. This dissertation investigates how the growing availability and use of data has influenced the performance and the strategic patterns in men’s professional tennis. In this view, analytics is treated as an information capability - it structures evidence, reduces uncertainty, and supports more consistent decision-making.

The study uses historical match data from the Jeff Sackmann ATP dataset, focusing on seasons with most broadly reliable match statistics. The central design choice is an era comparison between 1991–2005 and 2006–present. The year 2006 is used as a practical turning point because it aligns with the wider introduction of tracking and review technologies (e.g., Hawk-Eye), which increased the enhancement and credibility of tennis data and encouraged analytics-driven approaches at the professional level. While the dissertation is not measuring “analytics adoption” directly inside teams, it tests if the observable outputs of the game performance indicators and match characteristics shift meaningfully after this change in the sport’s data environment.

The analysis is organized into two parts. First, descriptive statistics make a summary on how key performance indicators evolve between eras and contexts (surface, tournament level, match format). Core measures include serve effectiveness (ace and double-fault rates), pressure-point performance (break-point save and conversion rates), return effectiveness, and match structure (tiebreak frequency). Distributional patterns are examined to better understand the behavior of tennis events and to support suitable inference.

Second, hypothesis testing (including t-tests, welch-test) evaluates whether the differences between eras are statistically significant and practically meaningful, reporting effect sizes alongside p-values.

Overall, the expected contribution is an evidence-based assessment of whether and how is the “data era” corresponding to measurable changes in performance and strategies in tennis.

Keywords

Tennis, Data Analysis, Hypothesis Testing, Descriptive Statistics

“Η Επίδραση της Ανάλυσης Δεδομένων στην Απόδοση και
Στρατηγική στο Επαγγελματικό Τένις: Μια Προσέγγιση από την
Οπτική των Πληροφοριακών Συστημάτων Επιχειρήσεων”
(in Greek)

“Στυλιανός Κουρόγιωργας”
(in Greek)

Περίληψη

Το επαγγελματικό τένις εξελίσσεται διαρκώς σε ένα όλο και πιο data-informed περιβάλλον, όπου τα στατιστικά αγώνων και βασικοί δείκτες απόδοσης υποστηρίζουν αποφάσεις που παλαιότερα βασίζονταν κυρίως στην εμπειρία και τη διαίσθηση. Η συγκεκριμένη διατριβή ερευνά το πώς η όλο και περισσότερη αυξανόμενη διαθεσιμότητα και αξιοποίηση δεδομένων έχει επηρεάσει την απόδοση και τα στρατηγικά πρότυπα στο επαγγελματικό ανδρικό τένις. Στο πλαίσιο αυτό, η ανάλυση δεδομένων δομεί την τεκμηρίωση, μειώνει την αβεβαιότητα και υποστηρίζει συνεπέστερη λήψη αποφάσεων. Η μελέτη αξιοποιεί ιστορικά δεδομένα αγώνων από το σύνολο δεδομένων Jeff Sackmann ATP, με έμφαση στις περιόδους όπου τα διαθέσιμα στατιστικά αγώνων θεωρούνται έγκυρα και πιο αξιόπιστα. Ο σχεδιασμός έχει γίνει συγκρίνοντας δύο εποχές: 1991–2005 και 2006–σήμερα. Το έτος 2006 επιλέγεται ως πρακτικό σημείο καμπής, καθώς συμπίπτει με την εισαγωγή τεχνολογιών παρακολούθησης και επανεξέτασης (π.χ. Hawk-Eye), οι οποίες ενίσχυσαν την αξιοπιστία των δεδομένων στο τένις και συνέβαλαν στην υιοθέτηση πιο σωστών προσεγγίσεων στο επαγγελματικό επίπεδο. Παρότι η διατριβή δεν μετρά άμεσα την «υιοθέτηση analytics» εντός των ομάδων, εξετάζει εάν τα παρατηρήσιμα αποτελέσματα του παιχνιδιού οι δείκτες απόδοσης και τα χαρακτηριστικά των αγώνων μεταβάλλονται ουσιαστικά μετά από αυτή τη μεταβολή στο περιβάλλον δεδομένων του αθλήματος.

Η ανάλυση οργανώνεται σε δύο μέρη. Πρώτον, μέσω περιγραφικής στατιστικής συνοψίζεται ο τρόπος με τον οποίο εξελίσσονται βασικοί δείκτες απόδοσης μεταξύ εποχών και (επιφάνεια γηπέδου, επίπεδο διοργάνωσης, μορφή αγώνα). Οι βασικές μετρήσεις περιλαμβάνουν την αποτελεσματικότητα του σερβίς (ρυθμοί άσων και διπλών λαθών, επιτυχία πρώτου/δεύτερου σερβίς), την απόδοση σε κρίσιμους πόντους (ποσοστά σωσίματος και μετατροπής break points), την αποτελεσματικότητα στην επιστροφή, καθώς και χαρακτηριστικά της δομής του αγώνα (συχνότητα tie-break). Επιπλέον, διερευνώνται κατανομές και μοτίβα των δεδομένων, ώστε να κατανοηθεί καλύτερα η συμπεριφορά των γεγονότων στο τένις και να τεκμηριωθούν κατάλληλες επιλογές στατιστικής συμπερασματολογίας.

Δεύτερον, μέσω ελέγχων υποθέσεων (συμπεριλαμβανομένων t-tests και Welch-tests) αξιολογείται κατά πόσο οι διαφορές μεταξύ των δύο εποχών είναι στατιστικά σημαντικές και πρακτικά ουσιαστικές, με αναφορά μεγεθών επίδρασης (effect sizes) παράλληλα με τις τιμές p.

Συνολικά, η συμβολή της διατριβής είναι μια τεκμηριωμένη και βασισμένη σε δεδομένα αποτίμηση ως προς το πόσο και με ποιον τρόπο η «εποχή των δεδομένων» αντιστοιχεί σε σημαντικές και μετρήσιμες αλλαγές στην απόδοση και στις στρατηγικές στο επαγγελματικό τένις.

Λέξεις – Κλειδιά

Τένις, Ανάλυση Δεδομένων, Έλεγχοι Υποθέσεων, Περιγραφική Στατιστική.

Table of Contents

Abstract	2
Περίληψη.....	3
Table of Contents	5
List of Tables.....	7
List of Abbreviations & Acronyms.....	8
1. Introduction.....	10
1.1 Context	10
1.2 Aim and Objectives of the Thesis	11
1.3 Contribution of the Thesis.....	12
1.4 Structure of the Thesis	13
2. Literature.....	14
2.1 Sports Analytics in General	14
2.1.1 Football	14
2.1.2 Basketball (NBA).....	16
2.1.3 How Analytics Transformed Sprinting and Ice Skating	18
2.2 Tennis Analytics.....	24
2.2.1 From Dispute to Data: How Electronic Officiating Anchored Tennis Analytics	24
2.2.2 Shot Value and Modern Matchcraft: Rewriting Tactics with Tracking and Models.....	25
2.2.3 Quantifying the Training Week: Workload, Periodization, and Wearables	28
2.2.4 Engineering the Serve: Biomechanics and Technique Transformation.....	29
2.2.5 Turning Surveillance into Prevention: Injury Epidemiology and Risk Control	31
2.2.6 Systems Thinking for Tennis Analytics: A Business Information Systems View	32
3. Methodology.....	33
3.1 Research Design and Purpose	33
3.2 Data Source and Dataset Description.....	34
3.3 Surface Context.....	35
3.4 Variable Construction and Key Performance Indicators (KPIs).....	36
3.5 Descriptive Statistics and Visualization Plan.....	37
3.6 Hypothesis Testing Strategy	38
3.7 Software, Workflow, and Reproducibility	39
3.8 Methodological Limitations (within this design).....	39
4. Results.....	41
4.1 Descriptive Statistics.....	41
4.2 Hypothesis testing	49
4.3 Discussion	52
5. Conclusion.....	53
References.....	55

List of Tables

Table 1: Analytics used per sport.....	23
Table 2: First-serve zones	26
Table 3: Pre 2026 data distribution	41
Table 4: Post 2026 data distribution.....	42
Table 5: Ace rate boxplots per surface per era.....	42
Table 6: Ace rate histogram per era	43
Table 7: Double fault rate boxplots per surface per era	44
Table 8: Double fault rate histogram per era	44
Table 9: First serve in rate boxplots per surface per era	45
Table 10: First serve in rate histogram per era.....	45
Table 11: First serve points won rate boxplots per surface per era.....	46
Table 12: First serve points won rate histogram per era	46
Table 13: Second serve points won rate boxplots per surface per era	47
Table 14: Second serve points won rate histogram per era.....	47
Table 15: Hold rate boxplots per surface per era	48
Table 16: Hold rate histogram per era.....	48
Table 17: Tie break time series	49
Table 18: Ace Rate test results.....	50
Table 19: Double fault test results	50
Table 20: Hold rate test results.....	50
Table 21: First serve in rate test results.....	51
Table 22: First serve points won rate test results	51
Table 23: Second serve points won rate test results.....	51

List of Abbreviations & Acronyms

ML: Machine Learning—computer models that learn patterns from data to make predictions.

LR: Logistic Regression—a statistical ML classifier for yes/no or multi-class outcomes.

RF: Random Forest—an ML ensemble of decision trees; strong for classification tasks.

SVM: Support Vector Machine—an ML classifier that finds the best boundary between classes.

KNN: K-Nearest Neighbors—an ML method that predicts using the most similar past cases.

xG: Expected Goals—the probability that a shot becomes a goal based on its context (location, angle, pressure, body part, etc.).

xGA: Expected Goals Against—the opponent’s total chance quality conceded by a team.

xGC/xGCA: Expected Goals Chain/Against—how possession sequences (“chains”) generate or concede chance quality.

PPDA: Passes Per Defensive Action—how many opponent passes are allowed before a defensive intervention in pressing zones; lower PPDA = more aggressive pressing.

GPS: Global Positioning System—wearable tracking of distance, speed, accelerations during matches.

EPV: Expected Possession Value—the expected goal value of the possession as it evolves, not just shot events.

EV: Expected Value—average expected points/goals for a decision (e.g., the true value of a shot including fouls).

GAM: Generalized Additive Model—a flexible statistical model that captures curved/nonlinear effects.

ROI: Return on Investment—wins or points gained relative to costs (salaries, staff, tax).

EV: (Expected Value) Average points you expect from a decision (e.g., a type of shot) once you account for all outcomes, including free throws drawn.

True-EV: (True Expected Value) Shot value that includes free-throw scoring attributable to the shot decision (e.g., fouled drives, fouled jumpers). This corrects early shot-chart models that ignored FT yield.

FT: Free Throws

ITF: International Tennis Federation

SPR: Stroke Performance Relevance

IMU: Inertial Measurement Unit

ITN: International Tennis Number

w_ace - winner's number of aces

w_df- winner's number of doubles faults

w_svpt - winner's number of serve points

w_1stIn- winner's number of first serves made

w_1stWon - winner's number of first-serve points won

w_2ndWon - winner's number of second-serve points won

w_SvGms - winner's number of serve games

w_bpSaved - winner's number of break points saved

w_bpFaced - winner's number of break points faced

l_ace

l_df

l_svpt

l_1stIn

l_1stWon

l_2ndWon

l_SvGms

l_bpSaved

l_bpFaced

1. Introduction

1.1 Context

The change that started about 20 years ago saw a change from a game based primarily on intuition and instinct and with a strong focus on subjectivity, to a data-driven game. Every type of sport is now at least partially incorporating performance statistics and measurements in an attempt to quantify at least part or all of the performances of the players. This is for the purpose of measuring the players and to use those measures to test and develop strategies, to work out workloads and to reduce the risk of errors.

In the same manner as the case above, in tennis, performance of the players, the games and the tactics employed by them are not only described and used for the purposes of gaining insights but they are actually evaluated for performance; the measures that are used in the evaluation form the basis of strategy, techniques and training. In the era before analytical statistics, they were mostly used for descriptions. In such case, the player was qualitatively evaluated by the coach in terms of his form, technique, fitness, and preparedness for the game. Strategy was also formulated and used based on the observation of performance, while the analysis would have been performed internally in the minds of players and coaches, rather than on the basis of any quantitative data. This, of course, was a useful way, but could not quantify the change in the performance and thus could not be compared to others or to determine which of these performance measures were more likely to lead to successes.

This is, then, not a story only of tennis, but also a chance to answer questions regarding more generic information systems challenges. For tennis, statistics and data are strategically important because they serve to establish the data-driven decision making, reduce the subjectivity and bias in decision making and turn subjective opinions into actionable information. For this reason, tennis has been chosen as an illustration of the effect of the information, how has the information collection, organization and usage evolved in sports sciences as well as more generally in business.

1.2 Aim and Objectives of the Thesis

Thus the present thesis aims to answer whether any major change has taken place in men's professional tennis game due to the "data era" and whether there are any significant difference between the two periods, pre and post data eras (using a large set of data from past ATP matches) and if the difference is statistically significant (using descriptive and inferential analysis). The objective of this research, therefore, is to establish whether there is any significant difference between the two eras and if the difference is due to a chance event. Specifically, performance before 2006 will be compared with the performance from 2006 onwards, with the year 2006 being chosen as a cutoff point for the starting of the data era as this seems to be a reasonable and realistic date.

The research objectives are:

First, use descriptive statistics and visualization to display any observed performance changes between the two time periods. In particular, the focus will be on the performance of serve, return (in other words, a measure of pressure on the service), and other performance measures that include the effect of other factors such as structural characteristics of tournaments (e.g. the occurrence of tie-breaks).

Second, a hypothesis test will be performed to determine whether any observed performance difference between the two time periods is statistically significant or due to a random event using the appropriate statistical method for both discrete and continuous variables.

