

EDITORIAL TEAM

EDITOR IN CHIEF

- Francesco Palumbo, Università di Napoli Federico II, Naples, Italy

CO-EDITORS ON A SPECIFIC SUBJECT

- Alessandro Celegato, AICQ Centronord - Quality and technology in production
- Adriano Decarli, Università di Milano, IRCCS /INT Foundation, Milan, Italy - Social and health studies
- Luigi Fabbri, Università di Padova, Padua, Italy - Surveys and experiments
- Vittorio Frosini, Università Cattolica del Sacro Cuore, Milan, Italy - Book review
- Antonio Giusti, Università di Firenze, Florence, Italy - Data Science
- Paolo Mariani, Università di Milano Bicocca, Milan, Italy - Social and economic analysis and forecasting

SCIENTIFIC COMMITTEE

- Thomas Aluja, UPC, Barcelona, Spain
- Paul P. Biemer, RTI and IRSS, Chicago, USA
- Jörg Blasius, Universität Bonn, Bonn, Germany
- Irene D'Epifanio, Universitat Jaume I, Castelló de la Plana, Spain
- Vincenzo Esposito Vinzi, ESSEC Paris, France
- Gabriella Grassia, Università di Napoli Federico II, Naples, Italy
- Michael J. Greenacre, UPF, Barcelona, Spain
- Salvatore Ingrassia, Università di Catania, Catania, Italy
- Ron S. Kenett, KPA Ltd. and Samuel Neaman Institute, Technion, Haifa, Israel
- Stefania Mignani, Università di Bologna Alma Mater, Bologna, Italy
- Tormod Naes, NOFIMA, Oslo, Norway
- Alessandra Petrucci, Università di Firenze, Florence, Italy
- Monica Pratesi, Università di Pisa, Pisa, Italy
- Maurizio Vichi, Sapienza Università di Roma, Rome, Italy
- Giorgio Vittadini, Università di Milano Bicocca, Milan, Italy
- Adalbert Wilhelm, Jacob University, Breimen, Germany

## ASSOCIATE EDITORS

- Francesca Bassi, Università di Padova, Padua, Italy
- Bruno Bertaccini, Università di Firenze, Florence, Italy
- Matilde Bini, Università Europea, Rome, Italy
- Giovanna Boccuzzo, Università di Padova, Padua, Italy
- Maurizio Carpita, Università di Brescia, Brescia, Italy
- Giuliana Coccia, ISTAT, Rome, Italy
- Fabio Crescenzi, ISTAT, Rome, Italy
- Franca Crippa, Università di Milano Bicocca, Milan, Italy
- Corrado Crocetta, Università di Foggia, Foggia, Italy
- Cristina Davino, Università di Napoli Federico II, Naples, Italy
- Loretta Degan, Gruppo Galgano, Milan, Italy
- Tonio Di Battista, Università di Chieti-Pescara “Gabriele D’Annunzio”, Pescara, Italy
- Tommaso Di Fonzo, Università di Padova, Padua, Italy
- Francesca Di Iorio, Università di Napoli Federico II, Naples, Italy
- Simone Di Zio, Università di Chieti-Pescara “Gabriele D’Annunzio”, Pescara, Italy
- Filippo Domma, Università della Calabria, Rende, Italy
- Alessandra Durio, Università di Torino, Turin, Italy
- Monica Ferraroni, Università di Milano, Milan, Italy
- Giuseppe Giordano, Università di Salerno, Salerno, Italy
- Michela Gnaldi, Università di Perugia, Perugia, Italy
- Domenica Fioredistella Iezzi, Università di Roma Tor Vergata, Rome, Italy
- Michele Lalla, Università di Modena e Reggio Emilia, Modena, Italy
- Maria Cristina Martini, Università di Modena e Reggio Emilia, Modena, Italy
- Fulvia Mecatti, Università di Milano Bicocca, Milan, Italy
- Sonia Migliorati, Università di Milano Bicocca, Milan, Italy
- Michelangelo Misuraca, Università della Calabria, Rende, Italy
- Francesco Mola, Università di Cagliari, Cagliari, Italy
- Roberto Monducci, ISTAT, Rome, Italy
- Isabella Morlini, Università di Modena e Reggio Emilia, Modena, Italy
- Biagio Palumbo, Università di Napoli Federico II, Naples, Italy
- Alfonso Piscitelli, Università di Napoli Federico II, Naples, Italy
- Antonio Punzo, Università di Catania, Catania, Italy
- Silvia Salini, Università di Milano, Milan, Italy
- Luigi Salmaso, Università di Padova, Padua, Italy
- Germana Scepi, Università di Napoli Federico II, Naples, Italy
- Giorgio Tassinari, Università di Bologna Alma Mater, Bologna, Italy
- Ernesto Toma, Università di Bari, Bari, Italy

- Rosanna Verde, Università della Campania “Luigi Vanvitelli”, Caserta, Italy
- Grazia Vicario, Politecnico di Torino, Turin, Italy
- Maria Prosperina Vitale, Università di Salerno, Salerno, Italy
- Susanna Zaccarin, Università di Trieste, Trieste, Italy
- Emma Zavarrone, IULM Milano, Milan, Italy

#### EDITORIAL MANAGERS

- Domenico Vistocco, Università di Napoli Federico II, Naples, Italy

#### EDITORIAL STAFF

- Antonio Balzanella, Università della Campania “Luigi Vanvitelli”, Caserta, Italy
- Luca Bagnato, Università Cattolica del Sacro Cuore, Milan, Italy
- Paolo Berta, Università di Milano Bicocca, Milan, Italy
- Francesca Giambona, Università di Firenze, Florence, Italy
- Rosaria Romano, Università di Napoli Federico II, Naples, Italy
- Rosaria Simone, Università di Napoli Federico II, Naples, Italy
- Maria Spano, Università di Napoli Federico II, Naples, Italy

#### A.S.A CONTACTS

##### **Principal Contact**

Francesco Palumbo (Editor in Chief)  
 editor@sa-ijas.org

##### **Support Contact**

Domenico Vistocco (Editorial Manager)  
 ijas@sa-ijas.org

#### JOURNAL WEBPAGE

<https://www.sa-ijas.org/ojs/index.php/sa-ijas>

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising: CLEUP SC, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: [info@cleup.it](mailto:info@cleup.it).

