

BIVARIATE MODELING OF TENNIS FIRST SERVE DEPENDENCE USING COPULA APPROACH: INSIGHTS FROM GRAND SLAM TOURNAMENTS

*(Permodelan Pergantungan Servis Pertama Tennis Menggunakan Pendekatan Copula:
Pandangan daripada Kejohanan Grand Slam)*

NUR AMIRA SARIATY RUSLAN, UMMUL FAHRI ABDUL RAUF* & ZURAINI ZAINOL

ABSTRACT

Modeling performance in tennis has traditionally relied on multivariate statistical techniques, often assuming normality in data distribution. However, these assumptions may not always hold, particularly in the case of binary match outcomes and interdependent performance metrics such as service and return success rates. This study introduces a copula-based approach to model the dependency structure between first serve performance and match outcomes using data from major Grand Slam tournaments: US Open, Australian Open, French Open, and Wimbledon. Applying various marginal distributions (Normal, Log-Normal, Gamma, Weibull, and Logistic) and selecting the best fit using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), we construct an optimal copula model for each tournament. We explore elliptical and Archimedean copulas to capture the underlying dependence between serve success rates and match results. Our findings indicate the Frank copula best fits all tournaments, demonstrating a significant positive dependence between serve performance and winning probability. The novelty of this study lies in the application of copula modeling to tennis match analytics, a domain where it has received limited attention, particularly in the context of women matches. Unlike conventional multivariate models, copulas allow for flexible dependence structures without requiring normality assumptions, providing a more accurate representation of real-world data. This approach offers valuable insights for coaches, analysts, and sports strategists by quantifying the impact of serve efficiency on match success. Our findings suggest that optimising first serve performance and return capabilities can significantly enhance a player's likelihood of winning, offering a data-driven framework for strategic decision-making in professional tennis.

Keywords: copula model; tennis match analytics; first serve performance; match outcome; dependence structures

ABSTRAK

Permodelan prestasi dalam tenis secara amnya bergantung kepada teknik statistik multivariat, selalunya menganggap data mempunyai taburan secara normal. Walau bagaimanapun, terdapat kemungkinan andaian ini tidak selalunya tepat, khususnya untuk keputusan perlawanan binari dan metrik prestasi yang saling bergantung seperti kadar kejayaan servis dan pukulan balas. Kajian ini memperkenalkan pendekatan berasaskan kopula untuk permodelan struktur pergantungan antara prestasi servis pertama dan hasil perlawanan menggunakan data daripada kejohanan Grand Slam; Terbuka AS, Terbuka Australia, Terbuka Perancis and Wimbledon. Menggunakan pelbagai taburan marginal (Normal, Log-Normal, Gamma, Weibull dan Logistik) dan memilih kriteria yang paling sesuai menggunakan Kriteria Maklumat Akaike (AIC) dan Kriteria Maklumat Bayesian (BIC), kami membina model kopula yang optimum untuk setiap kejohanan. Kami mengkaji kopula elips dan Archimedean untuk menerangkan pergantungan asas antara kadar kejayaan servis dan keputusan perlawanan. Dapatan kami menunjukkan kopula Frank yang paling sesuai untuk semua kejohanan dengan mempamerkan pergantungan positif yang ketara antara prestasi servis dan kebarangkalian menang. Kebaharuan kajian ini

adalah aplikasi pemodelan kopula untuk analisis perlawanan tenis, domain di mana ia telah mendapat perhatian yang terhad, terutamanya dalam konteks perlawanan tenis wanita. Tidak seperti model multivariat konvensional, kopula membenarkan struktur pergantungan yang fleksibel tanpa memerlukan andaian normal, memberikan perwakilan data dunia sebenar yang lebih tepat. Pendekatan ini menawarkan dapatan bermakna kepada jurulatih, penganalisis dan ahli strategi sukan dengan mengukur kesan kecekapan servis terhadap kejayaan perlawanan. Penemuan kami mencadangkan bahawa mengoptimumkan prestasi servis pertama dan teknik pukulan balas yang baik mampu memberikan peningkatan ketara kemungkinan pemain untuk menang, menawarkan rangka kerja dipacu data untuk membuat keputusan strategik dalam tenis profesional.

Kata kunci: model kopula; tenis analisis; prestasi servis pertama; kejayaan perlawanan; struktur pergantungan

1. Introduction

In sports analytics, a few sports have received analytical attention in increasing player performance and enhancing player tactical and technical aspects to obtain the prizes offered if the player can win the match. Despite the global popularity of tennis, most analytical studies have concentrated on men matches, leaving women tennis comparatively underexplored. Most existing tennis analytics research has focused on men matches and relied on regression based models that assume independence between performance variables. This leaves a gap in understanding the dependencies such as serve and return dynamics in women tennis match. Tremendous methods have been developed to analyse tennis matches, which will be discussed further in the literature section. Most of the previous papers focus in depth on men tennis, while women tennis has received less research attention. A regression model is one of the most popular methods for exploring this kind of data. Traditional approaches often designate the regression model for easy evaluation and testing, with clear interpretation. Various regression methods, such as logistic, log-linear, normal, and tobit models, can handle discrete, continuous, and mixed performance metrics. Traditional methods focus on individual performance measurements. However, they frequently fail to capture the complex relationship between multiple factors that might influence the match outcomes, including a player's mental state, playing surfaces, and the interaction between shot selection and strategy. These traditional methods are insufficient because they oversimplify tennis's complexity by neglecting to take into account dependencies or correlations between variables. For example, player fatigue may influence shot selection, affecting performance in ways that normal statistics cannot fully reflect. We found that most models tend to ignore the dependencies between the variables and assume that the data follows a normal distribution. Regression models, on the other hand, are generic types that typically give equal weight to all historical observations. Percy (2015) stated that a few stochastic models by assuming that the random vector $\mathbf{x}$ has a conditional probability mass or density function $f(\mathbf{x}|\theta)$ given a parameter vector θ with prior probability density function $g(\theta)$. They left an interesting discussion of how to select from a variety of copula forms and how to dynamically alter them to simulate dependence on random processes, which has attracted our attention to fill in the research gap mentioned. Upon examining the other studies, we observed that none of the existing research addressed the remaining question. Hence, we endeavored to address the gap identified by investigating the copula method in relation to tennis serve performance and match outcomes.

Since its initial introduction by Sklar (1959), copula theory has evolved into an essential tool for examining multivariate data and describing the interactions between random variables.

Copulas, which are vital multivariate functions that allow marginal distributions to be separated from the underlying dependence structure, serve as the foundation for this theory. Understanding the dependence structure is sometimes relevant when analysing high-dimensional data to achieve dimension reduction. The copula method has the advantage of reducing a multivariate problem to several univariate problems and has the advantage of retaining symmetry if required (Percy 2015).

The advent of copulas in sports statistics has transformed how analysts and fans perceive and assess player performance, team dynamics, and overall strategy. Copulas are multivariate probability distribution functions that describe the relationship between random variables in sports statistics. They connect multivariate distribution functions to one-dimensional marginal distribution functions, allowing for the modeling and analysis of the dependent structure between variables (Sulistianingsih & Satyahadewi 2025). Copulas help sports analysts model and evaluate the associations between multiple variables influencing performances, such as individual statistics, environmental conditions, team plans, and opponent strengths. They also enable analysts to better capture complicated relationships and correlations between various elements of the game.

In tennis analytics, copulas have been a game-changing development, providing a more comprehensive understanding of the complex interactions between factors affecting player performance and match results. They provide an advanced statistical framework for analysing the relationships between different aspects of a player's game and reveal the complex dynamics and strategies occurring on the court. By employing copulas, analysts studying singles matches can evaluate the combined distribution of important performance measures, allowing for an in-depth evaluation and understanding of their interactions and effects throughout the match.

Tennis analysts can assess the relationships between many elements of a player's game and discover winning patterns or techniques via copulas. Copulas, for example, allow analysts to examine the relationship between a player's effectiveness when serving and their capacity to win important points or games, taking into account other variables such as opponent strength, court surface, and match conditions. Copulas can also highlight behavioral trends in players, such as a propensity to perform better under duress or to specialise in particular game mechanics. Copulas also make it easier to analyse the dynamics of a match over time, recording changes in momentum, tactical tweaks, and the effect of mental toughness or exhaustion on performance. Coaches, players, and tennis lovers can enhance their comprehension of match strategies and create more successful game plans to maximise performance and success on the singles court by discovering these insights.

Based on the previous literature searches, Gorgi *et al.* (2019) mentioned in their study that during recent years, there has been a lot of interest in modeling and predicting tennis match results. Yu and Wang (2023) employed a logistic regression model to evaluate the strategies and techniques in the 2021 US Open men's singles, testing ten potential parameters and selecting the one that followed the normal distribution for the model development. This indicates that tennis has received great attention from researchers in finding any desired outcome related to the tennis data. Statistical models can be useful for describing the key elements of tennis matches and determining the degree of ability of tennis players in various conditions based on (Klaassen & Magnus 2001). Recently, there have been multiple attempts to use player data to forecast tennis match results. Barnett and Clarke (2005) indicate that the points scored on serve and during the reception of a serve varies fluctuate across matches and can be modeled as Gaussian distributed random variables. In their study, they perform Monte Carlo simulations based on these Gaussian distributed input variables to calculate the probability density functions for each player to win a match. They develop a stochastic Markov

chain model based on the realisation that a player's probability of winning a point on serve is not constant throughout a tournament but varies from match to match and is better modeled as a random variable whose probability density function closely resembles a Gaussian (around the sample mean).