Finally, differences in performance between playing surfaces (hard court, clay and grass) will be investigated based on the assumption that playing surface, and so performance profiles, vary with the type of playing surface.

Next, I review the implications of these results, and interpret them in light of the field of Business Information Systems, where the field of analytics is considered not as an independent area but as one particular capability of the broader business decision making arena.

1.3 Contribution of the Thesis

This dissertation offers both empirical and conceptual contributions to the literature. Empirically, the dissertation contributes to the literature by comparing performance measures in professional tennis across two eras within the same study using free to the public, publicly available ATP data. There is not much published work comparing the two eras using multiple measures, and the majority of work in sports analytics in the game of tennis has focused on individual case studies, on the impact of playing strategy, or on technology. The work presented in this dissertation is the first to compare multiple performance indicators across two eras, and includes an analysis of surface specific performance.

This dissertation provides a conceptual contribution. In addition to the empirical contributions, the work in this dissertation relates the literature on sports analytics and the game of tennis to the broader Business Information Systems literature, and does not seek to position the analysis and work as simply another analytics or technical article, but rather as an analytics article, where analytics is a particular capability of data, information, and the resulting knowledge, for planning, evaluation, and decision-making under uncertainty. The work in this dissertation offers one case study of how the changing business information systems landscape (i.e., the changing data infrastructure and practices in analytics) may change competitive and organizational outcomes in a particular arena. The work in this dissertation also offers a methodological contribution by laying out the performance measures included in the analysis, along with the overall analysis, and includes a fairly extensive description of the hypothesis testing approach, and the performance measures used in the tests. This methodological contribution offers an analytics article that is not only easily readable for the tennis fan, as well as for the non-fan reader, but offers a study that is replicable and assessable in ways that most published work on analytics is not.

1.4 Structure of the Thesis

The rest of the dissertation is organized as follows: the second chapter is the Literature Review, where I begin with a discussion of the growth of analytics in several sports, including football (soccer), the National Football League in the United States, basketball, the National Basketball Association in the United States, as well as track and field (sprinting), and the Olympic sport of speed skating. This literature review also includes an exhaustive review of sports analytics in tennis, which covers the fields of tracking and measurement technology, playing strategy and tactics, training loads, biomechanics, injury surveillance, and a more general analysis of Business Information Systems in tennis. In the third chapter of this dissertation I describe the methodology of this study, including the research design, the data source, the definition of a particular era in the context of the research, the variable construction for the study, the set of performance indicators chosen for analysis, and the overall analysis, including the approach to hypothesis testing and limitations. The fourth chapter of the dissertation presents the results, including the descriptive statistics and the results of the statistical analysis for the various hypotheses, both overall and separately by surface, in each era. The fifth chapter of the dissertation provides a discussion of the research results in the context of the review of the literature in Chapter 2, and places these results into the broader context of Business Information Systems, of sports more generally, and across the two eras of data in tennis. The sixth and final chapter is the Conclusion, which summarizes the results, as well as the research contributions and limitations, and includes a description of possible future research.

2. Literature

For years, sports were dictated by a coach's instinct, their history, and tradition. Coaches and players used to make decisions based on what they could see with their eyes, on anecdotes from games that they watched or on their personal beliefs. In tennis, a player's tactics used to be decided by their own knowledge of the sport, and coaches only had an idea of what they should do in training, and sometimes in matches. Coaches used to watch matches from the past opponents, yet the analysis that came out of it was incomplete and could be biased because of the lack of any data.

Measuring performance in a match was a lot more subjective than statistical. The vast majority of coaches would observe certain visual factors like how far a player can run, their physical structure, and how strong they are in pressure moments, rather than having hard numbers. Training sessions used to target overall physical attributes, with limited insight into the specific areas of a game that require improvement, such as shot timing, the efficiency of a serve, or the player's movements across the court. Player health was managed only by dealing with injuries, with almost no assessment of the training load on a player, nor with any study of their biomechanics.

Other sports also used past performance results to form their squad and tactics, not with analytical models based on these results. Scouting teams also used to rely on their own intuition and relationships and these were prone to many mistakes. And even the opinions of experts and fans were also a factor that dictated decisions in some cases.

Overall, due to the lack of data-driven analysis, it required more time and effort to reach any high level of achievement and it was not done so precisely and accurately. Athletes who were successful needed talent and a lot of experience with the subtleties of the sport, and achievements came by trial and error. While great players came into existence during this time, they did not know exactly what to do and when, as analytics provided. Numerous ways to achieve improvement are yet to be discovered.

2.1 Sports Analytics in General

2.1.1 Football

In football, analytics has replaced static formulas and gut feeling with contextual decisions dependent on game state, decisions that are transparent and consistently reproducible.

Substitutions, performance metrics (xG/xGA/xGC/xGCA and PPDA), and match-outcome prediction (EPV vs xG) all demonstrate this transformation.

Substitutions have become state-aware decisions

Managers have for decades made their substitution decisions according to a "60 minutes" rule, or by gut feeling. With the growth in datasets, analysts are turning substitutions into a classification problem. Mohandas, Ahsan, and Haider et al. [1] have taken over 51,000 events from 9,074 matches spanning 9 different leagues and used a series of machine learning algorithms to train on that data, namely a logistic regression, a random forest, a support vector machine and a k nearest neighbour algorithm, to find the timing and individual characteristics that maximise improvement in a game's outcome. The random forest model managed to correctly predict more than 70% of substitution situations, while the support vector machine had high precision values, implying that a model with variables about the state of play during the match (current score, player attributes, time in match, and game context) will be able to predict when to change players. This moves the change from a moment when change is expected to an informed decision that supports and adds to the manager's instincts and experience

From “work-rate” to chance quality and pressing efficiency

In performance analysis, distance and sprints have been popular metrics, yet they do not tell us what a team needs to do to win a game. In the same two seasons, Barthelemy and colleagues [2], analyzed 71 Ligue 2 matches, and found that in combination with the more sophisticated, technically- and tactically-derived statistics such as xG, xGA, xGC and xGCA, as well as PPDA (average number of passes allowed per defensive team action) and tracking-derived measures of high-speed running and accelerations/decelerations, they were able to predict the winning outcome more accurately than volume statistics. In particular, they found that teams were more likely to win when their xGA and xGCA were low, with most volume statistics having no statistical significance. This means that a better predictive measure than total distance is the quality of the chances that are either created or allowed, and the efficiency of a pressing game: a small shift in the way in which fitness is valued, but a bigger shift in terms of the choice of performance indicators.

Choosing the right KPI for forecasting vs auditing

At the start of a game, a manager will probably ask, "if we press here, and overload there, and push into this space, are we likely to win the ball?" In a study that used both match event data from the German Bundesliga and tracking data, Forcher [3] compared the shot-based xG to the event-based possession value (EPV). It shows that xG predicts the actual result better than does EPV (xG accuracy is around 0.656), while EPV better predicts pre-match outcomes compared to xG. In sum, EPV should be used before a match to understand how and if a team can win the ball back, and the subsequent outcome; xG should be used as a retrospective measure of a team's shooting performance. Separating these measures creates both analytical and match day preview measures that link directly into coaching work.

Putting it together (football)

At the start of a week, EPV-based game preview identifies how and where possession value will shift. The manager will plan his press trap strategies and his build-up strategies based on this assessment. On the day of the game, a change recommender using live xG, PPDA and GPS-based metrics indicating player fatigue will offer the best time for the substitution, and where the player should be positioned. Post-game analysis is grounded in xG, xGA, xGC, xGCA and PPDA. Instead of just looking at total distance, analysts identify those sequences and actions that resulted in xG for/against, and how effective the press was at stopping the opponent attacking chain. This ends-to-end approach allows for KPIs to be connected to decision-making moments, and helps address the disconnect between what the data reveals and what the coaching staff does.

2.1.2 Basketball (NBA)

The same applies to analytics progress in basketball. The main areas where we have seen improvement include the (1) re-definition of shooting in True Shooting percentage including free throws; (2) combining draft advice from different teams; (3) using the "money is the ultimate objective" mindset in both salary management and the hiring process for coaches and analysts alike.

Re-specifying shot selection: True-EV, not just 3-volume

Past shot charts have touted the three point line by showing its better points-per-shot ratio and recommending that teams attempt more threes, though they largely excluded free throws obtained on a shot decision (such as drives that draw fouls or jump shots that draw fouls). Ehrlich and Sanders [4] add the free throws made on the shot to the true expected value (EV) using game state analysis, fine-grained shot EV charts, and find there is a consistent decline in the value of the three point shot since the 2018–19 season. Their result suggests that teams are over shooting threes and not generating enough paint pressure or fouls. This work corrects an earlier, too simplified notion that three point shots are so valuable a portfolio of only three point shots is the optimal shot portfolio and that in reality portfolios of three point shots and mid shots and actions that draw free throws are superior when the true EV is calculated.

Drafting: consensus beats idiosyncrasy

Draft night is inherently noisy, a cacophony of draft experts’ mock draft boards, internal draft boards, and public consensus. Fisher and Montague propose rank-biased distance as a more sensible performance metric for evaluating a single ranked list and a rank choice aggregation (instant runoff voting) and find that consensus (aggregating many mock drafts into one list) consistently beats an individual mock draft in predicting player draft order in many years [5]. In practice, consensus should be used to improve draft pick order predictions and improve trading value predictions for draft pick movement (up or down) and is generally used to mitigate risk in the selection process via weighing evidence rather than individual author bias.

Wins-per-dollar: luxury tax and analytics staffing

The NBA is a soft cap. There are provisions for teams to exceed the cap, but only with an assessment on each dollar of overage. Do these additional dollars result in an increase in regular season games won? Pradhan and Leshchinskii (2011-12 – 2023-24 NBA seasons) utilize mixed effects models (in order to control for external determinants of regular season wins, such as roster talent distribution or market size) and find that, indeed, luxury tax paying teams are measurably winning more games than non-taxpaying teams [6]. The researchers conclude that “tax-paying teams have successfully monetized this policy for competitive gains, suggesting that tax payers should view the tax as an investment with

positive returns rather than a pure cost.” Wang, Sarker, and Hosoi take a similar approach in calculating the ROI of analytics staffs (2010-21 to 2018-19 NBA seasons) and find, using a two way fixed effects model and adjusting for confounds such as player salaries, roster injuries, schedule difficulty, team effects, and season effects, that 1 additional basketball operations analyst is worth one additional regular season win [7]. These analyses indicate that there are two means for a basketball operations department to increase the likelihood of victory in the regular season; roster spend and investment in analytics staff. The choice to pursue, depending on the team’s situation, should be that which leads to the most wins-per-dollar.

Putting it together (basketball)

On a weekly level, the basketball analyst produces an EV chart that shows how each player performs in their current lineup, including how many FTs they draw from the free throw line when the team executes certain actions or in certain situations. In turn, the EV chart highlights which lineup is underperforming on total shot EV (including free throws) while maintaining a high number of 3 point shooting. On the draft night level, the basketball analyst synthesizes all the public and private mock draft into a consensus, using a ranked choice voting algorithm. This consensus is then used to forecast which player is projected to go at each pick of the draft, as well as forecast trades on pick orders that will be executed throughout the draft. On a season level, as the basketball team makes decisions on whether it is worthwhile to spend more on roster building in order to reach the pay-or-play limit, or to use some of the budget for hiring new analysts, the analyst can calculate the marginal cost-benefit ratio using wins per dollar.

2.1.3 How Analytics Transformed Sprinting and Ice Skating

Prior to the analytics revolution, coaching for sprinters and ice skaters mostly relied on the trainer’s experience and intuition. Technique errors were communicated with vague cues, like “get your knees up higher” or “stay low in the turn”; training loads were based on intuition; injury mitigation was focused primarily on general conditioning and generic safety messaging. None of those tools told you what specifically to change about an athlete’s movement pattern (where forces were being generated, for instance); how large a change in that element needed to be to have a noticeable effect; nor did they make a strong case about what the causal or quantifiable, before-versus-after training impact was on performance or

injury. With analytics, instrumented measures of force/torque/center of pressure (COP)/training loads, with their accompanying systematic review and policy, have made technique modification, training prescription, and injury mitigation all more feasible, repeatable, and successful. Four studies, as follows, show the success of those approaches and give concrete findings for sprinting and ice-skating interventions.

Sprinting: Quantifying Technique, Targeting Torque, and Gaining Time

One study used a single-case experimental design tracking an elite sprinter over four competitive years. Nagahara and Yamashita (2025) investigated the specific kinetic and kinematic variables contributing to the sprinter’s increased maximum speed, using a series of force-plates and motion capture. The athlete’s 200-m best time was reduced by 1.17 seconds over that four-year period, an improvement large enough to likely be important in elite competition. When his sprint technique was analyzed, the performance improvement correlated with the athlete’s reduction of torque at the knee joints in the knee extension (the force the quadriceps muscle group produces to stabilize his knee, absorb the force of landing, and push off into a more extended phase of the sprint), during a given stride. There were also improvements in his maximum speed as the ankle plantar-flexion (the downward ankle pushing-off force) increased, which indicates the use of a torque redistribution strategy that can be specifically programmed during training rather than just using intuitive, or weight-training alone, approaches. The question in the pre-analytics period was the lack of clarity around the causal connection between movement pattern changes and sprinting speed; the solution in the post-analytic era is the quantification of the connection between specific torque profiles and speed improvement, with specific training to elements during the support phase (like the timing of foot-ground contact and ankle stiffness) now possible to further improve speed [8].