Rules for manuscript submission: <https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions>  
 Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (<http://web.it.nctu.edu/~qtqm/>)

SINERGIE – Italian Journal of Management



Statistica Applicata – Italian Journal of Applied Statistics (ISSN:1125-1964, E-ISSN:2038-5587) applies the Creative Commons Attribution (CC BY) license to everything we publish.

Published: August 2023

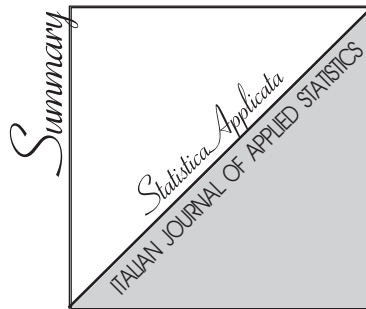
CLEUP SC

‘Coop. Libreria Editrice Università di Padova’

via G. Belzoni, 118/3 – Padova Italy

Phone +39 049 8753496 Fax +39 049 9865390

[info@cleup.it](mailto:info@cleup.it) – [www.cleup.it](http://www.cleup.it) – [www.facebook.com/cleup](http://www.facebook.com/cleup)



Vol. 35, Number 1

- 7 *Carpita, M., Metulini, R., Van Eetvelde, H.* *Thematic Issue on “Statistics for Performance and Match Analysis in Sports” - Editorial*
- 11 *Bonnini, S., Corain, L., Pesarin, F., Salmaso, L.* *Multivariate Permutation McNemar’s Test with Application to Performance Evaluation of Basket Players*
- 31 *Gjøen, P.S.-U., Hvattum, S.A., Moltubak, E.M., Hvattum, L.M.* *When is 2 Better than 3 in Basketball?*
- 47 *van der Wurp, H., Groll, A.* *Using (copula) Regression And Machine Learning to Model and Predict Football Results in Major European Leagues*
- 77 *Candila, V.* *welo: an R Package for Weighted and Standard ELO Rates*
- 95 *Epasinghe Dona, N., Swartz, T.* *A Causal Investigation of Pace of Play in Soccer*

## “STATISTICS FOR PERFORMANCE AND MATCH ANALYSIS IN SPORTS” - EDITORIAL

### **Maurizio Carpita**

*Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 25122, Brescia, Italy*

### **Rodolfo Metulini**

*Department of Economics, University of Bergamo, Via Caniana, 2, 24127, Bergamo, Italy*

### **Hans Van Eetvelde**

*Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281-S9, 9000, Ghent, Belgium*

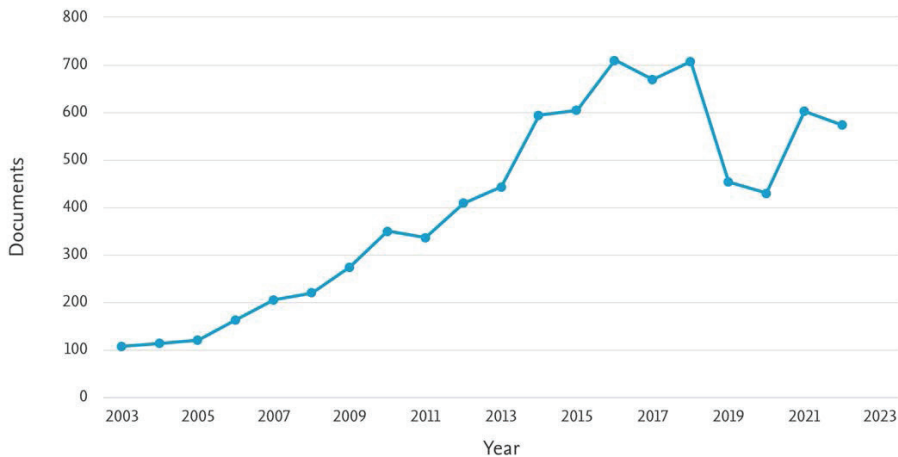
Statistics is more and more adopted in all sports with a variety of aims, ranging from predicting the outcome of a match of competition [3] and the analysis of performance [6, 7], to the prediction and prevention of injuries [9], amongst many others.

Statistical research and application in sports are fostered by the joint force of i) the increased availability of a large amounts of data of different types (e.g., play-by-play, trajectories, images) and from several sources (e.g., manual annotation, sensors, and tracking systems) and ii) the advances in information technologies and the computer storing capabilities [5, 8].

The interest in the topic is proved by the appearance of dedicated special issues [2, 4, 10], workshops, proceedings [1], and spontaneous contributions, also thanks to the birth of many research projects, such as the *BDsports* (<https://bodai.unibs.it/bdsports/>), which supports this thematic issue with the *ISI Special Interest Group on Sports Statistics* (<https://www.isi-web.org/isi-community/committees/sports-statistics>).

The invasiveness of this topic in scientific research is evident: In the last twenty years (2003-2022) a total of 8,080 articles with both the words “statistics” and “sport” in the title, in the abstract or in the keywords have been published by journals indexed in Scopus. As shown in Figure 1, in the early 2000s, the number of publications satisfying such a criterion was around 100 per year. These numbers increased to more than 700 in the year 2016 (with an average annual growth rate of 15.7%), they experimented a decrease to about 400 in 2019 and 2020 and they finally increased again reaching 600 articles per year in the last two years (2021 and 2022).

**Figure 1.** The number of articles published in journals indexed in Scopus with both “statistics” and “sport” in the title, abstract, or keywords.



This thematic issue follows up on methodological developments in the field, by collecting original contributions that focus on the application of up-to-date statistics and machine learning methods and techniques on sport-specific problems, as are the prediction of game outcomes, the evaluation of player/athlete’s performances and traits, the search for the optimal strategy and tactics to be adopted.

This thematic issue collects ten works about individual as well as team sports. Specifically, five of them are related to basketball, three refer to soccer, and two to tennis.

The first paper, by Bonnini, Corain, Pesarin, and Salmaso, investigates the application of the multivariate McNemar’s test for evaluating the effect of the field factor on the performance of basketball players. The proposed method is based on the nonparametric combination of permutation tests.

Gjøen, Hvattum, Moltubak, and Hvattum show through simulations that in basketball there are game situations where a strategy of taking fewer three-point attempts at the expense of more two-point attempts will improve the probability of winning the game.