According to Kovalchik (2016), point-based tennis forecasting models specify the probability of winning a point on serve and return. Using an independent identically distributed (IID) assumption for point outcomes, the chance of winning a match is a function of the probabilities of winning a point on serve and return, which may be expressed in closed form. Furthermore, Palayangoda *et al.* (2022) have been applying copula methods to analyse the bivariate data of runs scored and balls encountered by the cricket player. They implemented distributions such as Gamma, Weibull, and exponential, but found the Gamma distribution to be the most suitable distribution generally for all the cases, and they assumed the marginal distributions of both parameters follow Gamma distributions. Thus, they estimate parameters from the maximum likelihood approach using Gamma marginals and Frank copula. Within the copula technique, it is crucial to note that copula functions have the flexibility to select any continuous distributions as the marginals, depending on the goodness of fit of the data used.

Adapting to the previous literature search, we attempted to incorporate a few marginal distributions into our tennis dataset to enhance the accuracy of the model and tried to fit different copula methods to estimate the parameter. Copula applications have been demonstrated to be a helpful tool for assessing cricket player performance, therefore, Palayangoda *et al.* (2022) proposed that their article may raise interest in the topic and create an opportunity for further research. Hence, this work aims to propose a bivariate copula model using additional marginal distribution and copula function that best fits the serve and return of serve in the tennis first serve. This present tennis study will utilise data obtained from the Grand Slam Tournament held in the year 2022 for women match because assessing the previous research found that women tennis is underexplored compared to men. The event included in the analysis are Australian Open, French Open, Wimbledon, and US Open, which took place throughout the year. This paper will focus on the tennis first serve, as this was the most important variable that was selected using the feature extraction process. The main objective of this work is to develop a bivariate copula model that relates the serve and return serve of the first serve to the outcome of the match. The bivariate copula will be evaluated between serve and return serve indicated by the percentage of first serve won (PFSW) and percentage of first serve return won (PFSRW) on the tennis match outcome. The primary goal of this study is to establish the bivariate copula model between serve and return serve of first serve towards the match outcome in the different tournaments.

This study fills a gap in the literature by focusing on the relationship between serve and return dynamics of the tennis first serve using data from the 2022 Grand Slam tournaments. The proposed bivariate copula model offers a more thorough understanding of match outcomes by emphasising these important performance aspects, particularly on the effectiveness of the tennis first serve. This study builds on previous research in tennis analytics by using copula-based modeling to explore the interdependence between first serve and return of serve performance in relation to match outcomes. Beyond the court, this work may result in the development of prediction tools that analysts and tournament organisers may use to predict match outcomes, improving the viewing experience for both stakeholders and spectators.

Following is the organisational structure of the paper. Section 2 provides an overview of the methodology used. Section 3 presents the results and discussion, and Section 4 gives the conclusion of the research presented.

2. Methodology

2.1. Data description

This subsection presents a comprehensive description of the variables utilised in the tennis dataset collected from the Grand Slam tournaments.

Table 1: Description of tennis variables

Variable	Definition	Description	Unit	Datatype
PFSW	Percentage First Serve Won	The proportion of points won by a player when the served is successfully executed	Percentage (%)	Continuous
PFSRW	Percentage First Serve Return Won	The proportion of points won by a player when returning the opponent's first serve	Percentage (%)	Continuous
PSSW	Percentage Second Serve Won	The proportion of points won by a player on second serve attempts	Percentage (%)	Continuous
PSSRW	Percentage Second Serve Return Won	The proportion of points won when returning an opponent's second serve	Percentage (%)	Continuous
Aces	Aces	The total count of serves that results in immediate points without being touched by the opponent	Count	Discrete
UE	Unforced Error	The total number of errors occurred by a player in a non-pressured situation	Count	Discrete
Status	Match Status	A binary outcome that represent the match outcome of the player either win or lose	Categorical	Binary

Table 1 consists of all the parameters involved in the Grand Slam Tournaments involving the serve and return serve performance of the players. The output variable is the match outcome, while the others are the input variables. The dataset used in this study was obtained from the official Women's Tennis Association (WTA) match statistics database and the respective Grand Slam tournament websites. These sources provide detailed match-level serve performance statistics for professional women's tennis players. Prior to analysis, the dataset was carefully cleansed and pre-processed to ensure data quality is fulfilled. Matches with missing or incomplete information were excluded, and all variables were verified for consistency. The final dataset represents a complete and reliable sample for the purpose of the analysis.

2.2. Copula theory

Copulas were first introduced by Sklar (1959), which was used to join the univariate distribution function, hence creating a multivariate distribution function. Copula has been extensively applied in hydrology, medical science, construction, accident analysis, insurance, geosciences, irrigation, safety analysis, chemical engineering, banking, artificial intelligence, weather, climate change, biometrics, and the aging population (Bhatti & Do 2019; Almetwally *et al.* 2024). According to Dewick and Liu (2022), parametric copula modeling can be carried out systematically, beginning with identifying and determining the model's distributions, then transforming the distributions to marginal distributions (typically uniform), and finally selecting copulas. Another advantage of the copula is its utility in understanding tail dependencies in risk and survival contexts.

From the previous literature search, we know that the copula method is widely used in the finance (Dewick & Liu 2022), climate (Chen *et al.* 2024), engineering, industry, economics, technology, weather, drought, sports, medical science (Liu *et al.* 2024) and many more (Almetwally *et al.* 2024). To convey the methodology used for the sports data in which the outcome is in binary response, either 'win' or 'lose', the tennis point to the opponent was described using the copula method instead of the commonly traditional approach of the

multivariate model. We tried to apply the copula method, which is described in terms of the probability approach, to demonstrate that the method used is also suitable for the binary response instead of continuous and count data. The data used was from a tennis match in the Grand Slam Tournament, consisting of the US Open, Australian Open, French Open, and Wimbledon for the year 2022.

To understand more, let X and Y be two random variables with joint distributions of $F_{X,Y}(x, y)$ and marginal cumulative distribution function $F_X(x)$ and $F_Y(y)$ respectively. According to Sklar theorem (Nielsen 2006), there exists a copula function $C(x, y)$ defined on the unit square such that

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \quad (1)$$

This copula function is unique if both $F_X(x)$ and $F_Y(y)$ are continuous. Conversely, given $F_X(x)$ and $F_Y(y)$ and a copular function $C(x, y)$, then Eq. (1) defines a joint distribution function with marginal distribution $F_X(x)$ and $F_Y(y)$ respectively.

If the probability density function (pdf) $f_X(x)$ and $f_Y(y)$ of X and Y exist, then the joint pdf of the two random variables can be expressed as

$$f_{X,Y}(x, y) = c(F_X(x), F_Y(y))f_X(x)f_Y(y) \quad (2)$$

where $c(u, v)$ is defined as

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v} \quad (3)$$

2.3. Marginal distributions

Finding the appropriate marginal distribution for each serve characteristic is a crucial step in fitting copula for tennis outcome characteristics. In this study, the first serve won and first serve return won outcome characteristics are taken into consideration. The distribution functions tested in this study are Normal, Log Normal, Gamma, Weibull and Logistic distributions. The following are the equations for the probability density functions (pdfs) of the five distributions and their domains.

2.3.1. Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x > 0 \quad (4)$$

where μ is the mean and σ is the standard deviation;

The normal distribution is typically employed when data is symmetrically distributed around a mean and tends to cluster in the center with a few extreme outliers on either end. This is especially effective for modeling performance measurements that do not have considerable skew or heavy tails. In tennis, if a player's serve accuracy or consistency is relatively stable and follows a typical "bell curve" pattern, the normal distribution will appropriately model this behavior.

2.3.2. Log-Normal Distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0 \quad (5)$$

where μ is the mean and σ is the standard deviation of $\ln(x)$.

The log-normal distribution is used when the data is skewed, particularly when the data represents multiplicative processes or is bounded at 0. For example in tennis, the percentage of first serve points won, often exhibit a skewed distribution where extreme values are less frequent but still significant.

2.3.3. Gamma Distribution

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} e^{-\frac{x}{\beta}} x^{\alpha-1}, x > 0 \quad (6)$$

where α is the scale parameter and β is the shape parameter.

Gamma distribution is often applied to continuous, positive data with a skew, especially when modeling the time to an event or counting the number of occurrences until a certain threshold is reached. It is flexible and can model data that does not follow a simple symmetric distribution, which is common in tennis performance data. For example, the number of points won on serve could be well-modeled by the gamma distribution, especially if these events are counted over time or across several points in a match.

2.3.4. Weibull Distribution

$$f(x) = \left(\frac{a}{b}\right) \left(\frac{x}{b}\right)^{a-1} e^{-\left(\frac{x}{b}\right)^a}, x > 0 \quad (7)$$

where a is the scale parameter and b is the shape parameter.

Weibull distribution is often used in modeling “time-to-event” data or survival analysis, particularly when the event rate is not constant over time. This distribution is especially useful for data that can exhibit varying hazard rates, making it highly relevant for tennis match analysis, where the probability of winning a point on serve or return may change depending on match conditions such as player fatigue or shifting in momentum. For example, a player’s chances of winning on first serve may decrease if they are fatigued or under pressure during critical moments of the match, and the Weibull distribution can capture these varying hazard rates over time.

2.3.5. Logistic Distribution

$$f(x) = \frac{e^{-\frac{(x-\mu)}{s}}}{s \left(1 + e^{-\frac{(x-\mu)}{s}}\right)^2}, x > 0 \quad (8)$$

where μ is the location parameter and s is the scale parameter.

Logistic regression is widely used when the outcome of interest is binary, such as the probability of winning or losing a point on serve or return. This approach models the likelihood of a binary outcome as a function of various predictors. In tennis, logistic regression is particularly relevant for modeling outcomes like the probability of winning a point on serve or the likelihood of successfully returning a serve. The model can help capture how various factors influence the probability of winning a point. For example, a player with a strong first serve will likely have a higher probability of winning a point on serve, and logistic regression helps quantify this relationship.