Injuries

Sprinting is another example of the clear success of an analytic-informed training method. In a review of 12 studies, Tedeschi and colleagues (2025) found a consistent record of athletes successfully incorporating specific amounts of high-speed running at 80-90% of max speed, eccentric hamstring work, and correction of key sprinting biomechanics, like pelvic positioning and hip extension. The review showed that those efforts resulted in reduction of hamstring injury rates anywhere from 56 to 94 percent. Why that was successful

is also discussed in that review, which cites significant changes in injury-related variables among athletes, like longer biceps femoris fascicle length (which rose ~20 percent over the course of a training cycle), or muscle strength improvement (15-20 percent), all of which can be tracked longitudinally by a coach. Before, coaches thought fast, high-speed running by speed athletes was mostly “harmful” to the hamstring muscles. Therefore, many coaching programs chose to avoid doing high speed sprinting very often; this limited speed improvement from time to time. However, now that analytics is more available, the tendency seems to have reversed. Coaches rely on workload exposure curves, muscle architecture and technique checkpoints in order to control the amount of high-speed running done to help them develop speed in athletes for better performance and to prevent injuries. Or, in other words, the usage of analytics enables coaches to train speed, not reduce the volume of sprinting or be extremely cautious about the process [9].

To summarize, in short these 2 sprinting papers show a closed loop: measure technique, prescribe a specific change (torques, angles, contact duration) to confirm increased speed, and measure anatomy/forces, prescribe eccentric exercise for the sprinter to confirm lowered risk of injury. This is what sprinting looks like after analytics are implemented.

Ice Skating: Mapping Push-Off Physics and Reducing Risk with Multi-Component Programs

To increase performance, van der Kruk et al. (2019) used skates fitted with sensors, which are known as force sensors, to look at pushing force (forces exerted sideways and downward on the skaters) and the center of pressure (COP) path followed by elite short-track speed skaters to establish technique indicators for fast skaters. Specifically, faster male skaters (based on faster personal best times) had a more rearward COP during critical turn-to-straight transitions ($r = 0.82$ to 0.86) and showed a tendency of creating more conservative sideways peak forces when entering a turn, and this finding agrees with a better directionality and directionality of force and less loss of force to the side. Previously, this difference in technique was described through phrases like “be balanced on the blade” and “don’t let the skating entry slide,” but without analytics, they were unable to be proven or prescribed for each individual skater. In the current world, coaches are able to identify and measure the application of force when the push occurs, visualize where the COP goes in a stroke, and tell which technique modifications (like when the edge changes and the control of sideways forces peaks) lead to performance improvements [10].

Injuries

Regarding the amount of injuries that occur in skating/snow skating sports and in competitive skating, Fan et al. (2024) conducted a systematic review and meta-analysis that consisted of 15 studies with a total combined sample size of 26,123 and determined that the injury rate in the group that received a data-based multi-component intervention was roughly 50% lower (pooled risk ratio (RR) = 0.50; 95% confidence interval (CI) 0.42–0.63). Significantly, changes in the regulations (e.g., change in laws and ice/track rules) proved the most effective at lowering risk (RR = ~0.28), whereas the other programs were education and skill training (RR = ~0.50), education videos (RR = ~0.53) and wearing of safety gear (RR = ~0.64) for the purpose of reducing risk. The problem here during the era before analytics is still the current one: skating facilities and teams give safety advice without a real idea of which skaters are getting injured, the kind of injuries they are incurring, the types of movements that are causing the problem, which safety guidelines are being abided by and the extent of adherence with the equipment that is prescribed. The current day analytics based approach is the multi-part approach (that consists of teaching about changing behavior, required protective equipment and regulations) that has been proven to effectively reduce injury and in turn increase the potential for skating practice in the max capacity and for competition [11].

To sum up, a table below shows for each sport, team or individual, what statistic methods were used:

Sport (Type)	Statistics / Analytics Used	Typical Data Inputs (as described)
Football / Soccer (Team sport)	Match-state substitution modeling using ML classifiers (Logistic Regression, Random Forest, SVM, KNN) to predict substitution timing/personnel; chance quality & defensive quality metrics: xG, xGA, xGC/xGCA; pressing intensity: PPDA; forecasting	Event + match context features (scoreline, time, performance indicators), plus technical-tactical metrics and physical/tracking-derived measures (e.g., accelerations/decelerations, high-speed running mentioned alongside xG/PPDA)

	vs auditing metrics: EPV compared to xG for prediction vs post-match review	
Basketball (NBA) (Team sport)	Shot selection analytics using player tracking + shot charts; True Expected Value (TrueEV) models that include free-throw yield; GAM-based refined shot charts; Draft analytics via rank aggregation (rank-biased distance; ranked-choice aggregation); Front-office ROI models: “wins-per-dollar” (salary cap/luxury tax) and analytics staffing ROI (fixed-effects designs estimating wins gained per analyst investment)	Player tracking/shot location + outcome data; foul/free-throw outcomes; ranked lists (mock drafts); team financial/payroll and performance data across seasons
Sprinting (Individual sport)	Biomechanics + performance analytics using long force-platform arrays and motion capture; focus on kinetic/kinematic determinants of maximum speed; torque-based insights (e.g., knee-extension torque, ankle plantar-flexion torque) linked to performance change; injury-risk analytics using	Force-plate data, motion capture variables, torque/kinematic outputs; training exposure measures; injury surveillance/intervention evidence; strength/architectural measures (e.g., fascicle length)

	structured interventions: sprint exposure at ≥ 80 –90% max speed, eccentric hamstring strengthening, biomechanical optimization; monitored adaptations (e.g., fascicle length changes, strength gains)	
Ice Skating (Individual sport)	Technique analytics using instrumented skates (force sensors) measuring push-off forces and center-of-pressure (COP) trajectories ; identification of performance-linked technique markers (e.g., COP positioning; lateral force peaks); injury prevention analytics via multi-component intervention programs (education/training, videos, protective equipment, and policy/rule/environment changes) evaluated with risk ratios (RR)	Skate force sensor data; COP trajectories; force-time measures by phase (curve entry/transition); injury data across large participant samples; intervention design components

Table 1: Analytics used per sport

2.2 Tennis Analytics

Before the mid-2000s, the elite game of tennis relied heavily on basic statistics—first serves percentage, winners, and unforced errors, plus the call booth and the eyes of the on-court line judge. Strategy—from serve placement and return positioning to training loads—was largely a matter of coach and player experience. Line call challenges were abundant, primarily with regards to high-speed serves or balls. The lack of accuracy, utility, and format of usable data was a significant limitation.

In March 2006, the USTA, WTA, and ATP announced the introduction of Hawk-Eye Electronic Line Calling starting with the Miami Masters tournament in April 2006; Hawk-Eye continues today to dominate as the tennis industry-standard. By the end of 2006, electronic replays and the challenge system were in place, followed by multi-camera systems, ball tracking system accuracy standards, and real-time adjudication of line calls at the courtside. This became the beginning of an infrastructure for collecting data for the purpose of analyzing player and ball performance and using the technology (wearing the same technology) for load monitoring of a player and serving as an official database of data regarding injuries at the Grand Slams (eg: Roland Garros).

2.2.1 From Dispute to Data: How Electronic Officiating Anchored Tennis Analytics

Electronic line calls did not just reduce friction between the line judge and player, the ELC became the method of collection and the data source for collecting data on the ball through ball tracking, ball trajectory calibration and impact reconstruction, the measurement of accuracy to millimeters, and the definition of the validity of a Hawk-Eye challenge, as dictated by the rules of tennis. The ITF, specifically, established a validation process (2003–2005) that included three stages for the system, lab, simulation, and on court trials before it was deemed approved to be implemented at tournaments as a challenge mechanism. The tennis tours implemented a system for a limited number of challenges. Broadcasts of the tour and matches started to visualize data on the tracking ball that was collected. Over time, this data source became credible among coaches and players, as well as the sports science community, that highly accurate data was being captured and could be evaluated at a point-level basis.

It is valuable to consider this from a BIS perspective as the platform layer of the electronic line calls became the basis for this collection and data source on which all analytics in tennis will now be constructed. This platform layer is composed of a multi-camera ball tracking system that is the official ball tracking system of record with an operating precision standard of millimeters that produces ball tracking standardized output data with multiple software versions of the Hawk-Eye system allowing for iterative improvements to the system to meet ITF rules and the sanctioning of Hawk-Eye as the standard ball tracking system of record for the purposes of resolving disputes in the most important tennis tournaments worldwide. Now that the platform layer was operational, this opened up two possibilities to BIS. First, researchers were now able to produce shot-level metrics of shot success and shot value by utilizing anonymized ball-tracking trajectory data and event-log data. Second, the adoption of ball tracking allowed for a change from qualitative coaching to quantitative coaching, enabling coaches to utilize a stronger data foundation when making decisions, rather than relying on anecdotal evidence that is limited to a single video clip or coding exercise [12].

In terms of path dependence, if the technology of line-calls had not been implemented, the impetus to create ball-tracking systems of record for each point or player would likely not have occurred. As well, there would have been less available data from a scientific perspective to conduct research on points, shots, and players. By placing the focus on data accuracy, data accuracy evaluation, and data validity in the official decision-making framework, tennis set the foundation for the data revolution.

2.2.2 Shot Value and Modern Matchcraft: Rewriting Tactics with Tracking and Models

Fitzpatrick (2023) explains that shot-level tracking and assessment has allowed researchers and scholars to move from an understanding of tennis points at a box-score analysis level to the geometry and physics of a tennis point. Drawing upon the Hawk-Eye dataset from the hundreds of matches played in Wimbledon, Fitzpatrick investigated the locations of serves and returns to determine trends in serving and returning strategies. Fitzpatrick found that serves in the left (A) and right (D) positions proved more successful than those in the middle zones (B and C), and returns in the left (A) and right (D) positions were more successful than returns into the middle zones (B and C). Furthermore, male winners were found to have been more precise in terms of delivering serves and returns into A and D areas than male losers. However, female winners proved to have greater precision than female losers only in serving. Thus, by using the data, analysts and researchers can adjust their thinking

regarding where to target serves and how to prepare for returns to focus on designing probabilities rather than implementing predetermined plays. [12]

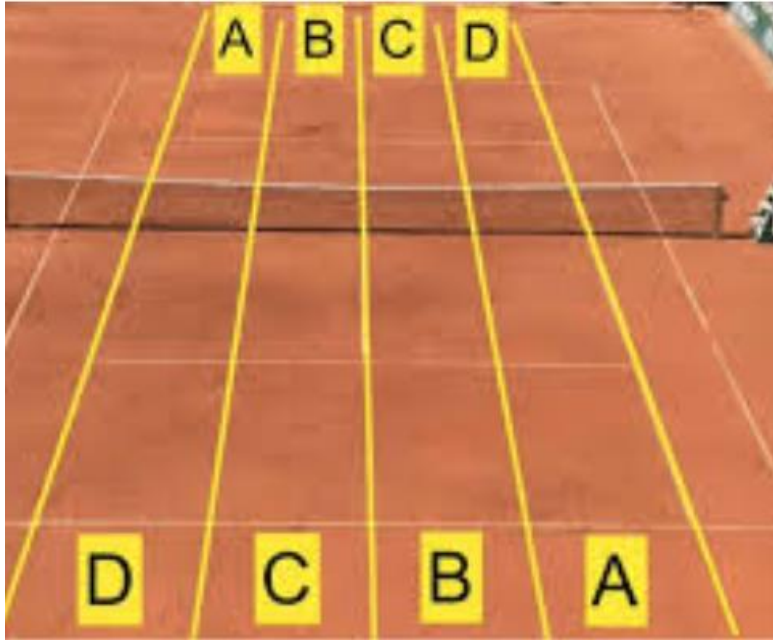


Table 2: First-serve zones

There have been studies conducted over tens of thousands of rallies, including ATP/WT A and Slam matches (262,596 in total), to determine the impacts of serve speed, location, court type, gender, and spin on the likelihood of the serving party winning the point. According to Mecheri, coaches have the potential to shift from their reliance on hunches (“hit harder on a faster surface”) prior to big-data analytics to utilizing the expected value of a combination of target, position, and speed on a variety of surfaces for individual players, i.e., the essence of data-driven strategy, which seeks to maximize the expected point value given a given risk. [13]

Decision modeling has relied upon these earlier works. According to Peter Tea, bayesian hierarchical models can help the researcher make inferences about the intended serve direction (Wide/Body/T) and find out the patterns in a player’s serve choices under different pressure (for example, players choosing a body serve more on second serves than first serves or other high leverage situations). This kind of information can be useful in helping the

returner better predict the server as well as in providing the server with a more robust serving strategy, particularly against skilled returners. [14]

Kovalchik stated that a Markov approach could also be used to distinguish between player’s decision making and skill; rally decision-making simulation using shot and player tracking data showed that as the probability of executing the chosen shot correctly decreases, the shot chosen becomes more conservative, that an error on the backhand leads to a larger expected loss than does an error on the forehand, and that the choice between the serve and return has a larger expected value than does the choice of shots within the rally. This new technique is not intended to find optimal shots that would work for a perfectly skilled player, rather it is to find the optimal shot given a player’s skills and preference in any given situation, effectively connecting training to tactics. [15]

Tennis match-level analyses have been used to determine which match factors are more predictive of match outcomes and it has been shown that short rallies, first serve points and baseline rallies have the greatest correlation to win or lose, regardless of whether the match has a clear winner or a more close match result. Fitzpatrick indicates that no adjustments need to be made to training based on opponent level, that match preparation should focus on the three important components of rally, first serve and baseline skills and that these skills are more critical in all match results. [16]