The paper of van der Wurp and Groll compares classical univariate regression approaches with copula models explicitly accounting for the dependency structure as well as with modern machine learning techniques in the context of modeling and predicting football results in the major European leagues.

A description of the characteristics of the R-package “welo” is given by Candila. The package is dedicated to calculating the weighted and unweighted

Elo rates for tennis players. It allows the user to obtain the Welo and Elo rates easily and quickly, as well as the predicted probabilities of winning.

Dona and Swartz introduce two quantitative definitions of pace in soccer, whose calculations are facilitated through the availability of player tracking data. Their study investigates the influence of playing pace on the number of shots taken by a team.

The paper by Biancalani, Gnecco, and Metulini studies whether, for a basketball player, obtaining a large salary can be explained by its average marginal contribution to the team performance, measured using generalized Shapley values. The study is applied to players in the NBA.

Wu and Swartz have developed automatic methods that analyze the activities of players that are “off-the-ball” in soccer. They introduced a metric that measures defensive anticipation, based on the velocity of a defensive player in a given situation. The analysis is facilitated through player tracking data.

The work by Macis, Manisera, Sandri, and Zuccolotto studies which skills are associated with the probability for a basketball player of scoring a certain number of points during an NBA season segment, by applying a stepwise Cox regression model and a Lasso-Cox regression.

Tracking data systems gain a lot of interest in football, but they are still expensive. Broadcasting videos provide an alternative for tracking data, but they are of less quality and are censored. Therefore, the study by Karlis and Kontos explores interpolation methods for retrieving the missing information about players and ball positions and rectifies the effect of censoring.

The thematic issue concludes with the study by Milekhina, Breznik, and Restaino, which aims to investigate the existence of professional tennis players’ psychological traits. For this purpose, datasets on tennis matches of professional male and female tennis players were collected and dynamical network analysis was applied using the RSiena program.

Finally, many thanks to all the reviewers that made this special issue possible.

*The Guest Editors*

Maurizio Carpita, University of Brescia, Italy  
Rodolfo Metulini, University of Bergamo, Italy  
Hans Van Eetvelde, Ghent University, Belgium



**REFERENCES**

1. Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (2018). *Machine Learning and Data Mining for Sports Analytics*. Cham: Springer. doi: 10.1007/978-3-030-17274-9
2. Brefeld, U., Zimmermann, A. (2017). Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, 31(6), 1577– 1579. doi: 10.1007/s10618-017-0530-1
3. Bunker, R., Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A Review. *Journal of Artificial Intelligence Research* 73:1285–1322. doi: 10.1613/jair.1.13509
4. Groll, A., Liebl, D. (2022). Editorial special issue: Statistics in sports. *Advances in Statistical Analysis*, 1-7. doi: 10.1007/s10182-022-00453-9
5. Gudmundsson, J., Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, 50(2), 1-34. doi: 10.1145/3054132
6. Lord, F., Pyne, D.B., Welvaert, M., Mara, J.K. (2020). Methods of performance analysis in team invasion sports: A systematic review. *Journal of Sports Sciences*, 38:2338–2349. doi: 10.1080/02640414.2020.1785185
7. Sarlis, V., Tjortjis, C. (2020). Sports analytics — Evaluation of basketball players and team performance. *Information Systems*, 93:101562. doi: 10.1016/j.is.2020.101562
8. Stein, M., Janetzko, H., Seebacher, D., Jäger, A., Nagel, M., Hölsch, J., Kosub, S., Schreck, T., Keim, D.A., Grossniklaus, M. (2017). How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, 2(1), 2. doi: 10.3390/data2010002
9. Van Eetvelde, H., Mendonça, L.D., Ley, C., Seil, R., Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics*, 8:27. doi: 10.1186/s40634-021-00346-x
10. Zuccolotto, P., Manisera, M., Kenett, R. (2017). Statistics in sports: Project BDSports (Big Data Analytics in Sports) bodai. unibs.it/BDSports. *Electronic Journal of Applied Statistical Analysis*, 10(3). ISSN: 2070-5948

## MULTIVARIATE PERMUTATION MCNEMAR'S TEST WITH APPLICATION TO PERFORMANCE EVALUATION OF BASKET PLAYERS

**Stefano Bonnini**<sup>1</sup>

*Department of Economics and Management, University of Ferrara, Ferrara, Italy*

**Livio Corain**

*Department of Management and Engineering, University of Padova, Padova, Italy*

**Fortunato Pesarin**

*Department of Statistical Sciences, University of Padova, Padova, Italy*

**Luigi Salmaso**

*Department of Management and Engineering, University of Padova, Padova, Italy*

**Abstract** *The McNemar test can be considered the extension of the one-sample test on proportions to the case of two dependent samples or a special case of the sign test for paired data. In this paper we focus on the multivariate McNemar's test by considering an unusual but interesting application of basket analytics. The application is related to the evaluation of the effect of the field factor in the performance of basket players. The proposed method is based on the nonparametric combination of permutation tests.*

**Keywords:** *Basket analytics, McNemar's test, Multivariate analysis, Permutation test.*

### 1. INTRODUCTION

The McNemar test provides a nonparametric solution to a very popular problem. It can be considered the extension of the one-sample test on proportions to the case of two dependent samples or a special case of the sign test for paired data (McNemar, 1947). Medical applications are very widespread (Eliasziw and Donner, 1991; Gonen, 2004; Lachin, 1992). However the fields of application are numerous and very heterogeneous: computer science (Shao et al., 2021), marketing (Bonnini et al., 2014), genetics (Akazawa et al., 2021), engineering (Ibrahim et al., 2021), education (Stransky et al., 2021), behavioral ecology (Pembury Smith and Ruxton, 2020) and many others. McNemar's test is also suitable for comparing classification rates of multiple predictive models (Demsar, 2006; Durkalski et al., 2003; Leisenring et al., 2000; Lyles et al., 2005).

<sup>1</sup>Stefano Bonnini, email: stefano.bonnini@unife.it

Let us consider a binary response variable with paired observations. For example, let us take into account a sample of basket players and a dichotomous variable  $X$  representing the players' performance in a given season.  $X$  takes value 1 if the performance is good (or positive) and 0 if the performance is bad (or negative). We are interested in the distinction between *home* and *away* matches and the observed data can be represented by a  $2 \times 2$  table whose rows correspond to good and bad performance in the *home* matches and the columns to good and bad performance in the *away* matches. The hypothesis that the performance of basket players is not affected by the so-called "field factor" is equivalent to the equality of the marginal probabilities of good performance in the *home* and *away* matches. We will see that, in order to test the significance of field factor's effect, we must compare the number of discordant paired observations. This is the typical goal of McNemar's test. Several versions and improvements of the test have been proposed over time to have powerful solutions suitable for the specific framework of the study, nature of the data and research objectives.