2.4. Copula families

The other step involved in fitting copulas is estimating the parameters of the copula. Two types of different copula families will be implemented in this study which are elliptical copula and Archimedean copula. Archimedean copula has gained greater attention in numerous applications due to the ability to represent multivariate joint distribution with a limited number of parameters. Archimedean copulas were selected due to their flexibility in modeling bivariate dependence structures, especially when both positive and negative dependence between variables are important (Nelsen 1999). Additionally, Archimedean copulas are versatile and simple to build. Examples of Archimedean copulas used in this paper are Frank, Clayton, and Gumbel. The symmetric characteristics of the Frank copula and its ability to estimate the joint distribution using only one parameter estimation have made it a popular choice in the literature (Palayangoda *et al.* 2022). Besides being able to cater the symmetric data, it is also suitable in which cases of positive and negative correlation were observed. Furthermore, Archimedean copulas have been widely applied in other areas such as hydrology (Genest & Favre 2007), reliability analysis (Joe 2014), and sports (Palayangoda *et al.* 2022) demonstrating their robustness and interpretability in low dimensional dependence modeling.

Additionally, this paper will present the application of the elliptical and Archimedean copulas family in determining the marginal distribution of the tennis serve indicators between the first serve to find out the existing relationship. The elliptical copula is defined via Sklar's theorem as the dependence structure of a related elliptical distribution and obtained from the respective multivariate distribution function by standardising the univariate marginal laws. The most often used copulas are elliptical. In essence, the elliptical copula models are those in which an elliptical distribution joins the univariate margins. Equivalent to all other copulas, elliptical copulas are expressed in the following form:

$$H(x) = C(F_1(x_1), \dots, F_d(x_d)) \quad x \in \mathbb{R}^d \quad (9)$$

where, H is the multivariate elliptical distribution function and $F_1, \dots, F_d$ correspond to the univariate margins.

There are two types of elliptical copulas which are Gaussian or Normal copula and Student t copula. The Normal copula is implemented in such a way that all univariate marginal distributions are normal, and the multivariate joint distribution is a multivariate-normal distribution. Normal copula do not model tail dependencies very well as their tails are flat. A Normal copula is defined as

$$C_{\Sigma}(u, v) = \theta_{\Sigma}(N^{-1}(u), N^{-1}(v)) \quad (10)$$

where Σ is the covariance matrix which is the parameter of the Gaussian copula and N^{-1} is the quantile (inverse cdf) function.

The Student t copula is implemented in such a way that all univariate marginal distributions are student, and the multivariate joint distribution is a multivariate student distribution. One thing to note is that student t copulas have fatter tails as compared to Normal copulas. A student t copula is defined as

$$C_{\Sigma, \nu}(u, v) = t_{\Sigma, \nu}(t_{\nu}^{-1}(u), t_{\nu}^{-1}(v)) \quad (11)$$

where Σ and ν are the covariance matrix and degrees of freedom which describe the Student t copula and t_{ν}^{-1} is the quantile (inverse cdf) function.

Akin to all other copulas, Archimedean copulas are defined by its generator $\varphi(\cdot)$, a continuous strictly decreasing function from $[0,1]$ to $[0,\infty)$ such that $\varphi(1) = 0$. If $\varphi^{-1}(\cdot)$ represents the inverse function of $\varphi(\cdot)$, the Archimedean copula is defined by

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) \quad (12)$$

Three types of Archimedean copulas will be used in this paper, which are Clayton, Frank, and Gumbel copulas. Clayton copula is a copula that allows any specific non-zero level of (lower) tail dependency between individual variables. This copula function captures the stronger dependence in the lower tail. In tennis match, it is useful when poor first serve performance is strongly linked to losing a match. In example, if a player's serve percentage drops significantly, their chance of losing rised sharply. Clayton copula is thus suited for investigating the poor performance that happens in tennis. A Clayton copula is defined as

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \quad (13)$$

In the bivariate case of Frank copula, its parameters can interpolate between a lower limit of $-\infty$ (countermonotonicity) and an upper limit of ∞ (comonotonicity). Frank copula is symmetric, which is able to capture both positive and negative dependence. In tennis, it can fit situations where both good and bad serve-return combinations reflect in a balanced way. This often can happen in real tournament data, in which both advantage and disadvantage in serving affect outcomes, but not only at the extremes. Frank copula can manage both situations, making it ideal for balanced dependency in serve-return dynamics. A Frank copula is defined as

$$C_\theta(u, v) = -\theta^{-1} \log \left(\frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right) \quad (14)$$

Next, the Gumbel copula is the copula that allows any upper tail dependency between the individual variables and is exchangeable. It captures stronger dependence in the upper tail. In tennis, it is useful when a dominant serve performance is strongly tied to winning. For example, Serena Williams, during her peak form, when she serves extremely well, her likelihood of winning is almost guaranteed. Gumbel copula is excellent for analysing peak-performance scenarios in which winning serve and match points are closely associated. The Gumbel copula is defined as

$$C_\theta(u, v) = \exp \left[-((-\log u)^\theta + (-\log v)^\theta)^{1/\theta} \right] \quad (15)$$

Maximum Likelihood Estimation (MLE) method is one of the commonly used methods for estimating the parameters of these marginal distributions. The method used for the estimation is called Maximum Pseudo-Likelihood (MPL). Two parametric estimate techniques, namely MLE and Inference for Margin (IFM) with Pseudo-maximum Likelihood (PML), were compared by Kim *et al.* (2007). They demonstrated that when IFM and MLE were employed as the techniques, margin misspecification had a major effect on parameter estimates. However, PML is preferred over the other two approaches because it is resilient to the misspecification of margins. When dealing with multivariate multiparameter copulas, however, MLE is the most effective method as well as the most effective method overall. They suggested implementing the PML approach. Hence, in this paper, despite using a commonly used method of IFM, we are trying to implement the PML method to estimate the parameter.

Assume that the marginal cdfs of X and Y , depending on unknown vector parameters α_1 and α_2 , are $F_1(x; \alpha_1)$ and $F_2(y; \alpha_2)$ respectively. Denote the corresponding pdfs by $f_1(x; \alpha_1)$ and $f_2(y; \alpha_2)$. Let $\hat{\alpha}_1$ and $\hat{\alpha}_2$ be the MLEs of α_1 and α_2 obtains from their marginal distributions.

The joint pdf of a copula-based parametric model for a random vector $Z = (X, Y)$ is

$$f(z; \alpha_1, \alpha_2, \theta) = c(F_1(x; \alpha_1), F_2(y; \alpha_2); \theta) f(z; \alpha) \quad (16)$$

where $f(z; \alpha) = f_1(x; \alpha_1), f_2(y; \alpha_2)$. The log-likelihood function, given observed sample of $z_1, z_2, \dots, z_n$ is

$$L(\theta, \alpha_1, \alpha_2) = \sum_{i=1}^n \log f(z_i; \alpha_1, \alpha_2, \theta) \quad (17)$$

The MPL estimators of the copula are obtained by maximising Eq. (17) with α_1 and α_2 replaced by $\hat{\alpha}_1$ and $\hat{\alpha}_2$, with respect to θ i.e. by solving the following first order conditions and showing that the corresponding Hessian matrix is negative definite:

$$\frac{\partial L(\theta, \hat{\alpha}_1, \hat{\alpha}_2)}{\partial \theta} = 0 \quad (18)$$

Suppose $X_i = (X_{i,1}, \dots, X_{i,p})$ are random vectors and R_{ij} the rank of X_{ij} . Then by applying $\hat{U}_{ij} = \frac{R_{ij}}{n+1}$ the pseudo observations can be calculated. Therefore, the pseudo log-likelihood is given by

$$\log L(\theta) = \sum_{i=1}^n \log c_\theta(\hat{u}_i) \quad (19)$$

where $\hat{u}_i = (\hat{U}_{i,1}, \dots, \hat{U}_{i,p})$. For a bivariate copula this simplifies to

$$L(\theta) = \sum_{i=1}^n \log \left\{ c_\theta \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \right\} \quad (20)$$

Table 2: Probability density function of distribution

Family of Copula	$\mathcal{C}_*(\mathbf{u}, \mathbf{v})$	Generator $g(t)$	Kendall tau, τ	Parameter Space
Normal	$\Phi_\theta[\Phi^{-1}(u), \Phi^{-1}(v)]$	$\frac{1}{\sqrt{2\pi}} e^{-t/2}$	$\frac{2}{\pi} \arcsin \theta$	$\theta \in (-1, 1)$
Student t	$T_{\theta, v}[T_v^{-1}(u), T_v^{BV-1}(v)]$	$\frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{\pi v}} \left(1 + \frac{t}{v}\right)^{-\frac{2+v}{2}}$	$\frac{2}{\pi} \arcsin \theta$	$\epsilon(2, \infty)$
Gumbel	$\exp\left(-[(-\ln u)^\theta + (-\ln v)^\theta]^{\frac{1}{\theta}}\right)$	$(-\ln t)^\theta$	$1 - \frac{1}{\theta}$	$\theta \in (1, \infty)$
Clayton	$[max(u^{-\theta} + v^{-\theta} - 1, 0)]^{\frac{-1}{\theta}}$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$\frac{\theta}{\theta + 2}$	$\theta \in (0, \infty)$
Frank	$\frac{-1}{\theta} \ln \left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right]$	$-\ln \frac{(e^{-\theta t} - 1)}{(e^{-\theta} - 1)}$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$	$\theta \in \mathbb{R}$

The MPL method is a semi-parametric approach, utilising both parametric and nonparametric processes. Similar to the IFM approach, the MPL method consists of two steps. The copula parameter is computed using the maximum likelihood method, a parametric

approach. However, the u_i and v_i are computed using order statistics from each sample data, a nonparametric approach. The MPL method is classified as semiparametric due to its use of both parametric and nonparametric approaches. In bivariate conditions, Eq. (21) provides the log-likelihood function of the MPL approach. The copula parameter, θ is derived by maximising the log-likelihood function.

$$LL = \sum_{i=1}^n \log c \left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) = \sum_{i=1}^n \log c (u_i, v_i) \quad (21)$$

given the R_i and S_i are the ranks of each variable X and Y .