Tennis has been an important area for the application of game theory, particularly in analyzing the serve. Zhang used metrics that take into account subsequent shots (Service Points Rating) have been found to be more accurate and precise in describing the effect of the first and second serves than the simple statistics of aces and double faults. In particular at the Wimbledon Championships played on the grass court surface it has been observed that the serve and return game appears to be in equilibrium with strategies that are played almost exclusively in pure strategies (players are almost always serving down the T or Wide). The match winning percentage is significantly correlated with serve strength (measured by service points rating) but this effect was not found using the service points rating, suggesting that the effect of the serve on the match outcome is much more significant than the service points rating (a more commonly used metric) suggests. These metrics can be used to analyze a player’s serve performance and can give insight as to how consistent the player has been in executing that serve during a match. Furthermore, it can aid the coach

in determining how well executed the player needs to vary serve locations in order to make the serve more effective. [17]

The prospect also exists for using modeling techniques to determine match momentum as well as predicting match outcomes at various stages in the game studied by Buhamra and Gupta. When these mid-game dynamics were modeled, and momentum was weighted by entropy of point scoring then the machine learning algorithm ensemble was used to predict match outcomes, and high-accuracy predictions could be obtained, suggesting that the psychological momentum of the game is something that can be modeled and quantified. Although there is no on-court coaching in tennis these back-room analysis techniques could provide players with information during a match to be used during change-overs and to provide tips for the player’s next opponent. [18][19]

2.2.3 Quantifying the Training Week: Workload, Periodization, and Wearables

The training landscape changed when validated sensors came to allow stroke counts and movement states to be captured in real time, removing any need for hand coding. Trunk- and neck-mounted IMU’s demonstrated very good accuracy at counting serves (98%) and drive ground strokes (94%) and were less capable of counting slices and volleys; wrist-worn IMUs demonstrated excellent test-retest reliability when counting total shots, with small error rates at capturing wrist speed and acceleration. It matters that we are able to validate this work, because it means that what would previously have only been coach estimates can now be turned into time series data that can inform periodization, load management, and return-to-play programs. [20][21][22]

National programs using Session RPE show that blocks impose a greater load overall than tournaments. However, weekly periodization in these settings is poor or inconsistent, consistent with the crowded calendar of tennis where travel, surface conditions, and draw lengths complicate the application of traditional periodization models. The key message from these data is to monitor load at the weekly level, reduce loads during tournaments where appropriate, and focus training activities on improving those aspects of performance that have been shown to most reliably impact shot value. [23]

The historical example of nine months of a periodized resistance training program in collegiate women showing improvement in strength, jump height, and also, importantly,

stroke velocities (e.g. serve, forehand, backhand) compared to those in a non periodized training group is of particular relevance for tennis. [24]

The performance outcomes of such studies (1 RM, angular velocity measures) are directly associated with technical skills on the court, which may translate into improved ball speeds. Data informed analysis and grouping methods also help identify the most effective approach for development programs. A 6 week training program in 10–14 year old tennis players saw improvements in ITN rating as well as short sprint performance and agility. However, cluster analysis revealed that players could shift from cluster to cluster across time and that players in different clusters demonstrated varying responses to the program. [25]

This suggests that training programs in tennis players must consider factors including the stage of player development and baseline player characteristics. More broadly it suggests that dose responses must be understood at the individual level, rather than just at the level of a generic program block. A review of 345 microcycle drills utilizing inertial tracking devices revealed that the majority of distance covered and the PlayerLoad metric were associated with baseline and all court situations, while serve drills were consistently among the lowest external load activities. Juniors covered more distance and accumulated higher loads than professionals in baseline and return baseline situations. Professionals, however, exhibited higher levels of acceleration and deceleration. [26]

2.2.4 Engineering the Serve: Biomechanics and Technique Transformation

Biomechanical research has identified technique and kinematic variables related to serve and groundstroke performance, replacing trial and error approaches with evidence based coaching targets. Multiple reviews have characterized the serve and the groundstroke (e.g., forehand, backhand) using an array of techniques and outcomes to characterize performance and movement. These include studies comparing the use of force plates to pressure sensitive insoles, or using optical motion based motion capture technologies to inertial sensors (e.g. IMU), as well as characterizing performance outcomes in terms of ball speed, ground reaction forces, and joint torque outputs. Although these studies may be disparate in design and methodology, several critical positions and movements have been identified to be associated with increased serve performance, which include the trophy position, racket low point position and the impact position. [27]

A review of studies examining these three key serve positions (e.g. trophy position, racket low point, impact position) identified specific joint angles associated with key performance

metrics. The review found mean values of lateral rotation of the shoulder of around 130° at racket low point; elbow flexion and thoracic lateral tilt angles at impact position and trunk angle and knee flexion angles at the trophy position. Such findings may be used by a tennis coach or sports scientist to develop targeted drills, such as those focusing on thoracic rotation and upper thoracic acceleration, while limiting the magnitude of shoulder external rotation and/or shoulder flexion, to optimize serve speed without compromising on safety and without relying solely on the athlete's own trial and error process. [28]

A recent study utilizing field measures of inertial sensing devices demonstrated that greater serve velocity is associated with increased shoulder and wrist angular velocities, greater trunk angular velocity at the point of ball contact, and increased shoulder external and internal angular velocities at the point of ball contact, which is supported by studies using motion capture technology to analyze performance at racket low point. Further, the analysis showed a significant effect on the association with serve velocity with greater hip angular velocities, which are also associated with the performance of the serve and backhand. [29]

In turn, these results suggest that a key mechanism for efficient serving is the sequential transfer of the kinetic energy up the kinetic chain in the movement. For coaches and practitioners working with tennis players, this suggests that the development plan should focus on the sequence of development: lower limb push off, then movement of the pelvis and torso rotation, followed by movement in the shoulder, elbow and wrist. Furthermore, monitoring angular velocities at different segments of the kinetic chain (e.g., at wrist, shoulder, torso, hips, etc.) can form the KPIs for performance improvement in the serving technique as well as other strokes of tennis.

Finally, while improving the biomechanical aspects of the serve can be expected to improve serve velocity, the overall movement load will likely increase. In addition to this, the biomechanical aspects of other strokes such as slice and volleys, which were previously not considered when using IMUs in research and practice, may not necessarily be monitored, leading to a potential imbalance in training load. This means that technical interventions associated with such strokes must be monitored with a load capture mechanism which may involve video assessments during the short term, and the long term transfer of technique should be re assessed after the completion of the periodized block in order to verify that there have been improvements in ball speeds of those strokes during the matches.

2.2.5 Turning Surveillance into Prevention: Injury Epidemiology and Risk Control

Longitudinal injury surveillance has enriched our understanding of injury trends and promoted injury prevention by anticipating needs. From 2011 to 2022, 750 injuries were recorded at Roland Garros, 62.5 injuries per event. The most prevalent were lower extremity injuries (hip, thigh, pelvis, ankle, foot) followed by the muscles, tendons and joints. The most frequent injuries were adductor, abdominal rectus and lumbosacral muscle strains; they are expected in a clay player that makes large movements involving the lower limb rotation. These results are helpful for the medical staffs to plan the number of staff, screening strategies, and the training to be developed in the clay period of the season. [30]

This review describes tennis injury epidemiology in all levels of play (i.e. professional, collegiate, junior) and highlights the progression made in the ATP monitoring for injury. Tennis has many risk factors related to the overuse injuries, which are the increased total volume of work and the variability in match time. Examples of injuries with overuse and repetitive pattern are in the lower spine, knee, ankle and shoulder. In 2009, monitoring for overuse injury was standardized, leading to more reliable data and a better ability to make decisions around risk. [31]

Risk factors allow to define thresholds for prediction of injuries. The best predictors were training duration and intensity, level of play (elite), number of days on clay, surface (clay) and type of injury (wrist and ankle on serve). Using these results on tennis training would require implementing a dose in tennis training with a defined dose threshold (e.g. number of ball hit, PlayerLoad threshold with IMU device), rotating the players between the surfaces, limiting the total number of serves during training period, defining the dose to the return-to-play (RTP) decision making process rather than to the traditional RTP protocols based on set criteria. [32]

A data infrastructure supporting the prevention of injuries should include those tools, which would require a data pipeline connecting wearable and session-RPE to obtain the dose; risk map of the monitoring system by surface and age group and objective indicators in RTP protocol (e.g. hitting load, movement intensity, and pain/function scales). In tennis, a health informatics system (BIS) will also be crucial for this data-driven prevention program, which include the availability of data to the medical personnel, security and privacy of the medical dataset and the design of specific dashboards and data visualization for coaches, S&C staff and medical personnel to make a more data-informed decision daily.

2.2.6 Systems Thinking for Tennis Analytics: A Business Information Systems View

All four fields have one common foundation, which is a strong data infrastructure. Officiating and governance provide credibility to data collection and auditing processes; monitoring and performance depend on a clear and well-structured data pipeline and reproducible statistical methods; workload management requires a proven set of wearable technologies and a standard data format; injury surveillance depends on longitudinal data registries; and biomechanics research is supported by the interchangeability between field and laboratory instrumentation systems [2]. On a mechanical level, this translates into measuring, data gathering, modeling and understanding. From a BIS standpoint the tennis industry and technology development for it has basically followed an adoption and maturity framework: instrumentation → Data governance and management → Analytic modeling → Decision support.

The important next phase for technology in tennis is the interoperability of data from many sources — HawkEye trajectories, wearables, weights room data and medical records — into one system that can inform every aspect of tennis. We can envision a combined player graph including performance metrics, capacity for load and injury risk, viewed via role-based dashboards and alerts to optimize daily decision making across the tennis organization. Special attention must be given to privacy, security and confidentiality to protect the privacy of data related to any player, coach or staff. Once the above challenges have been solved, there are opportunities to make breakthroughs in automated scheduling engines that can suggest workouts and travel schedules based on an individual player’s injury risk and performance goals, automated real-time alerts when an athlete is exceeding load limits or shows signs of fatigue, and “AI-performance coaches” that suggest serving strategies or technique fixes in real-time based on an individual player’s profile and their prior game results. Moreover, one must confirm that the protocols HawkEye uses to establish and maintain technical accuracy, version control and auditability in an official capacity can be applied to any new data pipelines used for the above use cases.

In 2006, technology in professional sports took a massive leap forward when a professional tennis tournament utilized video-game replays of every point of serve in order to assist the referees. The adoption of technology and analytics has completely changed the game. Analytics allows a coach to spot and rectify mistakes, wearable technology allows for the ability to track workload in order to help coach conditioning, biomechanical data gives

insight to how technical performance relates to injury risk, and we are finally able to monitor player health and workload during competition. Every one of these analytic methods gives scientific, evidence-based responses to questions that were once impossible to answer and can empower a coaching, training and medical staff to make better informed decisions on all aspects of the game from where to aim your serve and work on technique to risk management. Incorporating these methods into a contemporary Health Informatics system (BIS) might be the next step in the evolution of tennis and could lead to improved performance and well-being of the players. Further research in areas like interoperability, privacy preserving analytics, and data sharing programs could provide insights to assist data-driven tennis training and preparation programs, potentially leading to better player health.

3. Methodology

3.1 Research Design and Purpose

This dissertation seeks to empirically elucidate the impact of data availability on professional men’s tennis through a comparison of measurable match characteristics prior to and following a well-documented technological advancement: the launch of Hawk Eye electronic line calling in the professional tennis circuit in 2006. Hawk Eye first made its appearance on tour at the Miami tournament in 2006, then became a part of the Grand Slam competition at the US Open later that same season; the technology has since grown exponentially over various tournaments as an integral component of the increasingly data-driven environment in tennis. This is thus an arbitrary breakpoint, which provides a natural pre–post comparison.

This dissertation is quantitative, observational, quasi-experimental pre–post in nature and determines whether any differences exist between KPIs and patterns at the match level in the pre-data and post-data eras. The approach aligns with the business analytics methodology, which employs real-world operational data to search for structural differences between time periods designated by technology adoption. The data will be analyzed using

descriptive statistics (summary measures, plots, distributional analyses) and hypothesis testing (differences in means, proportions, distributions).

3.2 Data Source and Dataset Description

The data used in the study are from the Jeff Sackmann open tennis data repository. It is available for download from Jeff Sackmann github and offers match-level outcomes, tournament information, player characteristics, rankings, and (when available) serve and pressure statistics for the two players (winner and loser). The repository also has a series of files `atp_matches_YYYY.csv`, where YYYY is the season year.

The data unit for the analysis is an individual ATP match (i.e., men’s singles). There is one match per row with the following observations:

Tournament information: tournament ID/name, date, surface, tournament level, draw size, round;

Outcome: winner/loser; match score; match time in minutes if available; best-of-3 / best-of-5;

Player characteristics: winner/loser name and player ID, handedness, height, age, and nationality;

Rankings: winner/loser rank and rank points if available;

Serve and pressure statistics for winner/loser: aces, double faults, serve points won, 1st serves in, 1st serve points won, 2nd serve points won, service games won, break points faced and saved.

The analysis is including matches from 1985 and after. This choice is based on expected improvements in the completeness of detailed match statistics over time and is aligned with the practical requirement of using sufficient tennis oriented fields for key performance indicators.

The analysis uses data from 1985 and forward. This period is selected to ensure the availability of sufficiently granular match statistics in the dataset (i.e., the expectation is that the match-level statistics improved progressively over time), as well as to meet the requirement to have at least one tennis-specific variable for each key performance indicator used for the analysis.

The analysis is restricted to the ATP main tour level, using the tournament level codes from the dataset to select for the more elite contexts. The data sample retains matches from the following tournament categories::

- G (Grand Slams),
- M (Masters 1000),
- A (other tour-level events),
- F (tour finals / season-ending events).