Methodological proposals have been published for the application of McNemar's test on clustered binary data. Some of these contributions are based on scalar adjustments of the test statistic as if the assumption of independence on two variables is satisfied and a further adjustment by a factor in order to keep the null distribution approximately correct (Donald and Donner, 1987, 1990; Donner, 1992). Others are focused on the ratio estimator (Obuchowski, 1998; Rao and Scott, 1992). Wu (2018) proposes a method for power calculation of the adjusted McNemar test with clustered data.

For multiple comparisons of dependent proportions Westfall et al. (2010) proposes a stepwise testing approach, by using discrete characteristics for exact McNemar's tests. This is a valid solution to several applications and is also suitable in case of missing values, tests with different sample sizes, and other non-standard or complex problems. In addition, to keep into account the dependence structure, an approximate bootstrap method is also proposed. These methods control the familywise error rate in the strong sense.

For the case of two independent samples of paired univariate dichotomous variables, we mention the contribution of Feuer and Kessler (1989). The case of binary crossover data was addressed by Becker and Balagtas (1993). Agresti and Klingenberg (2005) present solutions for the comparison of two independent multivariate binary vectors for an overall comparative evaluation of marginal incidence rates in two populations. A multivariate extension of the McNemar test is developed by Klingenberg and Agresti (2006), by discussing Wald and Score-

Type tests, Generalized Estimating Equations approach, Likelihood Ratio and Ordinary Score Test.

In this paper we focus on the multivariate McNemar test by considering an unusual but interesting application of basket analytics. This application concerns the evaluation of the effect of the field factor related to the performance of basket players. The proposed method is based on the nonparametric combination (NPC) of dependent permutation tests (Pesarin and Salmaso, 2010). The rest of the paper is organized as follows. In Section 2, we present the classic univariate version of the McNemar test. Section 3 is dedicated to introduce the application of basket analytics, concerning the performance evaluation of basket players by comparing *home* and *away* performance. We will consider a review of the literature specialized on this topic in order to determine a suitable multivariate response that represents the performance of basket players. In Section 4 we describe the multivariate permutation McNemar test and we apply it to the problem of basket analytics. Conclusions are provided in Section 5.

## 2. MCNEMAR TEST FOR PAIRED DATA WITH BINARY RESPONSES

The McNemar problem is also called *test for marginal homogeneity*. The reason of this name will soon be clear according to the following description. Let us assume that the dataset consists of  $n$  independent observations of the bivariate response variable  $(X_{i1}, X_{i2})$ , the determinations of which are  $\{(x_{i1}, x_{i2}), i = 1, \dots, n\}$ , where the two marginal responses can take only two categories, conventionally denoted by 0 and 1. For example, the couple  $(X_{i1}, X_{i2})$  could represent the presence/absence of two characteristics on the  $i$ -th statistical unit. Another example concerns classifications according to a dichotomous scale by two evaluators of  $n$  objects, subjects or items. Marginal homogeneity is equivalent to equality of the marginal distributions of the bivariate response or the agreement between the two evaluators. Data are assumed to be determinations of a bivariate *Bernoulli* random variable. The joint probability distribution can be represented as in Table 1, where  $\theta_{rs}$  denotes the probability of occurrence of the couple  $(r, s)$ , with  $r, s \in \{0, 1\}$ . The hypotheses under testing are  $H_0: \theta_{\bullet 1} = \theta_{1 \bullet}$  and  $H_1: \theta_{\bullet 1} \neq \theta_{1 \bullet}$ .

The joint frequency distribution can be represented by Table 2, where  $f_{rs}$  denotes the absolute frequency of the couple  $(r, s)$  in the observed sample, with  $r, s \in \{0, 1\}$ . Note that this table, being not related to independent samples, is not properly a contingency table; hence the typical techniques for contingency tables cannot be applied.

The more similar  $f_{00} + f_{01}$  and  $f_{00} + f_{10}$  are (i.e. difference between  $f_{01}$  and

**Table 1: Probability distribution of the bivariate Bernoulli random variable.**

		$X_1$		
		0	1	
$X_2$	0	$\theta_{00}$	$\theta_{01}$	$\theta_{0\bullet}$
	1	$\theta_{10}$	$\theta_{11}$	$\theta_{1\bullet}$
		$\theta_{\bullet 0}$	$\theta_{\bullet 1}$	1

$f_{10}$  close to zero) the greater the empirical evidence in favor of the hypothesis of marginal homogeneity (null hypothesis) and vice-versa. Hence, a suitable test statistic for such problem might be based on  $(f_{01} - f_{10})$ . For small sample sizes the test statistic (conditional on the marginal frequencies) might equivalently be

$$T = f_{01}.$$

In fact, the sum  $f_{01} + f_{10} = n - f_{00} - f_{11} = s$  is fixed and the test assesses disparity of the discordants  $f_{01}$  and  $f_{10}$ . Therefore  $f_{01} - f_{10} = 2f_{01} - s$  and, consequently, there is an exact linear relationship between the two test statistics. Thus, they lead to the same  $p$ -values. When  $f_{01} + f_{10} \leq 20$ , approximate distributions are not required and not valid, and the exact distribution of one of the two equivalent test statistics can be used for the inferential purpose. Under marginal homogeneity,  $T$  follows a binomial distribution with parameters  $f_{01} + f_{10}$  and 0.5, that is  $T \sim \text{Bin}(f_{01} + f_{10}, 0.5)$ . The null hypothesis is rejected for either small or large values of  $T$ . When  $f_{01} + f_{10} > 20$  then

$$T = (f_{01} - f_{10})^2 / (f_{01} + f_{10})$$

is typically used as a test statistic (Kvam and Vidakovic, 2007).