According to Joo *et al.* (2020), the MPL approach applies the empirical distribution function using Weibull plotting position formula to each marginal distribution. The strategy may minimise the impact of incorrectly assuming the marginal distribution type during parameter estimation of the copula.

3. Results and Discussion

This section describes the results and discussion based on the analysis obtained in Section 2. The details of this section explained the scatterplot for each of the tournaments in the Grand Slam Tournament for the tennis first serve distribution of serve and return serve. Next, we attempted to fit in with a few marginal distribution functions and copula functions to identify the most suitable distribution for each of the tournaments that we could combined using copula functions.

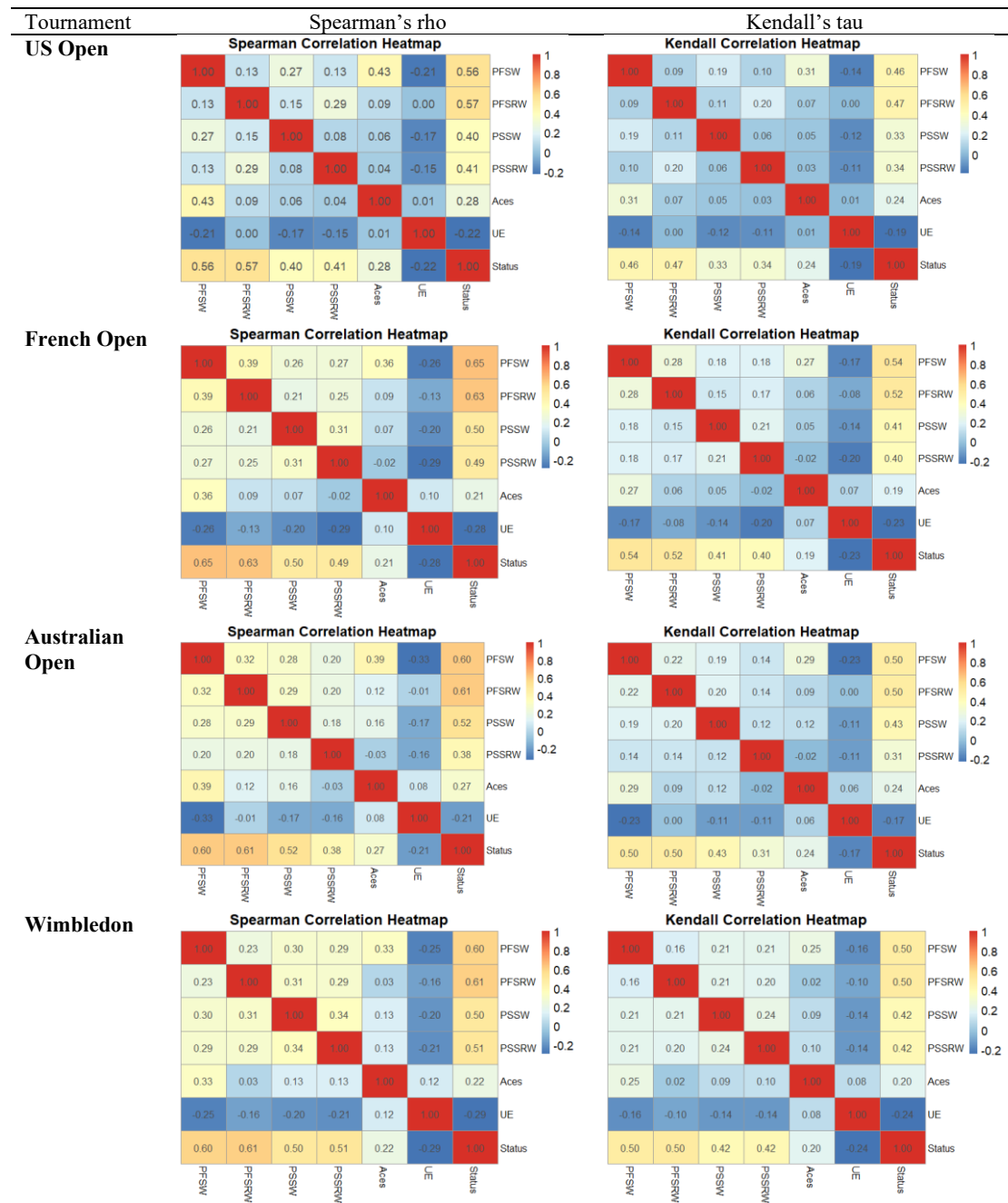
3.1. Correlation

The independent variables for the tennis serve performance had been screened previously using the feature selection techniques, and the top five parameters out of all parameters were chosen (Ruslan *et al.* 2024). Out of the parameters of serve extracted from the dataset, such as percentage first serve in, percentage first serve return won, percentage first serve won, percentage second serve won, percentage second serve return won, double faults, aces, first serve speed, and second serve speed, only the top five are chosen. Table 3 displays the correlation coefficients in the form of a heatmap for all serve performance variables that are correlated to each other. From there, we choose the highest correlated value as the best with the most significance ($p\text{-value} < 0.05$). According to the correlation values, the percentage of first serve return won (PFSRW) is the most correlated variable with the match outcome at the US Open, Wimbledon, and Australian Open. The percentage of first serves won (PFSW) has the highest correlation with match status at the French Open. Based on the results, we can conclude that the parameter first serve, which is the ability to control the first serve well and return the server's serve better, has an advantage in obtaining a match point over the opponent. These findings highlight important tactical differences across tournaments. At the French Open, which was played on a clay court, serve dominance is less pronounced, and extended rallies are common. Thus, the ability to consistently win points behind the percentage first serve won (PFSW) is critical for maintaining control of service games. By contrast, Wimbledon, US Open, and Australian Open are surfaces that reward faster play, which implies that success on return of first serve (PFSRW) becomes more decisive, as breaking serve can swing momentum quickly in shorter rallies. This demonstrates that both serve execution and return resilience are important, but also depend on the surface and the player's ability to control their serve better.

Since this work used both continuous and ordinal data, the appropriate correlation values to be used are Spearman's and Kendall's values (El-Hashash & Shiekh 2022). The assumptions of

Spearman are that because the two variables are measured on an ordinal scale, the scores on one must be monotonically connected to the other, and there are no ties observed (Etaga *et al.* 2021; Temizhan *et al.* 2022). Kendall's tau is a non-parametric measure, similar to Spearman's, in which it measures the relationship based on the difference between the probability of concordance and discordance of two observed variables, X and Y . Kendall's tau correlation coefficient examines the relationship between two ordinal variables and follows the same assumptions as the Spearman rank correlation

Table 3: Heatmap of the top 5 parameters using Spearman's rho and Kendall's tau



3.2. Marginal distribution

Table 4: Marginal distribution for tennis first serve and match outcome

Tournament		Normal		Log Normal		Weibull		Gamma		Logistic		Bernoulli
		$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{b}$	$\hat{a}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\mu}$	$\hat{s}$	$\hat{p}$
(i) Tennis First Serve												
PFSW	French Open	62.2284	10.0852	4.1174	0.1655	6.6239	66.5326	37.3347	0.6000	62.0969	5.8051	
PFSRW	US Open	33.9960	10.4454	3.4721	0.3457	3.6087	37.7385	9.3941	0.2763	33.9606	6.0564	
	Australian Open	35.1142	11.2928	3.4966	0.3755	3.4711	39.0568	8.2231	0.2342	35.2245	6.5772	
	Wimbledon	33.9567	9.92304	3.4754	0.3321	3.8261	37.5583	10.2185	0.3009	34.0656	5.7124	
(ii) Match Outcome												
a. Status	French Open											0.5000
	US Open											0.5000
	Australian Open											0.5000
	Wimbledon											0.5000

3.3. Goodness of fit parameter estimates

Table 5: Goodness of fit for parameter estimates of tennis first serve and match outcome

Parameter	Tournament	Distribution	Likelihood	AIC	BIC
PFSW	French Open	Normal	-947.42	1898.85	1905.92
		Log Normal	-949.34	1902.69	1909.76
		Weibull	-954.72	1913.43	1920.51
		Gamma	-947.62	1899.25	1906.32
PFSRW	US Open	Logistic	-950.94	1905.88	1912.95
		Normal	-948.81	1901.61	1908.67
		Log Normal	-964.87	1933.74	1940.80
		Weibull	-947.76	1899.53	1906.59
	Australian Open	Gamma	-954.75	1913.50	1920.56
		Logistic	-953.40	1910.81	1917.87
		Normal	-976.15	1956.30	1963.37
		Log Normal	-999.74	2003.48	2010.55
	Wimbledon	Weibull	-974.68	1953.35	1960.43
		Gamma	-986.12	1976.23	1983.31
		Logistic	-981.37	1966.74	1973.82
		Normal	-943.30	1890.61	1897.68
Status	US Open	Log Normal	-963.17	1930.35	1937.42
		Weibull	-942.30	1888.60	1895.67
		Gamma	-952.11	1908.21	1915.29
		Logistic	-946.96	1897.92	1905.00
	Australian Open	Bernoulli	-174.67	351.346	354.876
	French Open	Bernoulli	-176.06	354.119	357.656
	Wimbledon	Bernoulli	-176.06	354.119	357.656