We filter out matches when score information is absent or a match ends with a walkover (as score-related variables cannot be calculated with any confidence) and any required fields for a specific indicator are missing. We also define two epochs: (Pre vs Post “Data Enablement”). The pre-data era includes the years 1985–2005 and the post-data era includes 2006–the most recent year in the dataset. The breakpoint between the two is based on when Hawk Eye was officially introduced to the tour-level at Miami 2006 and subsequently adopted at the Grand Slams in that same year. As 2006 was a transition year (not all tournaments had adopted the technology), we include robustness checks in the methodology. To be conservative, we omit the year 2006 entirely (1985–2005 vs 2007+).

Match files in the repository usually feature extensive serve statistics, although not all of them have been recorded. This is why the empirical strategy has two more analysis cohorts. The first cohort (Full-match) employs variables that don’t need granular serve stats (like upset rates based on rankings, match length (where available), and score-related variables such as tie-breaks). The second cohort (Serve-stat) utilizes serve/pressure indicators (first-serve %, double fault %, ace %, break-point %, hold%). The serve-stat cohort only includes matches with complete serve statistics for both the winner and loser players.

3.3 Surface Context

Surface context significantly influences both the dynamics of a match and its outcome, necessitating a separate analysis across all surfaces. In general, grass courts tend to produce fast playing conditions (higher skid and lower bounce), thus favoring serve dominance (more tiebreaks, more aces, higher hold %). Hard courts are also fast, but can

vary from tournament to tournament and across time. Clay courts slow down the game (higher friction and bounce), thus elongating rallies and improving return efficiency (more tiebreaks, more aces, and a higher hold%). Carpet courts (not in this study) are surfaces consisting of a layer of synthetic textile or polymeric material rolled out over a concrete or wooden base. This surface will be excluded from the analysis because it will not have main event matches after the year 2026. The analysis is conducted surface-specific for both descriptive reporting and hypothesis-testing. Doing this isolates the effects of surface changes (e.g. when one era has a higher number of hard court matches) as a confounding factor and increases the interpretability of results.

3.4 Variable Construction and Key Performance Indicators (KPIs)

All engineered variables use fields from the match dataset; they are either interpretable or understood as tennis indicators (e.g. serve dominance, return pressure, match tempo, or competitiveness) or both; and comparable across era and surface. KPIs are constructed at the match level, combining statistics of the winning and losing players where possible in order to minimize any single player bias and capture match conditions more holistically.

Serve and Return KPIs (Serve-stat Cohort)

Serve volume

Total serve points: $w_svpt + l_svpt$

Represents total points played on serve in the match.

Serve aggression and risk

Ace rate: $(w_ace + l_ace) / (w_svpt + l_svpt)$

Double fault rate: $(w_df + l_df) / (w_svpt + l_svpt)$

First-serve quality and effectiveness

First-serve in %: $(w_1stIn + l_1stIn) / (w_svpt + l_svpt)$

First-serve points won %: $(w_1stWon + l_1stWon) / (w_1stIn + l_1stIn)$

Second-serve effectiveness

Second-serve points are $(svpt - 1stIn)$ for each player.

Second-serve points won %: $(w_2ndWon + l_2ndWon) / [(w_svpt - w_1stIn) + (l_svpt - l_1stIn)]$

Hold and break dynamics Using break point information:

Breaks conceded by a player can be approximated by $(bpFaced - bpSaved)$.

Service games held: $SvGms - breaks_conceded$.

Hold % (per player): $holds / SvGms$.

Match average hold %: average of winner and loser hold %.

Break point pressure and conversion

Break point chances per service game: $(w_bpFaced + l_bpFaced) / (w_SvGms + l_SvGms)$, interpreted as how often returners create break opportunities.

Break conversion % (match level): $total\ breaks / total\ break\ points\ faced$.

These KPIs will allow the dissertation to test the assumption that post-2006 matches follow patterns that indicate more strategic play and data-driven serving and returning tactics.

3.5 Descriptive Statistics and Visualization Plan

Descriptive analysis is used to summarize and communicate patterns before formal inference. The dissertation presents: Summary Tables (Pre vs Post)

For each KPI, report:

N (sample size), Mean and standard deviation, Median and interquartile range (IQR), Minimum/maximum, Missingness rate (where relevant). Missingness is referred to incomplete data within the dataset. In the specific dataset, the variables are not uniformly available through all years, having earlier matches showing some lower coverage. The analysis using matches with complete records.

Tables are provided overall and separately by surface.

Although inference is based on pre vs post groups, the methodology includes time trend visuals to show whether changes are sudden or progressive:

Yearly mean of each KPI (by surface), with a vertical reference at 2006. These plots are descriptive only (not causal) but help interpret whether 2006 aligns with visible shifts.

To assess shape, skewness, and outliers:

Histograms (with density curves) for key KPIs overall and by surface, boxplots in order to compare pre vs post (and by surface), kurtosis and skewness statistics, reported with comments (positive excess kurtosis suggests heavy tails/outliers, high skew suggests asymmetry and possible need for non-parametric tests.)

This distribution work supports the choice of hypothesis test and helps so as to validate the assumptions (normal approximation vs non parametric alternatives).

Moreover, pie charts (and/or stacked bars) will summarize sample composition: Surface distribution (hard/clay/grass) by era, tournament level distribution by era.

These charts clarify whether observed era differences might partially reflect calendar composition changes.

3.6 Hypothesis Testing Strategy

Hypothesis testing will take place by comparing the pre and post era for each KPI both at an overall level, as well as by surface. The hypotheses will look like this:

- Null hypothesis (H_0): No difference between eras in KPI (mean/proportion/distribution).
- Alternative hypothesis (H_1): A difference exists (two sided) or an expected direction exists (one sided, only when strongly justified).

Tests for Differences in Means (Continuous KPIs)

For KPIs such as ace rate, hold%, minutes per game:

- Use Welch’s two-sample t-test (robust to unequal variances) where approximate normality is plausible due to large sample size.
- A Welch's two sample t test will be run if normality can be reasonably assumed (e.g. central limit theorem holds due to sufficiently large sample size).

If distribution is skewed/heavy-tailed (judged by kurtosis/skewness, as well as the distribution plots), Mann-Whitney U test (non-parametric) results will also be reported, to test that the differences are not merely driven by outliers.

Reported outputs include:

- Mean difference (post – pre),
- 95% confidence interval,
- Test statistic and p-value,
- An interpretation sentence connecting results to tennis mechanisms.

Outputs include:

- Difference in proportions,
- Confidence interval for the difference,
- p-value.

3.7 Software, Workflow, and Reproducibility

All analyses were conducted in Python and can be fully reproduced. This includes ingesting and merging annual match CSV files, filtering data for year range, tour level, and data completeness flags, deriving KPIs and parsing result strings, creating descriptive plots and descriptive statistics tables, conducting hypothesis tests on the full dataset as well as for subsets by surface, and automatically saving summary tables and plots in a thesis-ready format. This is made possible by:

A standard directory structure (data/in and data/out);

An automated script to save figures and tables;

Documented definitions of KPIs and of data inclusion and exclusion criteria.

3.8 Methodological Limitations (within this design)

The research design contains several important limitations:

Observational Design: The pre-post analysis cannot confirm that Hawk-Eye caused the shift in tennis performance. Hawk-Eye also arrived around the same time that other factors were changing, including conditioning advances, racquet and string technology developments, and data analytics tools for player and tactical analysis. Hawk-Eye may have been the most high-profile and tangible marker of what we might call "the data revolution", but I will treat causality with care and frame the results as Hawk-Eye being associated with, rather than causing, the observed changes in performance.

Incomplete Data: Certain serving statistics were not reliably and consistently recorded prior to the Hawk-Eye era. While I will do my best to control for this by analyzing subsamples with adequate data coverage and reporting rates of missing data per subsample, it is not fully mitigated in my analysis design.

Surface Variation: While I stratified by surface type, not all hard courts (or clay or grass) are the same. Hard courts (indoor/outdoor, carpet, etc.) have different surface compositions; further, individual hard-court tournaments will vary in playing surface and playing conditions. Therefore, it may be more accurate to report differences in terms of tournament type than by broad surface categorization. Indeed, independent research has documented that the average tempo of play differs by tournament. I therefore frame the findings in terms of tournament and surface type, not individual matches.

Match-Level Detail Limitation: The current dataset does not record point-level details such as point type (e.g., winner, unforced error, winner off a forced error, etc.) or rally length. The results, then, may be interpreted in terms of the tactics that lead to those point-level categories, but not in detail at a point level.

Summary

This thesis employs a publicly available dataset and a well-defined historical breakpoint to examine professional tennis performance pre-vs-post Hawk-Eye implementation. A transparently documented, reproducible, and open data pipeline was built using Python, and this data was cleaned, aggregated, and prepared using the pipeline. Using basic descriptive statistics and plots, distributional diagnostic checks, and appropriate statistical hypothesis testing, the data are used to compare performance across the eras and surface types to assess serve dominance, return pressure, game length, and competitive balance. This work provides a data-driven framework for analyzing whether and how the data

revolution has impacted professional tennis performance, while documenting the research design and constraints in a fashion both useful for data scientists and useful for tennis professionals.

This work cannot conclude a cause and effect between Hawk-Eye and observed changes. I have taken a pre-vs-post design that can identify differences between eras but not the cause-and-effect between Hawk-Eye and those observed differences. However, the analysis has shown that the post-2006 era, when data is widely available, is statistically different than pre-Hawk-Eye and I caution readers to note those findings cannot confirm Hawk Eye as the cause of those performance changes.

4. Results

4.1 Descriptive Statistics

The analysis has been held on a large scale of data. The numbers of the sample regarding pre and post 2006 and by surface are the below:

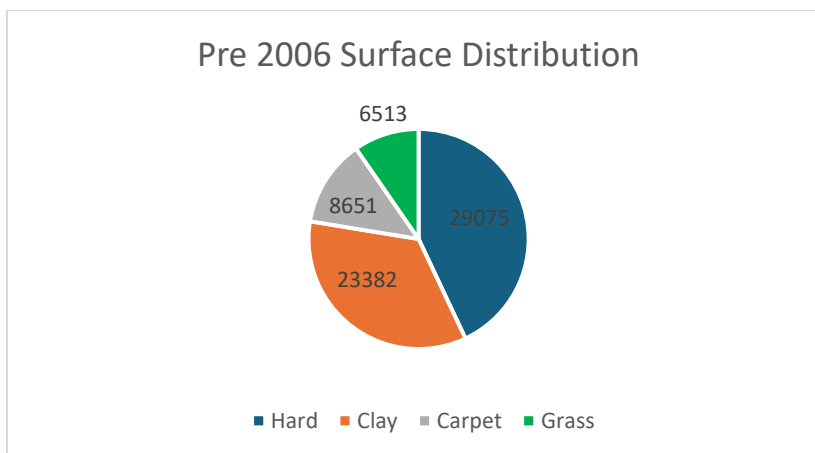


Table 3: Pre 2026 data distribution

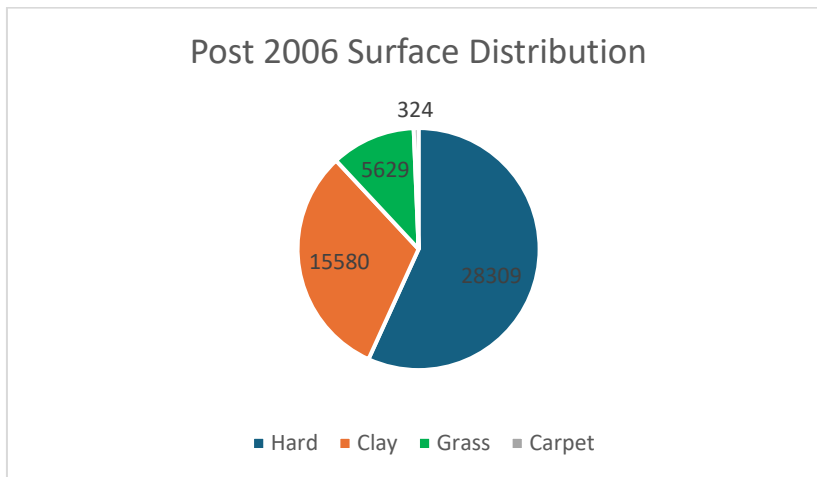


Table 4: Post 2006 data distribution

At first sight, the number of samples is fine other than the carpet surface which is excluded from the analysis

The descriptive analysis is based on the following KPIs:

Ace rate, Double fault rate, first serve in rate, first serve points won rate, hold rate, and second serve points won rate. Boxplots and histograms will be shown below in order to see the pre and post era.