Under  $H_0$  it approximately follows a  $\chi^2$  distribution with 1 degree of freedom. Some authors take into account the discontinuity correction:

$$T = (|f_{01} - f_{10}| - 1)^2 / (f_{01} + f_{10}).$$

**Table 2: Absolute frequency distribution of a bivariate binary response variable**

		$X_1$		
		0	1	
$X_2$	0	$f_{00}$	$f_{01}$	$f_{00} + f_{01} = f_{0\bullet}$
	1	$f_{10}$	$f_{11}$	$f_{10} + f_{11} = f_{1\bullet}$
		$f_{00} + f_{10} = f_{\bullet 0}$	$f_{01} + f_{11} = f_{\bullet 1}$	$n$

But, from the practical point of view, some experts think that, thanks to the computational capabilities of modern computers, this correction becomes not relevant (Kvam and Vidakovic, 2007). Simple changes to the decision rule must be considered for the one-sided problem. This test was proposed by McNemar (1947). Some variations were presented by Bennett and Underwood (1970); Mantel and Fleiss (1975); McKinlay (1975); Ury (1975).

The McNemar test can also be seen as the extension of the one-sample test on proportion to the case of two dependent samples. It can be also considered a special case of the sign test for paired data.

For example, let us consider the data about the performance of basket players in the 2016/2017 Italian championship (regular season). A reasonable measure of individual performance in a match is the ratio between the number of scored points ( $PTS$ ) and the actual played time in minutes ( $TIME$ ):  $PER = PTS/TIME$ . In the 2016/2017 regular season of the Italian championship, the general mean value of  $PER$  with respect to all the players and all the matches was 0.35. Hence, to determine whether the individual performance of a given player over the regular season has been good/positive ( $X = 1$ ) or bad/negative ( $X = 0$ ) we can consider the average value of the individual index and compare it with the general average 0.35. Formally

$$X_i = \begin{cases} 1 & \text{if } \overline{PER}_i \geq 0.35 \\ 0 & \text{otherwise,} \end{cases}$$

where  $\overline{PER}_i$  denotes the average of the values of  $PER$  over the regular season for

the  $i$ -th player.

A typical goal of the performance analysis of athletes playing round robin tournaments is whether the field factor affects their performance. In other words, the question is whether the probability of good performance in *home* matches is equal to the probability of good performance in *away* matches. Let random variables  $X_H$  and  $X_A$  represent the individual performance in the *home* matches and in the *away* matches respectively. Let  $\theta_H = P(X_H = 1)$  and  $\theta_A = P(X_A = 1)$ . We want to test  $H_0 : \theta_H = \theta_A$  versus  $H_1 : \theta_H \neq \theta_A$ .

In basketball, the distinction between functions and roles of the 5 different players of a team is not very evident and the tasks are often interchangeable. Anyway, there are some reference roles:

1. *point guard* (playmaker), with the task of calling the game patterns and dictating the rhythms of the ball
2. *shooting guard*, with the tasks of supporting the point guard, with whom he shares most of the characteristics and is usually the best shooter of the team
3. *small forward*, usually tall, fast and agile, he is interchangeable with the shooting guard and the power forward; he is important for the particular offensive peculiarities as well as for the defensive phase, especially in rebounding
4. *power forward*, occupies the same areas as the small forward, but he has a more marked physicality, less suited to running; he is one of the tallest players and is inclined to make space between the opposing defenders in the area, ready to receive and reject impacts with the opponents
5. *center* (pivot), typically the tallest and slowest player, has most of the points in his hands (especially in shots near the rim of the basket) and, in the defensive phase, he is the main protector of his team's area

In the individual performance analysis of basketball players the role is clearly a possible confounding factor and the distinction between roles must be considered, for instance through a suitable stratification. Since the distinction of the 5 roles presented above could be not suitable because the roles are not always so distinct and well defined, a more general classification, very common in U.S.A., can be considered:

- *backcourt* players during ball possession, take care of playing the ball in the back court; this category includes point guard and shooting guard

**Table 3: Absolute frequency distribution of 2016/2017 Italian basket regular season sample of players according to their (binary) performance as a function of the  $\overline{PER}$  index.**

Performance in away matches	Performance in home matches	
	Bad	Good
Bad	8	7
Good	1	8

- *frontcourt* players are responsible of scoring in the offensive half of the court; this category includes small forward, power forward and center.

Data about the 2016/2017 Italian basket regular season were collected. A stratified random sample of 24 players (12 backcourt and 12 frontcourt) from all the individuals who played at least 10 *home* and 10 *away* matches, was selected. For each of these 24 players, the seasonal average performance  $\overline{PER}$  in the *home* matches and in the *away* matches was computed in order to obtain the couples of binary data  $(x_{iH}, x_{iA})$ , where  $x_{iH}$  indicates whether the average performance of the  $i$ -th player in *home* matches was good or not and  $x_{iA}$  indicates whether the average performance of the  $i$ -th player in *away* matches was good or not. A synthesis of sample data, in the form of  $2 \times 2$  table, is shown in Table 3.

In *R*, for the application of McNemar test, the command `mcnemar.test(x)` is to be used, where  $x$  represents the  $2 \times 2$  table like Table 3 or the equivalent for other problems. If the significance level of the test is set at  $\alpha = 0.10$ , since the  $p$ -value of the test is 0.0703, then the null hypothesis of equal probability of performance in the *home* and *away* matches is rejected in favor of the hypothesis that the probability of good performance changes according to the field factor (the one-sided  $p$ -value for  $\theta_H > \theta_A$  is 0.0352).

### 3. PERFORMANCE EVALUATION OF BASKET PLAYERS

The analysis of the individual performance of basketball players has been the subject of a vast scientific literature. Among the most recent contributions, we mention Page et al. (2007), Cooper et al. (2009), Piette et al. (2010), Fearnhead and Taylor (2011), Ozmen (2012) and Deshpande and Jensen (2016). Some works focused on the prediction of the match outcomes (Brown and Sokol, 2010; Gupta,



2015; Loeffeholz et al., 2009; Lopez and Matthews, 2015; Ruiz and Perez-Cruz, 2015; West, 2006; Yuan et al., 2015). An interesting work about players positions and effectiveness of the shots from different areas of basketball court is that of Shortridge et al. (2014). Zuccolotto and Manisera (2020) present an overview of methods, models and *R* packages for the analysis of basketball data.

In the considered case study, related to the Italian basketball championship regular season 2016/2017, we select a stratified random sample according to the latter role classification.