3.4. Parameter estimate

Table 6: Goodness of fit for parameter estimates of tennis first serve and match outcome

Tournament	Copula	Estimate	Std Error	MLE	AIC	BIC
US Open	Normal	0.635	0.035	43.592	-85.184	-81.655
	Student t	0.635	0.035	43.582	-83.163	-76.104
	Gumbel	1.668	0.079	36.718	-71.437	-67.907
	Clayton	1.790	0.086	19.555	-37.110	-33.581
	Frank	4.400	0.564	44.999	-87.998	-84.468
Australian Open	Normal	0.661	0.032	50.311	-98.621	-95.084
	Student t	0.661	0.032	50.295	-96.589	-89.514
	Gumbel	1.731	0.080	42.202	-82.404	-78.867
	Clayton	2.028	0.117	19.688	-37.376	-33.839
	Frank	4.839	0.570	52.867	-103.734	-100.197
French Open	Normal	0.687	0.027	57.163	-112.325	-108.788
	Student t	0.687	0.027	57.132	-110.263	-103.189
	Gumbel	1.804	0.082	48.238	-94.476	-90.939
	Clayton	2.326	0.142	19.072	-36.144	-32.607
	Frank	5.343	0.583	61.671	-121.342	-117.804
Wimbledon	Normal	0.663	0.032	50.698	-99.395	-95.858
	Student t	0.663	0.032	50.696	-97.393	-90.318
	Gumbel	1.732	0.083	42.782	-83.564	-80.026
	Clayton	2.036	0.102	17.493	-32.986	-29.448
	Frank	4.750	0.617	52.086	-102.172	-98.635

Table 4 displays the marginal distribution for all four tournaments in the Grand Slams Open for the percentage of first serve won (PFSW), percentage of first serve return won (PFSRW), and match outcome of the tennis first serve either won or lose the match against the opponent. All the parameters for each of the different tournaments used are tested using five different marginal distributions of Normal, also known as Gaussian, Log Normal, Weibull, Gamma, and Logistic for the tennis serve performance and using Bernoulli for the match outcome since the trial is only once for each match point. All these methods are estimated from the data sets using the method of Maximum Likelihood Estimation (MLE) as in Table 5. For each tournament, the best-fitted distribution for each tennis first serve is selected based on the value of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which is shown by the ‘bolded’ entries that showed the best-fitted marginal distribution with the lowest AIC and BIC values. It is noted that from Table 5, for instances the difference in AIC between the Normal distribution and the Gamma distribution for PFSW at the French Open is relatively small ($\Delta AIC = 0.4$). According to Burnham and Anderson (2002), models with ΔAIC less than 2 are considered to have substantial support, indicating that both distributions provide a comparable fit to the data. Therefore, although the Normal distribution was selected as the best-fitting model due to its lowest AIC value, the Gamma distribution may also be considered as an adequate alternative. But the Normal distribution was ultimately preferred due to its relative simplicity and ease of interpretation in subsequent modelling.

These results indicate clear differences for all four tournaments, the percentage of first serve return won (PFSRW) of the US Open, Australian Open, and Wimbledon are best fitted with a Weibull distribution, reflecting a skewed performance pattern where most players clustered around moderate return percentages, but a few achieve exceptionally high return values. This result aligns with the fast surfaces of these tournaments, where excellent elite serve return stands out against the majority. By contrast, the percentage of first serve won (PFSW) of the French Open is best fitted by a Normal distribution suggesting a more symmetric distribution of performance. This is aligned with the slower clay surface, which reduces serve dominance and balances outcome across players. For the tennis match outcome, since the data is binary with only a single trial, we only proposed the Bernoulli distribution for the marginal distribution.

The Frank copula is characterised by a symmetric dependence structure, meaning that it assigns equal strength of dependence in both upper and lower tails of the joint distribution, and does not exhibit tail dominance. In other words, the association between serve and return serve performance remains consistent across both high performance and low performance levels. In the tennis study, this symmetric dependence is meaningful because it suggests that the relationship between serve and return serve performance is balanced rather than driven primarily by extreme performances. Specifically, players who perform well in serving are also likely to perform consistently in return serving, and vice versa. This indicates that serve and return serve abilities are not independent specialised skills only evident during exceptional performance, but instead reflect an overall performance level of the player.

This finding advances the existing sports analytics literature by providing statistical evidence that the dependence between serve and return serve performance is symmetric rather than upper-tail or lower-tail dependent. Many previous studies analyse serve and return serve separately or focus on mean-based relationships, without examining the dependence structure across the full joint distribution. By applying copula modelling, this study demonstrates that the Frank copula is more appropriate than Gumbel copula (which emphasises upper-tail dependence) or Clayton copula (which emphasises lower-tail dependence), indicating that extreme dominance in only one component (serve or return) does not solely drive the

dependence. Instead, the dependence exists consistently across all performance levels. This provides a more comprehensive understanding of player performance structure and highlights the importance of modelling joint performance using flexible dependence models such as copulas.