1) Ace rate:

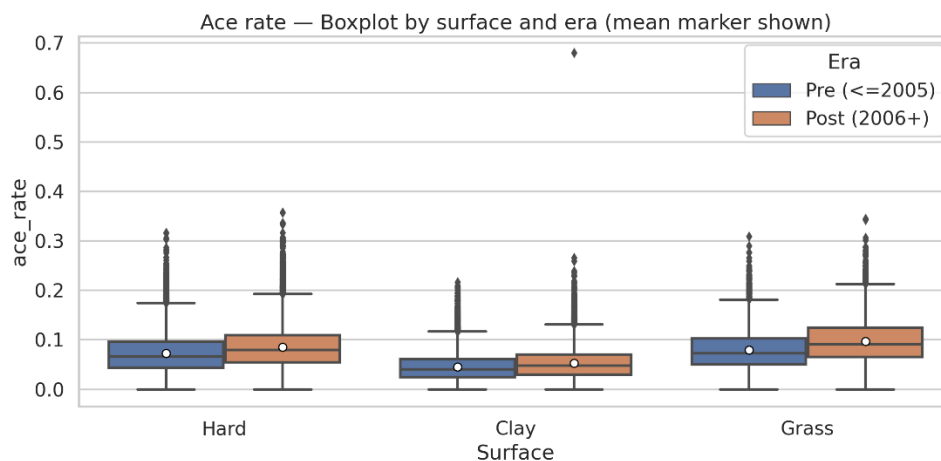


Table 5: Ace rate boxplots per surface per era

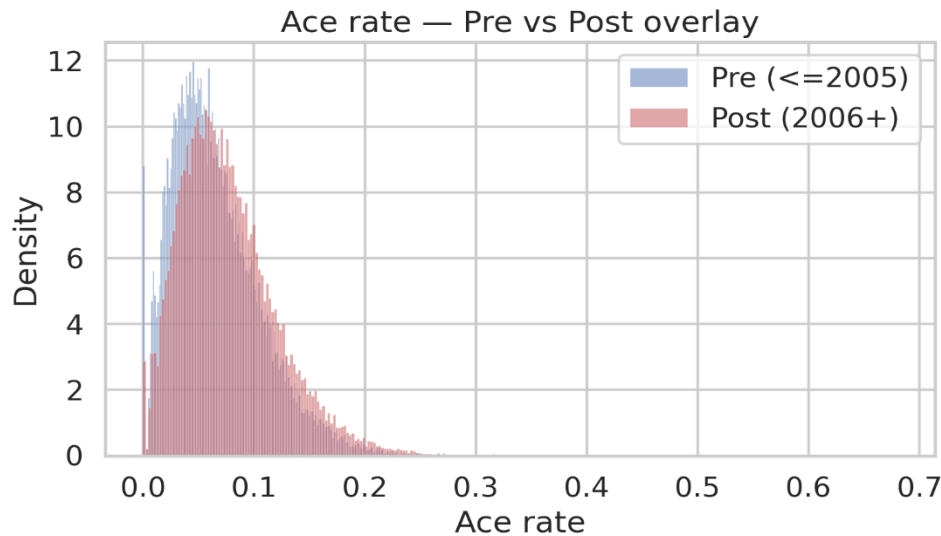


Table 6: Ace rate histogram per era

Figure X of the boxplot compares the distribution of the KPI ace rate, separated by court surface. It is clear that in all surfaces, post 2026 distributions are upwards in comparison to pre-2026 distributions. This is shown due to higher medians and higher mean markers (white circles). The ace rate ranking remains consistent. Grass has the higher rate due to the “slippery” surface, clay has the lowest rate due to the “slow” surface and hard along with carpet are in between, which makes absolute sense. In all surfaces the interquartile ranges (IQR) shifts slightly upwards indicating that the increase is not derived from few extreme matches, but it reflects a general upward trend. Of course, outliers exist something which is very normal. In addition to the boxplot, which depict a higher median/mean ace rate in the post-2006 era, the histogram shows that the entire distribution lays slightly to the right after 2006. Moreover, the shape is remaining strongly right tailed in both eras, which means than ace-heavy matches still exist but the-post 2006 period has a larger share of matches in the higher ace rate range.

2) Double fault rate:

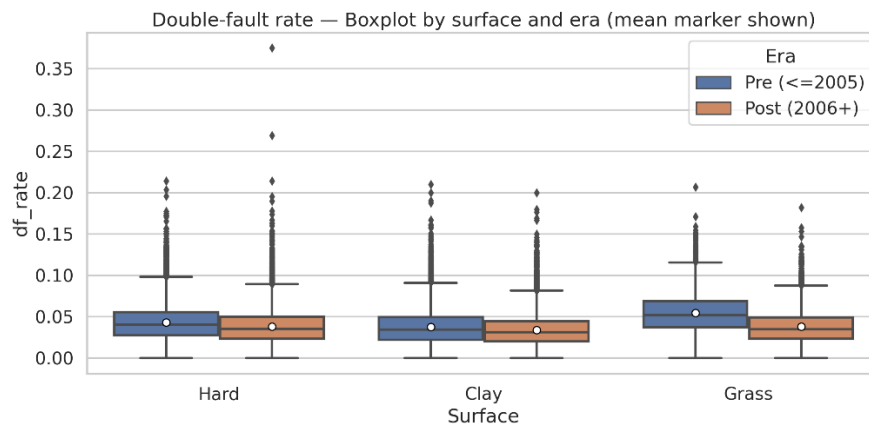


Table 7: Double fault rate boxplots per surface per era

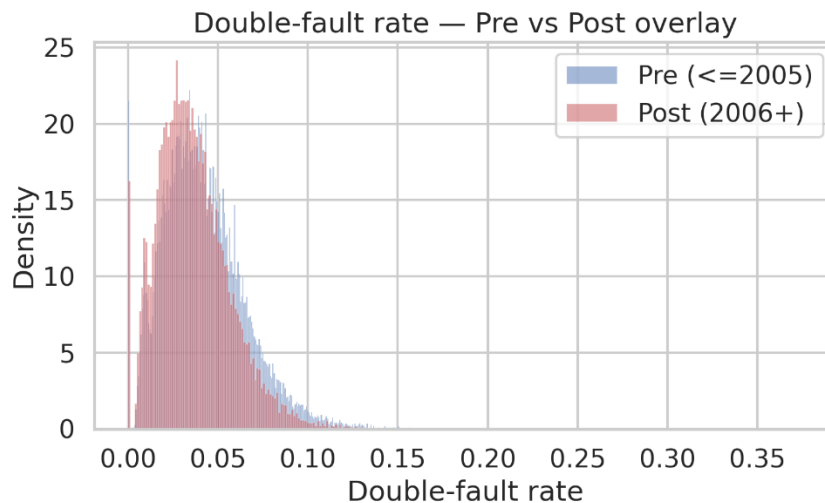


Table 8: Double fault rate histogram per era

Figure X (double fault rate by surface and era) clearly suggests that the post-2006 period has slightly “cleaner” serving faults across all surfaces. The largest visible reduction is depicted in grass surface in comparison to the rest surfaces which the reduction is smaller. This is due to the fact that grass is a “faster” surface and the ace rate is much higher than rest surfaces. This implies that even in the second serve, tennis players were going for a “powerful” serve in order to achieve an ace. Regarding outliers, all surfaces show many high outliers with hard surface having the most. Overall, the main visual message is not a dramatic structural change in spread, but a gradual reduction of post-2006 tendency. In addition to the boxplot which shows an eradication of double faults, the histogram validates this pattern which is depicted from the density that is slightly concentrated more toward lower values.

3) First serve in rate:

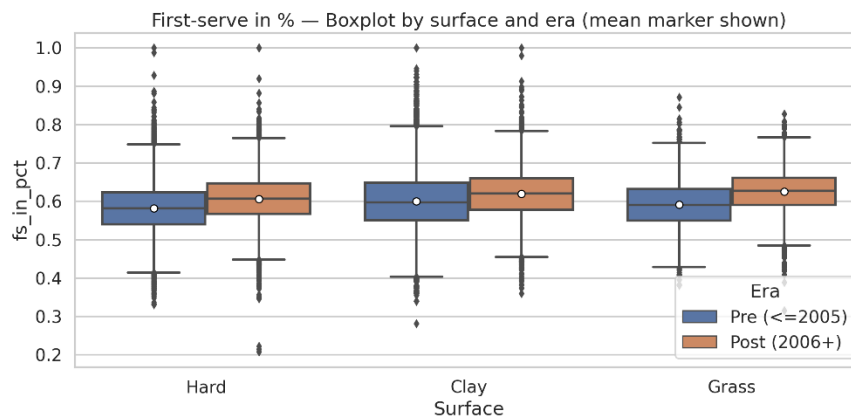


Table 9: First serve in rate boxplots per surface per era

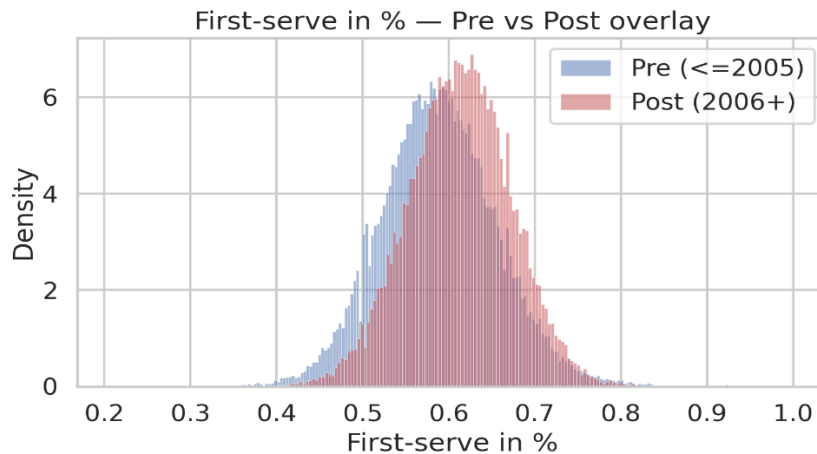


Table 10: First serve in rate histogram per era

Figure X (first serve in %) is showing a clear upward shift in first-serve consistency in the post-2006 era across all surfaces. The boxplots in post era sit higher than pre 2006 ones, having higher medians and mean markers. This suggests that a typical match after 2006 shows a marginally higher proportion of first serves landing in. At the same time the (IQRs) stay similar implying that there is not a major change in variability. Regarding outliers, on the upper side, several matches show unusual high first serve in percentages 0.9 and 1. This happens because some players have extremely accurate serve and/or when the match has relatively few serve points. On the lower side, there are scattered outliers between 0.2 and 0.4 indicating that some matches had poor first serve accuracy. A comment about the histogram is that it show a small rightward shift in the distribution after 2006 and in both

eras it displays a spread in both sides of the main peak, implying very low and high accuracy matches remain possible regardless of the improvement.

4) First serve points won rate:

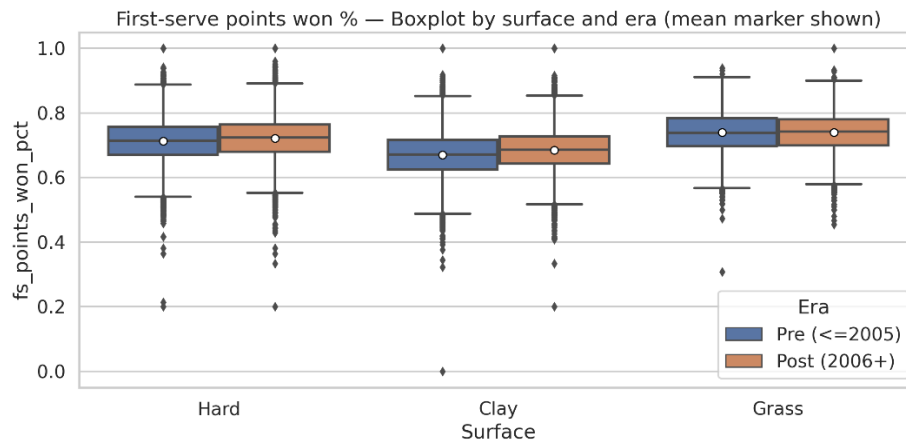


Table 11: First serve points won rate boxplots per surface per era

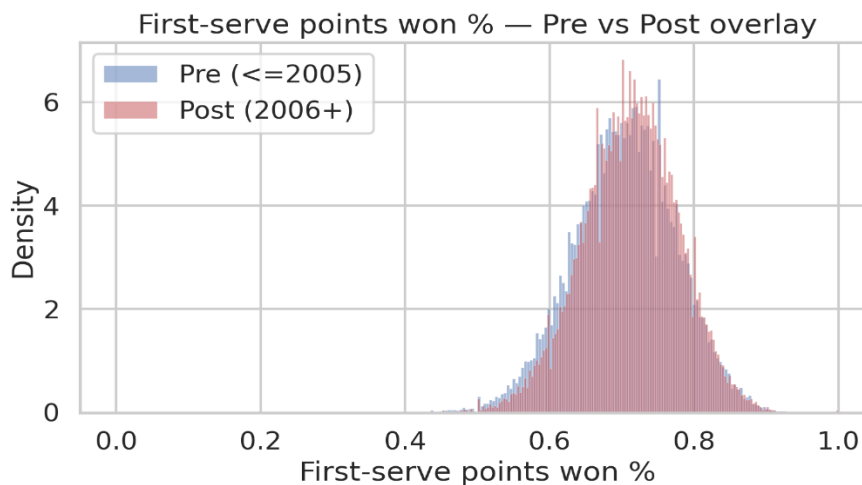


Table 12: First serve points won rate histogram per era

Figure X (first serve points won %) indicates that players have the tendency to win a slightly higher share of points behind their first serve which is a pattern across all surfaces. This is a KPI related to some of the above because serving better provides an “easier” point. The surface ordering remains intuitive. Grass and carpet show the highest values having hard surface rates very close. Clay surface sits lower which implies more returnable serves and longer rally points. Outliers have same distribution as in the First serve in rate, showing

some very high and very low. The histogram here indicates that the distribution overlaps substantially, suggesting the change is incremental rather than structural, with outliers on both ends continuing to appear in both periods.