Typically, there are two approaches of performance analysis in basketball analytics: the *bottom-up* approach starts from the individual contributions of each athlete to predict the team's performance or the final result of a match; the *top-down* approach uses the overall contribution of the team to determine the individual contributions of players. Our contribution, although not specifically aimed at calculating the team's performance, is compatible with the *bottom-up* approach of which it could be a preliminary step. Since the starting point and the raw data refer to the individual performance, let us consider some scientific contributions about performance measures of individual players.

The ratio between the number of scored points *PTS* and the played time in minutes *TIME* mentioned in the previous section is a simple, reasonable but in many cases not adequate performance measure of a player in a match. Typical more sophisticated measures are:

- *Player Efficiency Rating (PER)*: it takes into account and weighs the number of 3-points shots, of 2-points shots and of free shots, the number of assists, the stolen balls, the blocks and other quantities. It is a reliable measure of performance only for the offensive phase and the reference values change season by season
- *Win Shares*: they measure the contribution of each single player to the team's overall victories, by distinguishing and summing the offensive and the defensive contribution. It is not suitable for small tournament such as the Italian championship with a total of only 30 matches in the regular season.
- *Tendex*: proposed by the sports journalist Dave Heeren in 1959, the *Tendex Rating* is a measure of efficiency based on a weighted algebraic sum of partial indices such as *PTS*, number of rebounds, number of assists, number of stolen balls, number of blocks, turnovers, free throws made, field goals made and personal fouls. This index is used to determine the *efficiency*

*rating* used still today, especially in the United States, as an efficiency assessment index and based on the ratio between Tendex and number of played matches. It is very popular because it uses simple variables, usually included in the box-scores, and takes into account both offensive and defensive performance.

- *Performance Index Rating (PIR)*: it can be considered the European version of Tendex. In 1991 it appears, for the first time, in the Spanish ACB League. It is still used today to determine the most valuable player (MVP) of the week in the Spanish national league and in the EuroLeague. It includes in the algebraic sum the same variables of Tendex with, in addition, *fouls drawn* and (with negative sign) *shots rejected*.
- *Offensive Efficiency Rating (OER)*: this is another very popular index defined by Dean Oliver as the number of points done by a player per 100 total possessions or simply the ratio between *PTS* and number of total possessions (*PO*).

The goal of this work is not to determine an optimal performance index but it is evident that each index has pros and cons and represents a partial aspect of a complex phenomenon. Consequently, the concept of *performance* of a basket player is multidimensional. In order to consider the multivariate nature of the response variable, we take into account the two most commonly used indices, *PIR* and *OER*, and we transform them with a logic similar to what we did with the *PER* index in order to compute a bivariate binary response variable representing the performance of a basket player in the regular season.

#### 4. MULTIVARIATE EXTENSION OF MCNEMAR TEST: PERMUTATION SOLUTION

Let us consider the multivariate extension of the problem illustrated above. The dataset consists of multivariate paired data with  $q$  binary variables. The data are assumed to be determinations of the random variables  $(X_{1ih}, X_{2ih})$  with  $i = 1, \dots, n$  and  $h = 1, \dots, q$ . Let  $\theta_{rs,h}$  denote the probability (or population proportion) of the couple  $(r, s)$  for the  $h$ -th response, with  $r, s \in \{0, 1\}$  and  $h = 1, \dots, q$ . The multivariate McNemar test can be defined as

$$H_0 : \bigcap_{h=1}^q [\theta_{01,h} = \theta_{10,h}],$$

against

$$H_1 : \bigcup_{h=1}^q [\theta_{01,h} < \neq > \theta_{10,h}],$$

where, in the overall alternative hypothesis, some of the partial hypotheses can be two-sided and some others one-sided. Each partial testing problem can be solved with the binomial test based on the test statistic  $T_h = f_{01,h}$  which, when  $H_0$  is true, follows a binomial distribution with parameters  $f_{01,h} + f_{10,h}$  and 0.5, where  $f_{rs,h}$  denotes the sample absolute frequency of the couple  $(r, s)$  for the  $h$ -th variable, with  $r, s \in \{0, 1\}$  and  $h = 1, \dots, q$ .

Equivalently, we can consider the following data transformation

$$Y_{ih} = g(X_{1i,h}, X_{2i,h}) = \begin{cases} +1 & \text{if } X_{1i,h} < X_{2i,h} \\ -1 & \text{if } X_{1i,h} > X_{2i,h} \\ 0 & \text{otherwise,} \end{cases}$$

and apply the permutation test for paired data based on the test statistic

$$T_h^* = \sum_{i=1}^n Y_{ih} S_i^*$$

with  $S_i^* = +1$  with probability 0.5 and  $-1$  with probability 0.5 under  $H_0$ . The application of the NPC methodology for multivariate permutation tests provides a solution to this testing problem (Pesarin and Salmaso, 2010).

The procedure requires the examination of all  $2^n$  possible permutations. In practice, when this number is large ( $2^{24} = 16\,777\,216$ ), their complete examination may become unpractical. Thus, according to the literature (Pesarin, 2001; Pesarin and Salmaso, 2010), especially in the  $q$ -dimensional case, we suggest considering a random sample from the set of permutations consisting in carrying out  $B$  independent permutations. In other words, this is realized by a random generation of  $B$  sets of  $n$ -dimensional vectors of signs (note: the same permutation of signs jointly for all  $q$  variables). To emphasize that the  $B$  permutations are taken conditionally on the given dataset, this procedure is named "Conditional Monte Carlo" (CMC). Once the  $q$  partial tests are carried out, the related  $q$  partial significance level functions are to be combined by means of a suitable combining function through the NPC methodology (Pesarin, 2001). According to the null permutation distribution of the combined test statistic, the  $p$ -value can be computed and compared with the significance level  $\alpha$  in order to take the final decision about either rejection or acceptance of the null hypothesis  $H_0$ . This method can be considered a particular case of the more general permutation test for multivariate paired observations. Suitable combining functions are:

- Fisher combining function:  $T_F = -2 \sum_h \log(\lambda_h)$ ,
- Liptak combining function:  $T_L = \sum_h \phi^{-1}(1 - \lambda_h)$ ,  $\phi^{-1}$  being the standard normal quantile function,
- Tippett combining function:  $T_T = \max_h(1 - \lambda_h)$ ,

where  $\lambda_h$  is the partial  $p$ -value.