3.5. Scatterplot

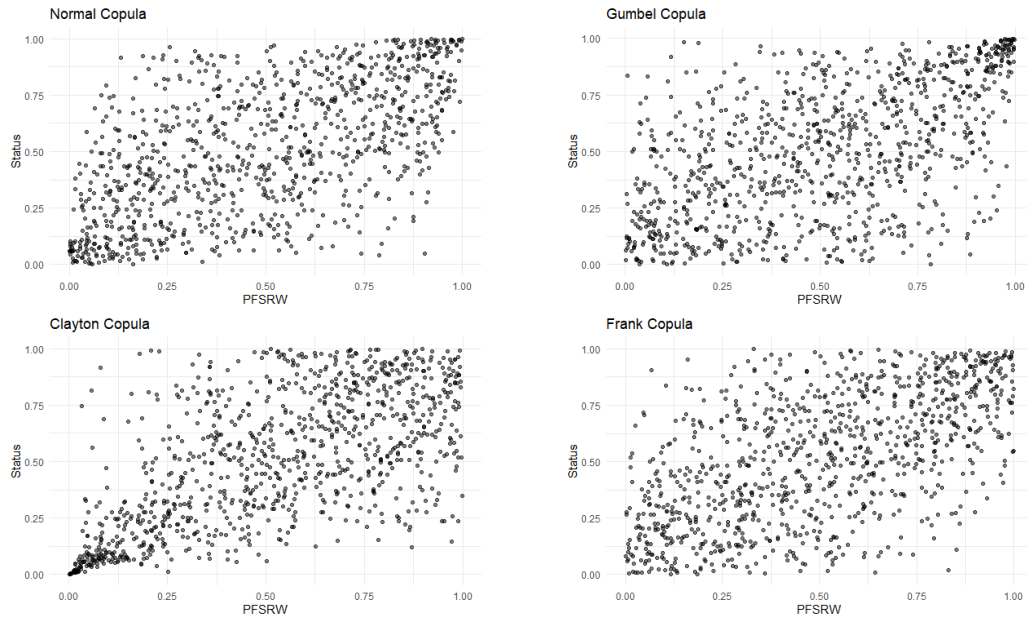


Figure 1: Scatterplot of Normal, Gumbel, Clayton and Frank for US Open

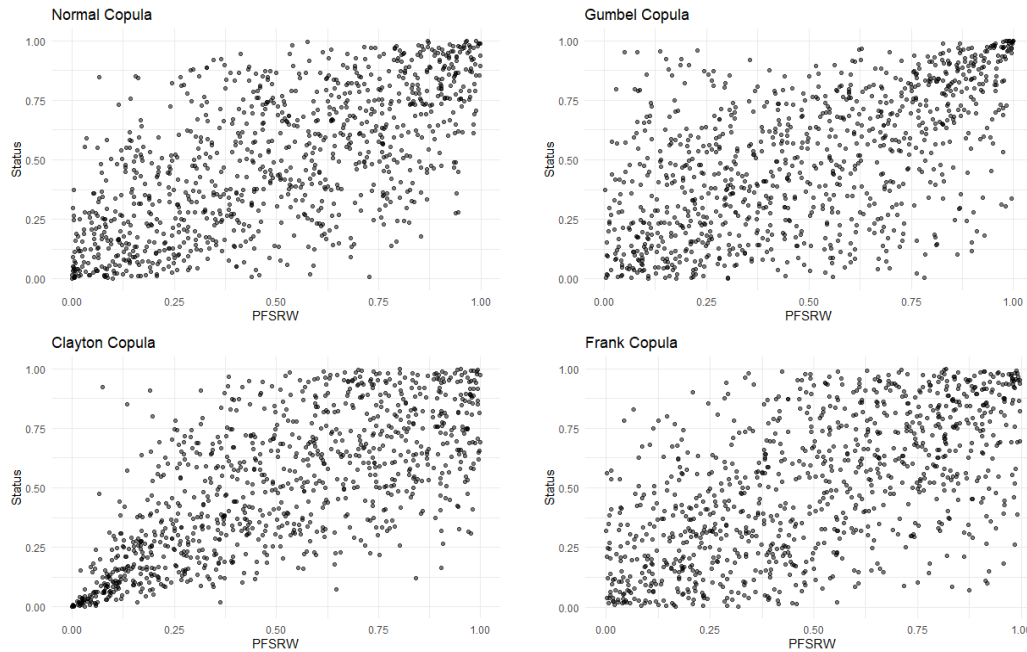


Figure 2: Scatterplot of Normal, Gumbel, Clayton and Frank for Australian Open

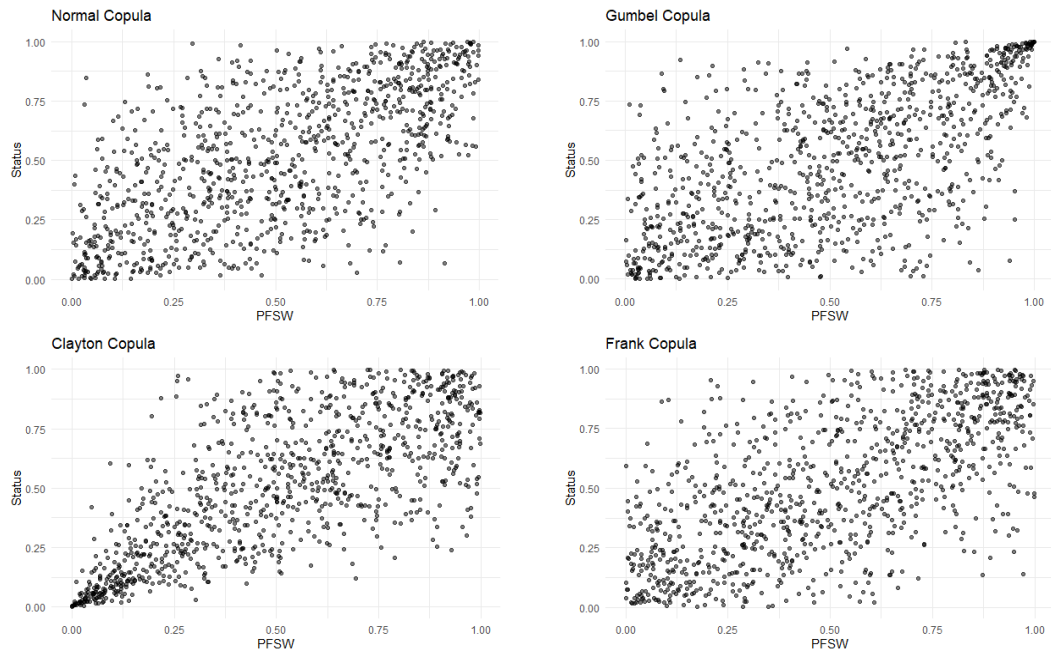


Figure 3: Scatterplot of Normal, Gumbel, Clayton and Frank for French Open

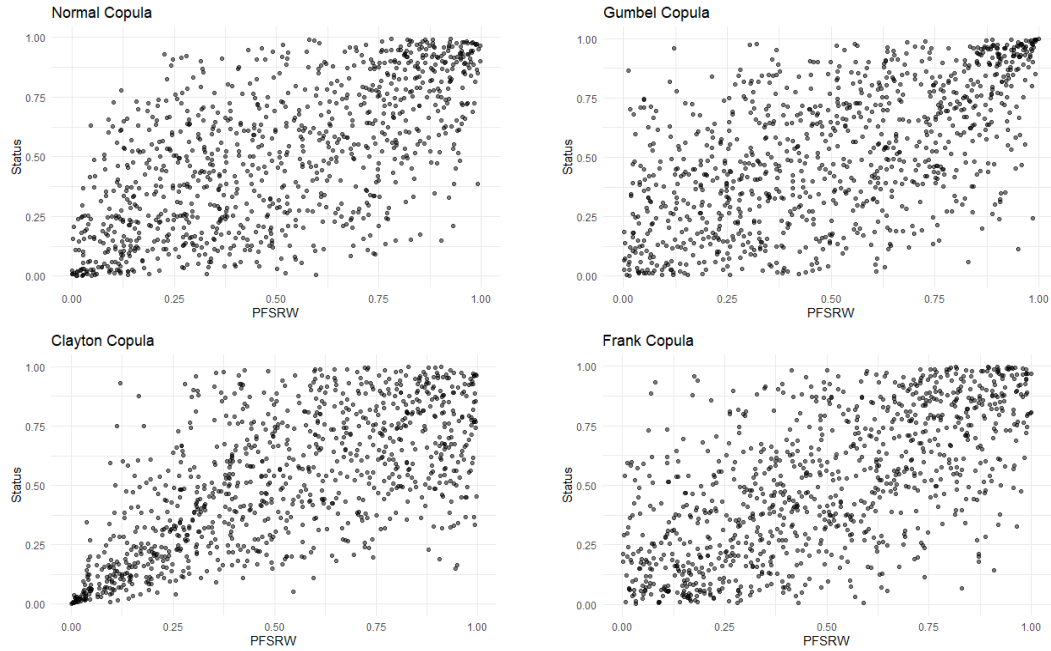


Figure 4: Scatterplot of Normal, Gumbel, Clayton and Frank for Wimbledon

A scatterplot is used to display a relationship between two sets of data, and a dot in the visual will represent a single data point. Plotting several data points reveals a trend on the graph. The scatterplot in Figures 1 to 4 shows the correlation that exists between the bivariate variables in the four different tournaments with differences in the court surface. It is well-known that the Australian Open and US Open are played on hard courts, Wimbledon on grass courts, and the French Open on clay courts. According to Prieto-Lage *et al.* (2023), different court surfaces give differences in the bounce of the ball that are treated as mediating factors, which might influence how each player's skill in controlling the bounce of the ball might affect the outcome of the match. For example, for a slow surface such as a clay court, the ball will be facing a higher coefficient of friction, decreased horizontal speed, and an increase in the height of the ball bounces, and vice versa for a fast surface such as a hard court. Observing the plot, there is a strong positive linear correlation of the tennis first serve between serve and return of serve on match status. As the first serve performance increases, the result of the match will also increase towards winning. This indicates that increasing the serve performance is important as it will help to player to score in the match and guarantee a point won for them. The scattering of the dot is different based on the graphs observed might be influenced by the serve performance on the different surface used.