5) Second serve points won rate:

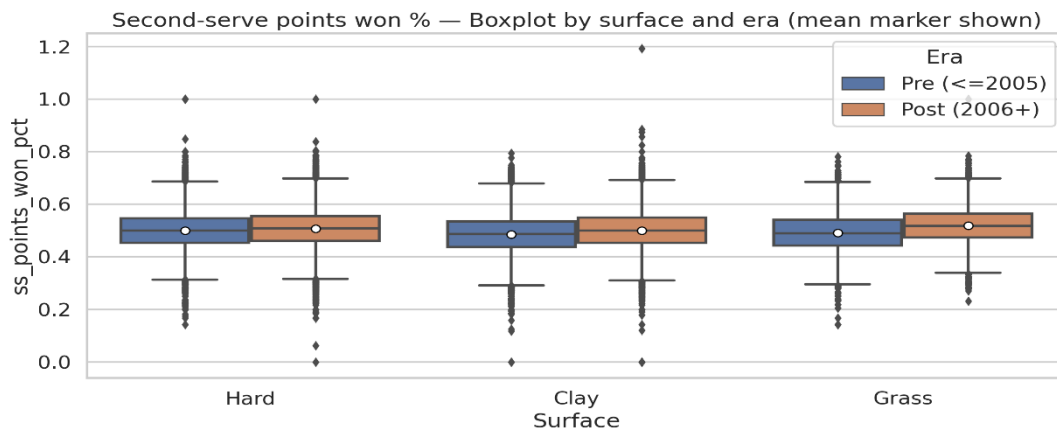


Table 13: Second serve points won rate boxplots per surface per era

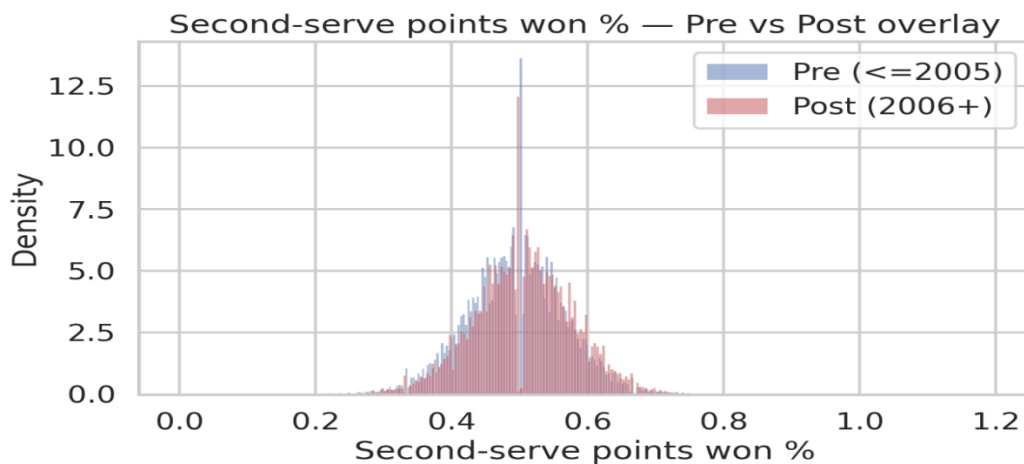


Table 14: Second serve points won rate histogram per era

Figure X (second serve points won%) shows a moderate but stable post-2006 improvement in second serve effectiveness. Post 2006 medians and mean markers sitting above the pre-2006 era across all surfaces. This implies that after 2006, a higher share of second serve points exist which is related to a better second serve, a batter point placement, a better spin of the ball or a better strategy within the point. The surface ordering remains the same as the aforementioned KPIs which makes clear sense.

6) Hold % (match average)

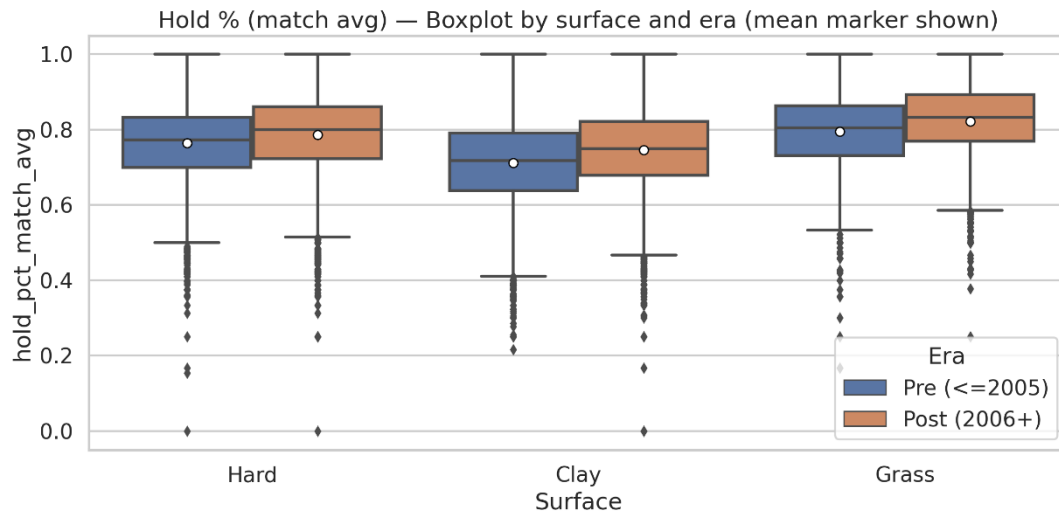


Table 15: Hold rate boxplots per surface per era

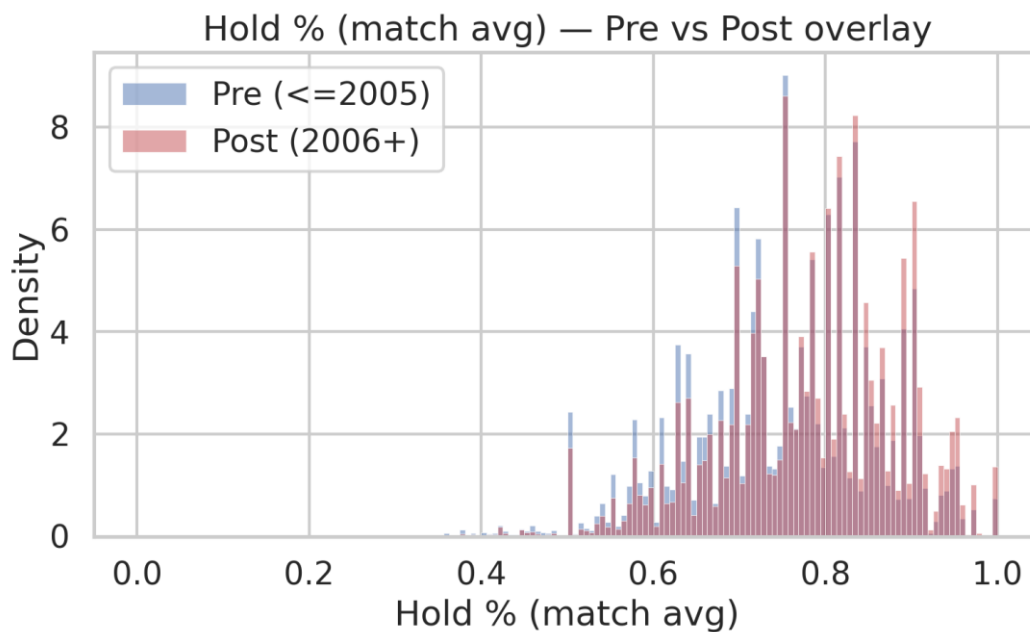


Table 16: Hold rate histogram per era

Figure X (hold%) suggests that the post era is characterized by higher service game control across all surfaces. Medians and means sit higher in the post 2006 era. The pattern is the same, having higher indexes in grass, carpet and hard while clay has slightly lower which is due to the “slower” and more “returnable” surface. Outliers show that some matches have

perfect (100%) serving hold which is very normal while some lower outliers show that some matches have 0% hold which is explained by the difference between the players’ quality.

Finally, having all the above KPIs improved it is implied that the match has longer duration and the service breaks are difficult to be made. This leads to the result of the tendency of having more tie breaks. It is clearly shown below on the time series which are depicted by surface. As it is known, the carpet surface has no data after 2006 but if it had, the tendency would have been the same.

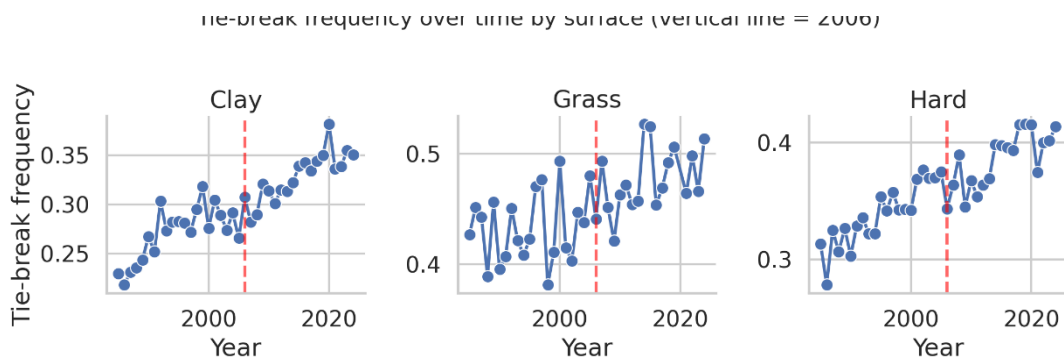


Table 17: Tie break time series

4.2 Hypothesis testing

Following the analysis made with boxplots and histograms, some hypothesis tests will be held in order to provide additional support for the underlying assumption.

The test is the same across all KPIs. The assumption that needs to be tested is if the KPI is the same in the pre-era and in the post-era.

H_0 : The KPI is the same in the pre-era and post-era

$H_0: \mu_{pre} = \mu_{post} \text{ or } p_{pre} = p_{post}$

H_1 : The KPI differs between eras

$H_1: \mu_{pre} \neq \mu_{post} \text{ or } p_{pre} \neq p_{post}$

where μ is mean and p is proportion.

The tests that is held is a t test for all KPIs. A t-test is a statistical method used to determine if there is a significant difference between the means (averages) of two groups, which may

be related to certain features. It helps determine if a result is statistically significant—unlikely to have occurred by chance—or if a hypothesis should be rejected.

The null hypothesis will be rejected if the p-value is less than 0.05. However, the significance level of 0.01 will be used due to the large sample size of data and the high quality. This will reduce the likelihood of false positives and ensure more robust statistical conclusions.

H1: Ace rate

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	Test stat	p-value	Decision
Hard	20,919	28,177	0.07266	0.08562	+0.01296	-34.389	5.68e-256	Reject H0
Clay	16,119	15,502	0.04574	0.05281	+0.00708	-20.323	3.15e-91	Reject H0
Grass	4,441	5,625	0.07949	0.09735	+0.01786	-20.989	1.00e-95	Reject H0

Table 18: Ace Rate test results

Ace rate is higher after 2006 on all surfaces, with the largest absolute increase on grass, consistent with a broad shift toward stronger serve outcomes in the post era.

H2: Double fault rate

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	Test stat	p-value	Decision
Hard	20,919	28,177	0.04293	0.03791	–0.00503	25.740	5.04e-145	Reject H0
Clay	16,119	15,502	0.03746	0.03379	–0.00367	15.776	7.48e-56	Reject H0
Grass	4,441	5,625	0.05454	0.03777	–0.01678	36.406	5.38e-270	Reject H0

Table 19: Double fault test results

Double-fault rate is significantly lower post-2006 across all surfaces — especially on grass, where the drop is largest — suggesting improved risk control or second-serve management in the post era.

H3: Hold%

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	Test stat	p-value	Decision
Hard	20,919	28,177	0.76482	0.78683	+0.02202	-23.665	4.68e-123	Reject H0
Clay	16,119	15,5	0.71148	0.74563	+0.03414	-27.875	5.94e-169	Reject H0
Grass	4,441	5,625	0.79452	0.82226	+0.02774	-14.588	1.14e-47	Reject H0

Table 20: Hold rate test results

The post era shows higher holding on every surface, with the largest increase on clay and grass, aligning with the broader shift toward server control after 2006.

H4: First serve in %

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	t-stat	p-value	Decision
Hard	20,919	28,177	0.5820	0.6063	+0.0243	-43.133	0.0 ($\approx p < 0.001$)	Reject H0
Clay	16,119	15,502	0.6002	0.6193	+0.0191	-24.725	1.14e–133	Reject H0
Grass	4,441	5,625	0.5913	0.6255	+0.0342	-29.643	2.13e–184	Reject H0

Table 21: First serve in rate test results

First-serve consistency is significantly higher post-2006 on all surfaces. The largest improvement is made on grass having 3.42% increase.

H5: First serve points won %

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	t-stat	p-value	Decision
Hard	20,919	28,177	0.7129	0.7220	+0.0090	-15.307	9.43e–53	Reject H0
Clay	16,119	15,502	0.6703	0.6844	+0.0141	-18.820	1.40e–78	Reject H0
Grass	4,441	5,625	0.7394	0.7400	+0.0006	-0.466	0.6412	Fail to reject H0

Table 22: First serve points won rate test results

On hard and clay, players win a higher share of first serve points post-2006 ~ 1% more. On grass however, there is no statistically significant change. This pattern suggests that the post-2006 improvement in first-serve effectiveness is clearer on hard and clay, while grass (already highly serve-friendly) shows little measurable change in this KPI.

H6: Second serve points won %

Surface	N(pre)	N(post)	Mean(pre)	Mean(post)	Δ (post–pre)	t-stat	p-value	Decision
Hard	20,918	28,175	0.4995	0.5071	+0.0076	-11.479	1.85e–30	Reject H0
Clay	16,117	15,501	0.4858	0.5006	+0.0148	-17.641	2.57e–69	Reject H0
Grass	4,441	5,625	0.4910	0.5187	+0.0277	-18.995	5.85e–79	Reject H0

Table 23: Second serve points won rate test results

Second serve performance is significantly higher in post-2006 era across all surfaces. This indicates an improved and more careful quality of the second serve in the modern era.