The CMC procedure works as follows:

1. Compute the vector of observed values of the  $q$  partial test statistics as a function of the observed dataset  $\mathbf{X}$ :  $\mathbf{T}_{obs} = [T_1(\mathbf{X}), \dots, T_q(\mathbf{X})]' = [T_{1(0)}, \dots, T_{q(0)}]'$
2. Consider  $B$  random permutations and compute the values of the test statistics corresponding to each permuted dataset. For the  $b$ -th permuted dataset  $\mathbf{X}_{(b)}^*$  (with  $b = 1, \dots, B$ ), the test statistics are:  $\mathbf{T}_{\mathbf{b}}^* = [T_1(\mathbf{X}_{(b)}^*), \dots, T_q(\mathbf{X}_{(b)}^*)]' = [T_{1(b)}^*, \dots, T_{q(b)}^*]'$
3. Estimate the  $p$ -values according to the null permutation distribution:  $\hat{\lambda}_h = \hat{L}_h(T_{h(0)})$ ,  $\hat{\lambda}_{h(b)}^* = \hat{L}_h(T_{h(b)}^*)$ , with  $\hat{L}_h(t) = [\sum_{r=1}^B I(T_{h(r)}^* \geq t) + 0.5] / (B + 1)$  and  $I(A)$  being the indicator function of the event  $A$
4. Compute the observed value and the permutation values of the combined test statistic based on the combining function  $\psi$ ,  $T_\psi = \psi(\lambda_1, \dots, \lambda_q)$ :  $T_{\psi, obs} = \psi(\hat{\lambda}_1, \dots, \hat{\lambda}_q)$  and  $T_{\psi(b)}^* = \psi(\hat{\lambda}_{1(b)}^*, \dots, \hat{\lambda}_{q(b)}^*)$
5. Estimate the  $p$ -value of the combined test according to the null permutation distribution:  $\hat{\lambda}_\psi = \hat{L}_\psi(T_{\psi, obs})$

Since all partial tests are marginally unbiased, the combined test is unbiased. In other words, the probability of rejecting the null hypothesis in favor of the alternative, when the latter is true in at least one of  $q$  components, is greater than the significance level  $\alpha$  (Pesarin and Salmaso, 2010). Even if each partial test is distributed according to the binomial law, the multivariate (global) test is not multinomial. Moreover, when  $q > 2$ , the asymptotic approximation of the multivariate distribution cannot be considered, because the dependence relations among component binomials cannot be restricted to the  $q(q-1)/2$  pair-wise correlations coefficients (Joe, 1997; Pesarin, 2001). Indeed, also dependence three-wise, four-wise, etc. should be considered. Thus the described NPC by the CMC procedure based on  $B$  iterations is a suitable solution.

Let us consider again the example of the Italian championship regular season 2016/2017. The bivariate response variable is based on the indices,  $PIR$  and  $OER$ , transformed by a rationale similar to what we did with the  $PER$  index. In the 2016/2017 regular season of the Italian championship, the mean value of  $PIR$  with respect to all the players and all the matches was 8.5. Hence, in order to determine whether the individual performance of a given player over the regular season has been good/positive ( $X_1 = 1$ ) or bad/negative ( $X_1 = 0$ ) with respect to  $PIR$ , we can consider the average value of the individual index and compare it with the general average 8.5. Formally

$$X_{1i} = \begin{cases} 1 & \text{if } \overline{PIR}_i \geq 8.5 \\ 0 & \text{otherwise,} \end{cases}$$

where  $\overline{PIR}_i$  denotes the average of the values of  $PIR$  over the regular season for the  $i$ -th player. Similarly

$$X_{2i} = \begin{cases} 1 & \text{if } \overline{OER}_i \geq 0.84 \\ 0 & \text{otherwise,} \end{cases}$$

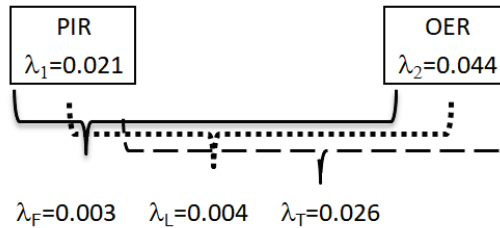
where  $\overline{OER}_i$  denotes the average of the values of  $OER$  over the regular season for the  $i$ -th player and 0.84 is the average mean over all players.

Let us consider a random sample of 24 players, stratified with respect to role (12 *backcourt* and 12 *frontcourt*). Data are shown in Table 4. We want to test if the proportion of good performances in the *home* matches is different from the proportion of good performances in the *away* matches for at least one of the two response variables.

The significance level of the test is set at  $\alpha = 0.10$ . The application of the combined permutation test, using  $B = 10\,000$  CMC runs with *Fisher*, *Liptak* and *Tippett* combining function provides the  $p$ -values 0.003, 0.004 and 0.026 respectively (see Figure 1). Hence, according to all three combined permutation tests, the null hypothesis of equal performance in the *away* and *home* matches is rejected in favor of the alternative hypothesis that the performance depends on the field factor. Note that the partial  $p$ -values of the univariate tests of the two components of the bivariate response ( $OER$ -based and  $PIR$ -based performance) are 0.021 and 0.044 respectively, as shown in Figure 1. To attribute the significance of the overall test to one of the two partial tests or to both of them, the  $p$ -values of the two partial tests must be adjusted. This is necessary to avoid the probability of type I error in the overall test exceeding the nominal significance level  $\alpha$ . The  $p$ -values of the two partial tests, adjusted with the well-known Bonferroni-Holm

**Table 4: Sample data about PIR-based and OER-based performance of players of the Italian championship in the 2016/2017 regular season, in the away and home matches.**

<i>Player</i>		$X_1(PIR)$		$X_2(OER)$	
<i>Name</i>	<i>Role</i>	<i>Away</i>	<i>Home</i>	<i>Away</i>	<i>Home</i>
Alibegovic	backcourt	0	0	0	1
Bushati	backcourt	0	0	0	0
Cournooh	backcourt	0	1	0	1
Dowdell	backcourt	1	1	0	1
Forray	backcourt	0	0	0	0
Harvey	backcourt	1	0	0	1
Mian	backcourt	0	0	0	1
Obasohan	backcourt	0	0	0	0
Randolph	backcourt	0	1	0	1
Spanghero	backcourt	0	0	0	1
Vitali	backcourt	0	1	0	1
Tonut	backcourt	1	1	1	1
Abass	frontcourt	0	1	1	1
Cusin	frontcourt	0	1	0	1
Fesenko	frontcourt	1	1	1	0
Iannuzzi	frontcourt	0	1	0	1
Kangur	frontcourt	0	0	0	1
Mazzola	frontcourt	0	0	1	1
Pascolo	frontcourt	0	1	1	1
Sacchetti	frontcourt	1	1	1	1
Thomas A.	frontcourt	0	1	0	0
Watt	frontcourt	1	1	1	1
Wojciechowski	frontcourt	0	0	1	1
Viggiano	frontcourt	0	0	1	0