3.6. Contour and density plot

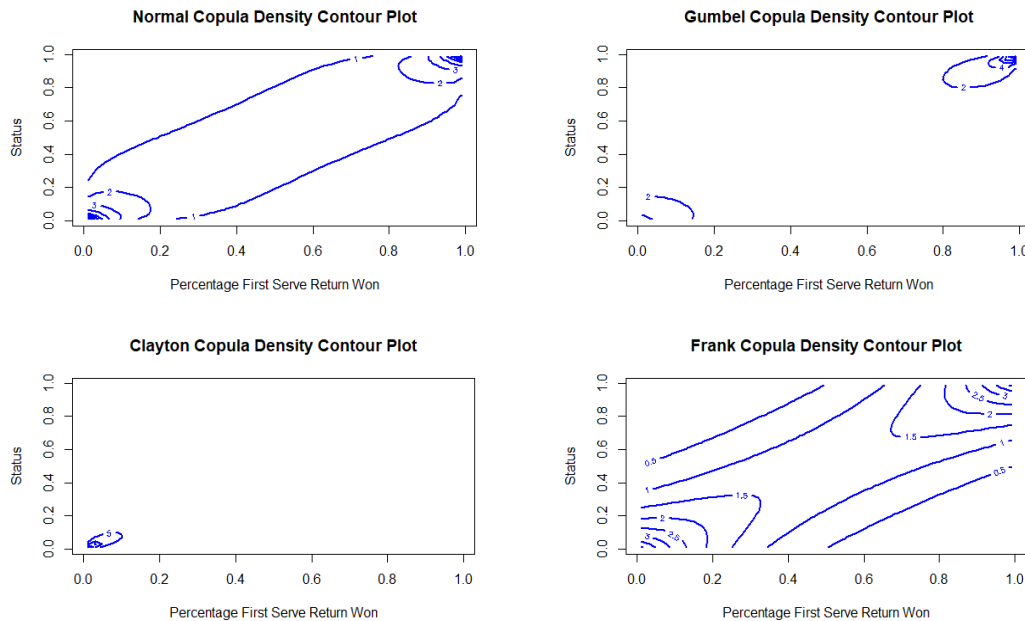


Figure 5: Contour plot of Normal, Gumbel, Clayton and Frank for US Open

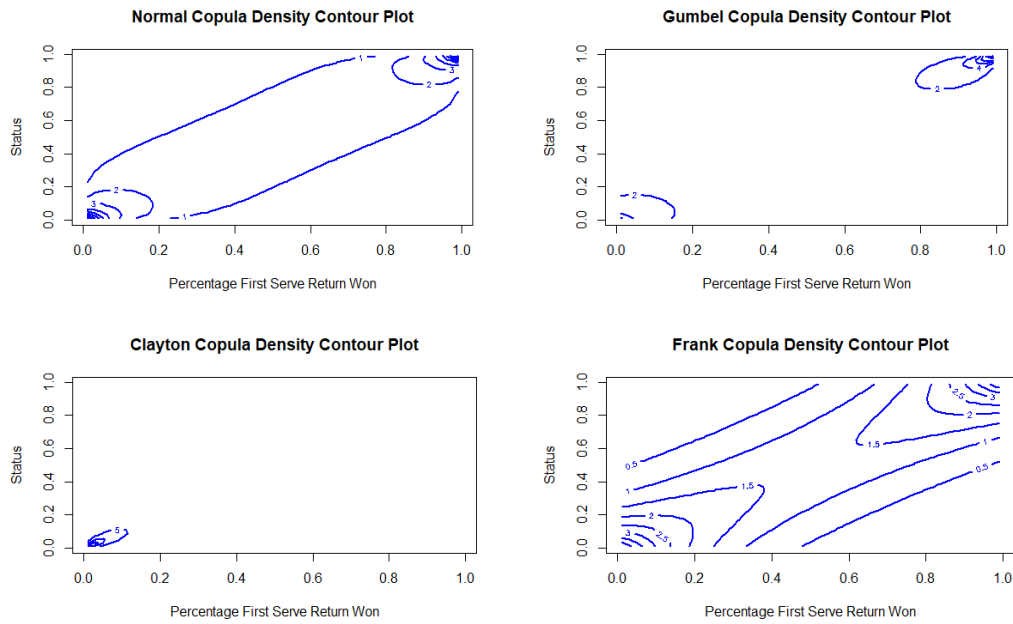


Figure 6: Contour plot of Normal, Gumbel, Clayton and Frank for Australian Open

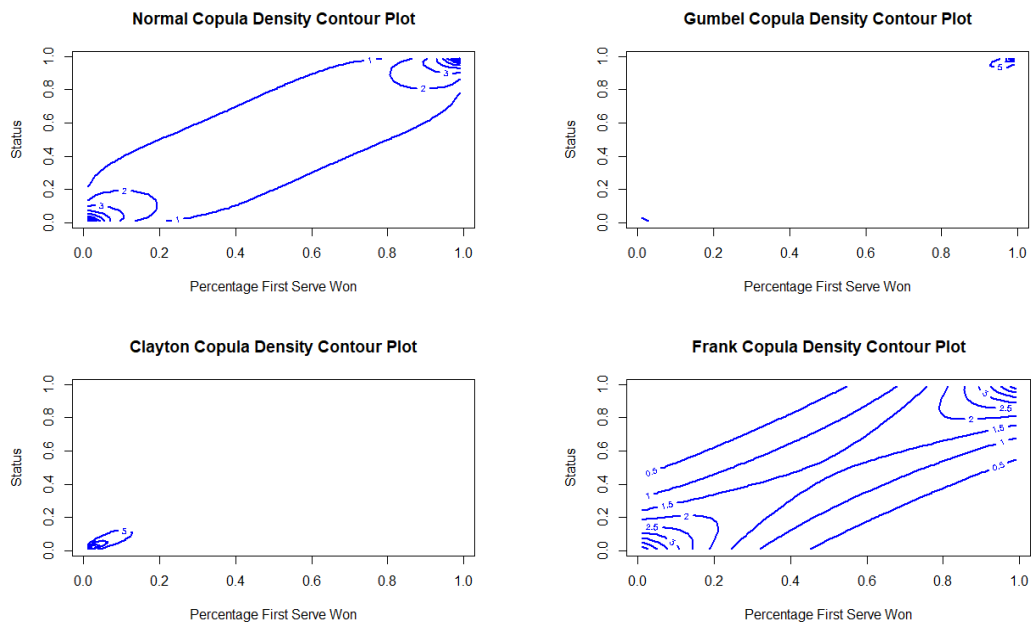


Figure 7: Contour plot of Normal, Gumbel, Clayton and Frank for French Open

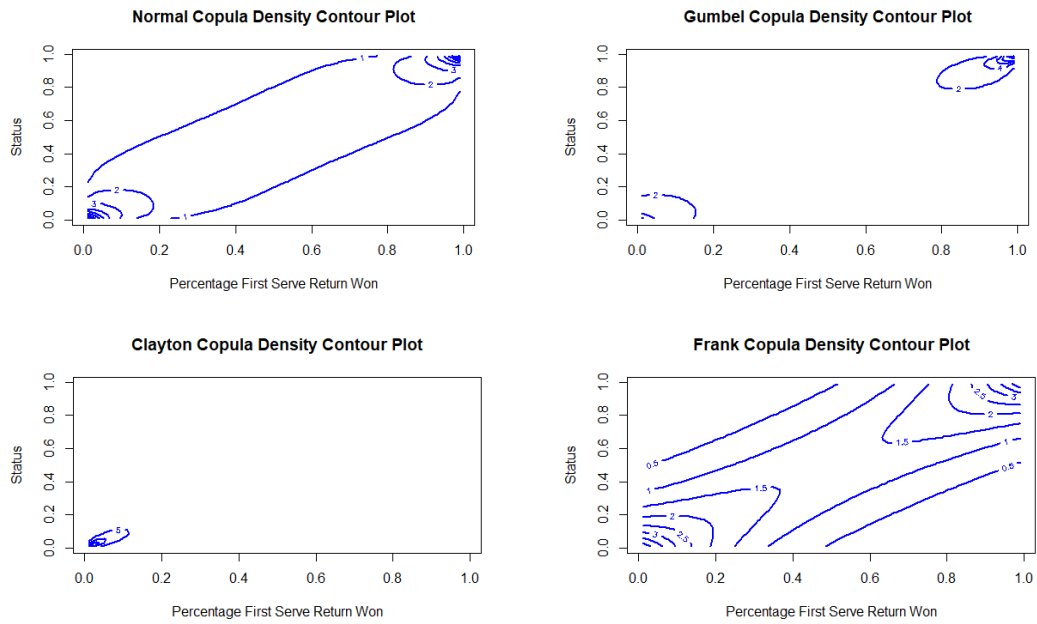
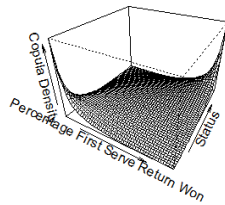
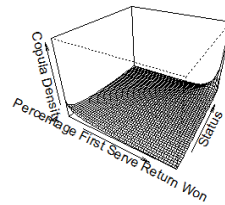


Figure 8: Contour plot of Normal, Gumbel, Clayton and Frank for Wimbledon

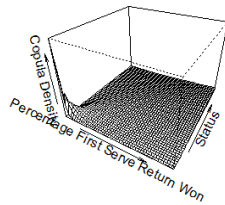
Normal Copula Density (Tail Dependence)



Gumbel Copula Density (Tail Dependence)



Clayton Copula Density (Tail Dependence)



Frank Copula Density (Tail Dependence)