4.3 Discussion

The statistical outcomes demonstrate that in the professional era after 2006 the serve as a whole is competitive and successful compared with the era prior to 2006, on all surfaces. The post 2006 era on all three surfaces – hard, clay, and grass – have higher average ace rate, first serve in percentage, second serve points won percentage, hold percentage, but lower double-fault rate. Taken collectively the key performance indicators do not show that any particular technical aspect of the serve has changed; instead, it shows that a trend in serve performance has emerged. This implies that in the post 2006 era players are able to serve with higher intensity, more accuracy, and with more consistency.

This trend might be a result of tennis becoming more analytical and gains in serve performance being achieved through the use of information in tactical preparation and training. This outcome is in line with tennis analytics and data-related literature that has become more focused on the areas of serve location, serve speed, return position, point construction and the formation of tactical patterns. As these results hold true for all surfaces, it is likely that coaches and players are better attuned to the perfect serve in order to increase the probability of an ace or a service point win. In particular, in the post 2006 era the players show lower double-faults, showing that although they serve more aggressively they are simultaneously more concerned about avoiding risk and mistakes in the modern game, therefore modern players have a greater capability to balance aggression and control than their predecessors.

Greater first-serve consistency and greater second-serve quality can also be attributed to literature on workload monitoring, wearable technology, biomechanics, and technical optimization that has provided the coaching methods for pro players with an evidence-based nature as a result of the access of relevant and appropriate data. There are no significant differences between the eras on the other surfaces except the first-serve points won percentage on grass, which can be seen to mean that, as grass is by default the most serve-friendly surface, there was limited scope for improvement in this performance indicator in 2006. So, developments today are demonstrated through service metrics such as higher first-serve points won percentage, second-serve points won percentage and a higher hold rate.

The business impact of this is that tennis analytics should not be used as a mechanism to report but should instead form part of a decision support framework. In particular, service

key performance indicators enable coaches and data analysts to measure improvement across different service elements, in terms of serve effectiveness, second-serve points won percentage, and service hold rate. The application of specific performance indicators for decision-making supports the argument that collection of performance data is a critical resource that tennis federations, academies, and performance departments ought to invest in. Further, from a management information systems perspective, our study shows that data could be helpful in minimizing uncertainty and enabling more consistent and systematic decision-making in the competitive and high stakes domain of professional tennis. In summary, results of the study are supportive of my main proposition, a data environment is associated with tennis performance and tactic changes which are significant in a statistical sense, although we are to understand such changes as an association instead of causality.

5. Conclusion

This thesis attempts to answer the question of whether a data environment in men's tennis is associated with changes in performance and the outcomes of the matches. The sample is composed of data containing results of historical ATP matches. I adopted a pre–post design with a break in the time series in the year 2006. I showed that there were statistically significant differences between the two eras’ data. The post-2006 era is marked with higher ace rates, higher percentage of first serves in, higher percentages of second service points won, higher percentages of holds in service games and lower rate of double faults for all major court surfaces in the sample. Results suggest that in men’s tennis the more recent development is associated with an improvement in serve efficiency and an increase in dominance of service games, and an optimization of trade-off between aggressiveness and safety. Hence, it indicates that in the development of the data environment in tennis a change in performance and tactics has occurred.

Apart from the empirical contribution, this thesis also makes a conceptual contribution that puts tennis analytics in line with the business information systems tradition. More specifically, within the framework, analytics is understood as a technology as well as a capability that supports decision making that is more uniform, data-driven and stable in the

domain. The results indicate that data not only improved the measurement in the domain, but also changed measurement of results and outcomes, strategies and competitive advantage.

On the other hand, there are several limitations to the research. Most importantly, the pre-post design adopted in this thesis is observational and thus does not allow for identification of causal effects. Thus, the results in the paper should be understood as showing association of tennis results with the development of a data environment. In addition, the post-2006 era has coincided, not only with more ubiquitous availability of Hawk-Eye and data-based technologies, but also with the general development of training methods, physical training and tennis equipment. In addition, detailed data on results of serves are not available for all years or for all tournaments in the database, thus making it possible that the comparisons of results presented in the thesis are biased. Finally, my sample was limited to data that contains match results without detailed point-by-point data, thus not allowing one to conclude whether and how tactics have changed in terms of, for example, shot placement, direction of serves and rally length. Rather the results show that serving results and match scores have changed in accordance with the assumption that serving results and match score contain information on changes in serving tactics.

The present thesis can be extended in several ways. For example, it could be extended by using point-by-point databases or ball-tracking databases to analyze the history of tennis more directly. Possible extensions would be, among others, analyzing serving strategies (direction and type), rally patterns, positioning on return or net approaches. Another possible extension would be to use more sophisticated longitudinal statistical methods, which would allow for identification of causal effects. Finally, results of this work could be extended to other tennis environments, like for instance lower-tier tournaments, analysis of single tennis players or the women’s tour. The results of this thesis combined with results from other research on the effects of data and analytics on tennis’ competitive performance, development and strategies would contribute to a greater depth of information on effects of data and analytics.

References

- [1] A. Mohandas, M. Ahsan, and J. Haider, “Tactically Maximize Game Advantage by Predicting Football Substitutions Using Machine Learning,” *Big Data and Cognitive Computing*, vol. 7, no. 2, p. 117, Jun. 2023, doi: 10.3390/bdcc7020117.
- [2] B. Barthelemy *et al.*, “Impact of Technical-Tactical and Physical Performance on the Match Outcome in Professional Soccer: A Case Study,” *J. Hum. Kinet.*, Jul. 2024, doi: 10.5114/jhk/185933.
- [3] L. Forcher, L. Forcher, A. Woll, and S. Altmann, “AI in Bundesliga match analysis—expected possession value (EPV) vs. expected goals (xG) to predict match outcomes in soccer,” *Front. Sports Act. Living*, vol. 7, Nov. 2025, doi: 10.3389/fspor.2025.1713852.
- [4] J. Ehrlich and S. Sanders, “Player Tracking, Shot Charts, and the National Basketball Association’s Three-Point Revolution: A Technology-Based Misspecification in a Primary Sport Management Setting?,” *Journal of Sport Management*, vol. 39, no. 5, pp. 323–341, Sep. 2025, doi: 10.1123/jsm.2024-0072.
- [5] J. D. Fisher and C. Montague, “Improving the aggregation and evaluation of NBA mock drafts,” *J. Quant. Anal. Sports*, vol. 21, no. 4, pp. 327–343, Dec. 2025, doi: 10.1515/jqas-2023-0100.
- [6] S. Pradhan and D. Leshchinskii, “Violating the salary cap: exploring performance gains in the National Basketball Association,” *Front. Sports Act. Living*, vol. 7, Jul. 2025, doi: 10.3389/fspor.2025.1625458.
- [7] H. Wang, A. Sarker, and A. Hosoi, “The Effect of Basketball Analytics Investment on National Basketball Association (NBA) Team Performance,” *J. Sports Econom.*, vol. 26, no. 6, pp. 668–688, Aug. 2025, doi: 10.1177/15270025251328264.
- [8] Ryu Nagahara and Masataka Yamashita, “DEVELOPMENT OF SPRINT RUNNING PERFORMANCE OVER 4 YEARS: A CASE STUDY,” *ISBS Proceedings Archive*, vol. 43, no. 1, pp. 1–3, 2025.
- [9] R. Tedeschi, F. Giorgi, and D. Donati, “Sprint Training for Hamstring Injury Prevention: A Scoping Review,” *Applied Sciences*, vol. 15, no. 16, p. 9003, Aug. 2025, doi: 10.3390/app15169003.
- [10] E. van der Kruk, M. M. Reijne, B. de Laat, and D. (H. E. J.) Veeger, “Push-off forces in elite short-track speed skating,” *Sports Biomech.*, vol. 18, no. 5, pp. 527–538, Sep. 2019, doi: 10.1080/14763141.2018.1441898.
- [11] Z. Fan, L. Min, W. He, Y. Yang, W. Ma, and J. Yao, “Efficacy of multicomponent interventions on injury risk among ice and snow sports participants—a systematic review and meta-analysis,” *BMC Sports Sci. Med. Rehabil.*, vol. 16, no. 1, p. 135, Jun. 2024, doi: 10.1186/s13102-024-00921-6.
- [12] A. ; S. Fitzpatrick J. A.; Choppin S.; Kelley J., “Analysing Hawk-Eye ball-tracking data to explore successful serving and returning strategies at Wimbledon,” *Int. J. Perform. Anal. Sport*, 2024, doi: 10.1080/24748668.2023.2291238.
- [13] S. ; R. Mecheri F.; Mantel B.; Kauffmann F.; Benguigui N., “The serve impact in tennis: First large-scale study of big Hawk-Eye data,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2016, doi: 10.1002/sam.11316.

- [14] P. ; S. Tea T. B., “The analysis of serve decisions in tennis using Bayesian hierarchical models,” *Ann. Oper. Res.*, 2022, doi: 10.1007/s10479-021-04481-7.
- [15] T. C. Y. ; F. Chan D. S.; Fernandes C.; Kovalchik S., “A Markov process approach to untangling intention versus execution in tennis,” *J. Quant. Anal. Sports*, 2022, doi: 10.1515/jqas-2021-0077.
- [16] A. ; S. Fitzpatrick J. A., “How are elite tennis matches won at Wimbledon? A comparison of close and one-sided contests,” *Eur. J. Sport Sci.*, 2023, doi: 10.1002/ejsc.12063.
- [17] R. ; Z. Ye D.; Zhang M.; Liu W.; et al., “Nash equilibrium and tennis serve performance: a game theory analysis,” *Int. J. Perform. Anal. Sport*, 2023, doi: 10.1080/24748668.2023.2256120.
- [18] N. Buhamra, A. Groll, and A. Gerharz, “Comparing modern machine learning approaches and different forecast strategies on Grand Slam tennis tournaments,” *Journal of Sports Analytics*, vol. 11, Oct. 2025, doi: 10.1177/22150218251338954.
- [19] B. Li, Z. Deng, G. Gupta, J. Li, and Y. Miao, “Predicting tennis match outcomes mid-game using machine learning on psychological and physical data,” *J. Big Data*, vol. 12, no. 1, p. 159, Jul. 2025, doi: 10.1186/s40537-025-01216-4.
- [20] V. ; G. Hadžić A.; Filipčić A., “Validity and reliability of a novel monitoring sensor for the quantification of the hitting load in tennis,” *PLoS One*, 2021, doi: 10.1371/journal.pone.0255339.
- [21] T. ; R. Perri M.; Murphy A.; Howle K.; Duffield R., “Validating an algorithm from a trunk-mounted wearable sensor for detecting stroke events in tennis,” *J. Sports Sci.*, 2022, doi: 10.1080/02640414.2022.2056365.
- [22] T. ; R. Perri M.; Murphy A.; Howle K.; Duffield R., “Prototype machine learning algorithms from wearable technology to detect tennis stroke and movement actions,” *Sensors*, 2022, doi: 10.3390/s22228868.
- [23] T. ; D. Perri R.; Murphy A.; Mabon T.; Reid M., “Periodisation in professional tennis: A macro to micro analysis of load management strategies within a cluttered calendar,” *Int. J. Sports Sci. Coach.*, 2023, doi: 10.1177/17479541221091087.
- [24] W. J. ; H. Kraemer K.; Triplett-McBride N. T.; et al., “Physiological changes with periodized resistance training in women tennis players,” *Med. Sci. Sports Exerc.*, 2003, doi: 10.1097/00005768-200301000-00024.
- [25] J. P. ; M. Oliveira D. A.; Sampaio T.; Kilit B.; Morais J. E., “Impact of a six-week training program on physical fitness and performance of young tennis players: a cluster analysis approach,” *Front. Sports Act. Living*, 2025, doi: 10.3389/fspor.2025.1571019.
- [26] F. J. ; C. Peñalva-Salmerón M.; Martínez-Gallego R.; Ramón-Llin J.; Guzmán J. F., “External Load in High-Level Tennis Training: Influence of Game-Specific Drills in Junior and Professional Players Across Playing Situations,” *Applied Sciences*, 2026, doi: 10.3390/app16010492.
- [27] J. ; M. Lambrich T., “Biomechanical analyses of different serve and groundstroke techniques in tennis: A systematic scoping review,” *PLoS One*, 2023, doi: 10.1371/journal.pone.0290320.
- [28] J. ; G. Jacquier-Bret P., “Kinematics characteristics of key points of interest during tennis serve: a systematic review and meta-analysis,” *Front. Sports Act. Living*, 2024, doi: 10.3389/fspor.2024.1432030.

- [29] A. V. ; F. Brito P.; Costa M. J.; Cardoso R.; Santos C. C.; Fernandez-Fernandez J.; Fernandes R. J., “The Influence of Kinematics on Tennis Serve Speed: An In-Depth Analysis Using Xsens MVN Biomech Link Technology,” *Bioengineering*, 2024, doi: 10.3390/bioengineering11100971.
- [30] B. ; G. Montalvan V.; Ramos-Pascual S.; van Rooij F.; Saffarini M.; Nogier A., “Epidemiology of Musculoskeletal Injuries in Tennis Players During the French Open Grand Slam Tournament From 2011 to 2022,” *Orthop. J. Sports Med.*, 2024, doi: 10.1177/23259671241241551.
- [31] M. C. ; E. Fu T. S.; Renstrom P. A.; Windler G. S.; Dines D. M., “Epidemiology of injuries in tennis players,” *Curr. Rev. Musculoskelet. Med.*, 2018, doi: 10.1007/s12178-018-9452-9.
- [32] B. ; C. Minghelli J., “Epidemiology of musculoskeletal injuries in tennis players: risk factors,” *J. Sports Med. Phys. Fitness*, 2019, doi: 10.23736/S0022-4707.19.09842-6.

Author’s Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.