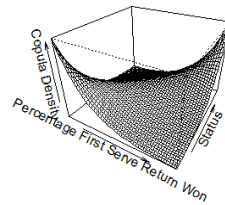


Figure 9: Density plot of Normal, Gumbel, Clayton and Frank for US Open

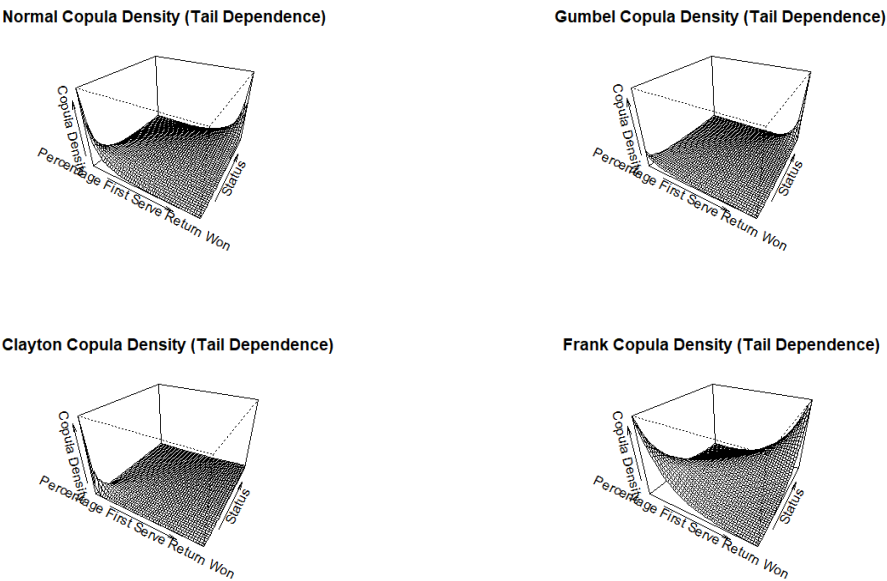


Figure 10: Density plot of Normal, Gumbel, Clayton and Frank for Australian Open

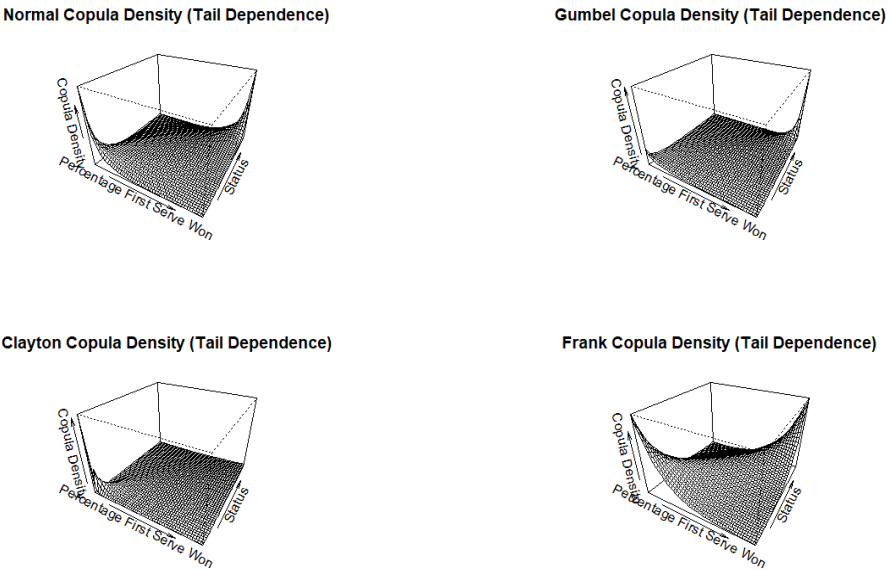


Figure 11: Density plot of Normal, Gumbel, Clayton and Frank for French Open

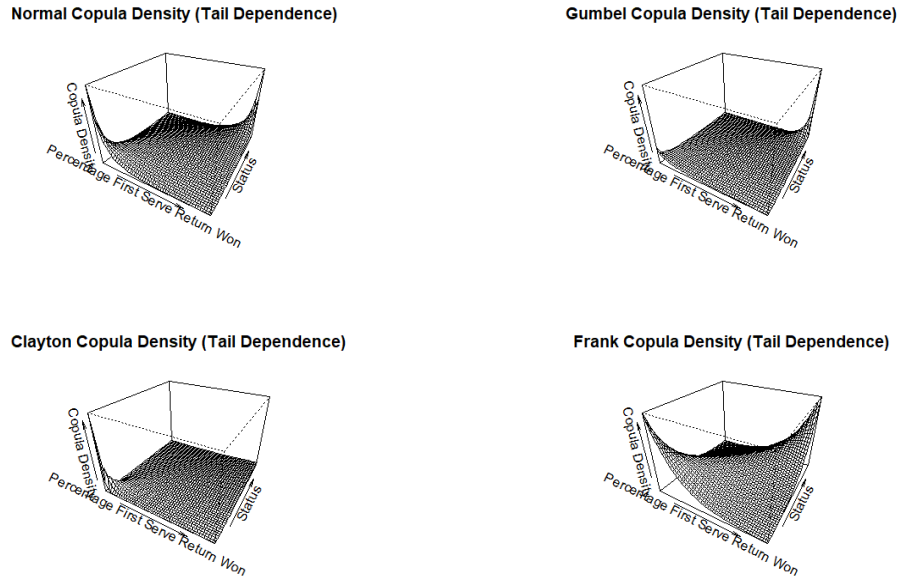


Figure 12: Density plot of Normal, Gumbel, Clayton and Frank for Wimbledon

Figures 5 to 8 show the contour plots for all four tournaments of Normal, Gumbel, Clayton, and Frank copulas. The plots show that the Frank copula fits the distribution of the tennis first serve performance and match outcome called status in all the tournaments involved. Looking at the contour plot of the Frank copula for all the tournaments, it can be seen that the contour lines are close together along the diagonal, indicating the variables have a significant positive dependence, which means they tend to increase or decrease concurrently at the value of 0 and 1. Towards the center of the plot, it can be observed that the contour lines are farther apart, indicating that weak dependence exists here. The theta value of the Frank is greater than zero, indicating that it has a stronger dependence, and the contour will be more concentrated towards the diagonal, reflecting stronger joint behavior. While observing the contour plot, it can be observed that the density plot for the Frank copula for all the tournaments involved is provided in Figures 9 to 12. The Frank copula is denser along the diagonal line $u = v$, demonstrating the positive correlation between the variables. When the copula is strong with a greater value of θ , the density is peaked near the diagonal, indicating that the variables are more likely to have similar values. As θ increases, the density becomes sharper and more concentrated around the diagonal, indicating that the two variables are more likely to exhibit similar behavior in which both are increasing or decreasing.

4. Conclusion

In conclusion, this study proposed a copula-based framework to model women tennis performance, providing an alternative to traditional multivariate models that assume normal marginal distributions. By allowing each variable to follow its appropriate marginal distribution, the copula approach offers a more flexible representation of tennis performance data. The goal is to suggest another applicable model that can be used to model the sports data when we do not always consider the data to follow the normal distribution. It is better in certain cases to follow the real distribution of the data to get the actual model closer to the real nature

of the data. Upon reviewing the tennis paper that is related to tennis serve and match status that was available in the previous research done, this paper is built upon this foundation by introducing the novel bivariate copula model that incorporates the marginal distributions of the serve performance and the copula fitting.

Based on the analysis of Grand Slam tournaments, the Frank copula was consistently identified as the best-fitting model, indicating the presence of a symmetric dependence between first serve and return serve performance in relation to match outcomes. This result suggests that these serve-related variables are statistically associated and jointly contribute to match performance within the scope of the tennis performance.

These findings contribute to the existing literature by demonstrating that copula models can be applied to women tennis performance data, which has received relatively less attention compared to men tennis. The proposed model provides an additional statistical framework that may be useful for analysing the dependence structure between performance variables and match outcomes. From a practical perspective, the model may offer useful insights for analysts and coaches by helping to quantify the relationship between serve and return serve effectiveness. Such information could support performance evaluation and training plans, although further research incorporating additional performance variables and broader datasets would be beneficial to strengthen its practical applicability.

Overall, this study highlights the potential of the copula approach as an alternative method for modeling sports performance data, particularly when the assumption of normality is not met. It is noting that predictive assessment and forecasting based on the selected copula model represent important directions for future research. Future work should also extend the model beyond serve statistics by incorporating factors such as serve speed, rally length, player fitness and ball velocity to improve model accuracy. The framework can also be adapted to national-level datasets, including Malaysian athletes, to support localised performance analysis and training strategies.

References

- Almetwally E.M., Fayomi A. & Qura M.E. 2024. Advanced copula-based models for type II censored data: Applications in industrial and medical settings. *Mathematics* **12**(12): 1774.
- Barnett T. & Clarke S.R. 2005. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics* **16**(2): 113–120.
- Bhatti M.I. & Do H.Q. 2019. Recent development in copula and its applications to the energy, forestry and environmental sciences. *International Journal of Hydrogen Energy* **44**(36): 19453–19473.
- Burnham K.P. & Anderson D.R. 2002. *Model selection and multimodel inference : A practical information-theoretic approach*. 2nd ed. New York: Springer.
- Chen T.X., Lyu H.S., Horton R., Zhu Y.H., Chen R.S., Sun M.Y., Liu M.W. & Lin Y. 2024. Using copula functions to predict climatic change impacts on floods in river source regions. *Advances in Climate Change Research* **15**(3): 406–418.
- Dewick P.R. & Liu S. 2022. Copula modelling to analyse financial data. *Journal of Risk and Financial Management* **15**(3): 104.
- El-Hashash E.F. & Shiekh R.H.A. 2022. A comparison of the Pearson, Spearman rank and Kendall tau correlation coefficients using quantitative variables. *Asian Journal of Probability and Statistics* **20**(3): 36–48.
- Etaga H.O., Okoro I., Aforka K.F. & Ngonadi L.O. 2021. Methods of estimating correlation coefficients in the presence of influential outlier(s). *African Journal of Mathematics and Statistics Studies* **4**(3): 157–185.
- Genest C. & Favre A.-C. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* **12**(4): 347–368.
- Gorgi P., Koopman S.J. & Lit R. 2019. The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society: Series A* **182**: 1393–1409.
- Joe H. 2014. *Dependence modeling with copulas*. Boca Raton: Chapman and Hall/CRC.
- Joo K., Shin J.Y. & Heo J.H. 2020. Modified maximum pseudo likelihood method of copula parameter estimation for skewed hydrometeorological data. *Water* **12**(4): 1182.

- Kim G., Silvapulle M.J. & Silvapulle P. 2007. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis* **51**(6): 2836–2850.
- Klaassen F.J.G.M. & Magnus J.R. 2001. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association* **96**(454): 500–509.
- Kovalchik S.A. 2016. Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports* **12**(3): 127–138.
- Liu C., Yang H., Yang J. & Wang H. 2024. Correlation analysis of diabetes based on copula. *Frontiers in Endocrinology* **15**: 1-11.
- Nelsen R.B. 1999. *An introduction to copulas*. New York: Springer.
- Newton P.K. & Aslam K. 2009. Monte Carlo tennis: A stochastic Markov Chain model. *Journal of Quantitative Analysis in Sports* **5**(3): 1-42.
- Palayangoda L.K., Senevirathne H.W. & Manage A.B. 2022. Modeling joint survival probabilities of runs scored and balls faced in limited overs cricket using copulas. *Journal of Sports Analytics* **8**(4): 277–289.
- Percy D.F. 2015. Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes. *Journal of the Operational Research Society* **66**(11): 1840–1849.
- Prieto-Lage I., Paramés-González A., Torres-Santos D., Argibay-González J.C., Reguera-López-de-la-Osa X. & Gutiérrez-Santiago A. 2023. Match analysis and probability of winning a point in elite men's singles tennis. *PLoS One* **18**(9): e0286076.
- Ruslan N.A.S., Zainol Z. & Abdul Rauf U.F. 2024. A comparative study of feature selection technique for predicting the professional tennis matches outcome in a grand slam tournament. *JOIV: International Journal on Informatics Visualization* **8**(1): 270–278.
- Sklar M. 1959. Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP* **8**(3): 229–231.
- Sulistianingsih E. & Satyahadewi N. 2025. Estimation of tail value at risk for bivariate portfolio using Gumbel copula. *JTAM (Jurnal Teori dan Aplikasi Matematika)* **9**(2), 615–628.
- Temizhan E., Mirtagioglu H. & Mendes M. 2022. Which correlation coefficient should be used for investigating relations between quantitative variables? *American Scientific Research Journal for Engineering, Technology, and Sciences* **85**(1): 265–277.
- Yu L. & Wang Y. 2023. Analysis of tennis techniques and tactics based on multiple linear regression model. *Applied Mathematics and Nonlinear Sciences* **8**(2): 2061–2068.

Department of Defence Science
Universiti Pertahanan Nasional Malaysia
Sungai Besi Camp, 57000
Kuala Lumpur, MALAYSIA
E-mail: 3231768@alfateh.upnm.edu.my

Department Centre for Defence Foundation Studies
University Pertahanan Nasional Malaysia
Sungai Besi Camp, 57000
Kuala Lumpur, MALAYSIA
E-mail: ummul@upnm.edu.my*

Department of Computer Science
University Pertahanan Nasional Malaysia
Sungai Besi Camp, 57000
Kuala Lumpur, MALAYSIA
E-mail: zuraini@upnm.edu.my

Received: 9 May 2025

Accepted: 28 March 2026

*Corresponding author