**Figure 1: P-values of the combined permutation McNemar tests with Fisher, Liptak and Tippett combination for the two-tailed alternative hypothesis**

method, are both significant (0.042 and 0.044 respectively). Hence, the performance of the players in the *home* matches is not equal to their performance in the *away* matches. This conclusion concerns both the *Performance Index Rating* and the *Offensive Efficiency Rating*.

It is worth noting that the method can also be applied to directional tests, i.e. with one-tailed alternatives. For example, the alternative hypothesis could be  $H_1 : [P(X_{1H} = 1) > P(X_{1A} = 1)] \cup [P(X_{2H} = 1) > P(X_{2A} = 1)]$ , where  $(X_{1H} = 1)$  and  $(X_{1A} = 1)$  mean that the seasonal performance according to *OER* in the *home* and *away* matches respectively is good and  $(X_{2H} = 1)$  and  $(X_{2A} = 1)$  have a similar meaning for *PIR*. In fact, it is reasonable to think that the performance at home is better than the performance away according to both partial indices. In other words, the probability of good performance at *home* is higher than the probability of good performance *away*. This multivariate test with restricted alternatives (one-tailed alternative hypotheses) admits a difficult asymptotic solution also for  $q = 2$ , where the normal approximation for the two marginal distributions would be assured but with an unknown approximation rate for finite  $n$ , such as  $n = 24$ . Therefore, the application of a parametric approach based on the assumption of (approximately) normal underlying distribution is not suitable because this assumption is not plausible with these sample sizes. Hence, in these conditions, the proposed solution is appropriate and valid because distribution-free and robust with respect to the departure from normality. For the one-tailed test with  $B = 10\,000$ , we obtained

the partial  $p$ -values  $\hat{\lambda}_{PIR} = 0.0201$  and  $\hat{\lambda}_{OER} = 0.0112$ , and the Liptak combined  $\hat{\lambda}_{TL} = 0.0013$ . Hence, we have empirical evidence that the performance at home is better (home-field effect) and this is true for both the performance measures considered in the study.

## 5. CONCLUSIONS

A solution to a multivariate version of the well-known McNemar test, has been proposed. The method is based on the NPC of dependent permutation tests. The case study relates to basket analytics. Specifically, the goal is to evaluate the performance of basket players of the Italian championship (2016/2017 regular season) in order to test the so-called field effect. In other words, the goal is to test whether, according to a given list of response variables, the proportion of good performant players in the *away* matches is equal to the proportion of good performant players in the *home* matches or not.

The proposed non parametric test is flexible, robust, unbiased and consistent with respect to departure from assumptions in at least one component of the multivariate distribution of the response. It is particularly interesting to underline that the NPC procedure does not require any specific assumption about the dependence structure of the dichotomous components of the multivariate response. Indeed, the dependence structure is implicitly considered without the need of modelling or estimating any unknown population nuisance parameters (Pesarin and Salmaso, 2010).

## References

- Agresti, A. and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. In *Applied Statistics*. 54:691-816.
- Akazawa, K., Kagara, N., Sota, Y., Motooka, D., Nakamura, S., Miyake, T. and Shimazu, K. (2021). Comparison of the multigene panel test and Oncoscan<sup>TM</sup> for the determination of Her2 amplification in breast cancer. In *Oncology Reports*. 46(4):1-8.
- Becker, M. and Balagtas, C. (1993). Marginal modeling of binary cross-over data. In *Biometrics*. 49:997-1009.
- Bennett, B. and Underwood, R. (1970). On McNemar's test for the  $2 \times 2$  table and its power function. In *Biometrics*. 26:339-343.



- Bonnini, S., Corain, L., Marozzi, M., and Salmaso, L. (2014). *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R*. Wiley, Chichester.
- Brown, M. and Sokol, J. (2010). An improved Lrnc method for Ncaa basketball prediction. In *Journal of Quantitative Analysis in Sports*. 6(3):1-23.
- Cooper, W., Ruiz, J. and Sirvent, I. (2009). Selecting non-zero weights to evaluate effectiveness of basketball players with dea. In *European Journal of Operational Research*. 195:563-574.
- Demsar, J. (2006). Statistical comparisons of classifiers over multiple datasets. In *Journal of Machine Learning Research*. 7:1-30.
- Deshpande, S. and Jensen, S. (2016). Estimating an Nba player's impact on his team's chances of winning. In *Journal of Quantitative Analysis in Sports*. 12:51-72.
- Donald, A. and Donner, A. (1987). Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. In *Statistics in Medicine*. 6:491-499.
- Donald, A. and Donner, A. (1990). A simulation study of the analysis of sets of 2x2 contingency tables under cluster sampling: Estimation of a common odds ratio. In *Journal of the American Statistical Association*. 85:537-543.
- Donner, A. (1992). Sample size requirements for stratied cluster randomization designs. In *Statistics in Medicine*. 11:743-750.
- Durkalski, V., Palesch, Y., Lipsitz, S., Philip, F. and Rust, P. (2003). The analysis of clustered matched-pair data. In *Statistics in Medicine*. 22:2417-2428.
- Eliasziw, M. and Donner, A. (1991). Application of the McNemar test to non-independent matched pair data. In *Statistics in Medicine*. 10(12): 1981-1991.
- Fearnhead, P. and Taylor, B. (2011). On estimating the ability of Nba players. In *Journal of Quantitative Analysis in Sports*. 7(3):11.
- Feuer, E. and Kessler, L. (1989). Test statistic and sample size for a two-sample McNemar test. In *Biometrics*. 45:629-636.
- Gonen, M. (2004). Sample size and power for McNemar's test with clustered data. In *Statistics in Medicine*. 23(14):2283-2294.

- Gupta, A. (2015). A new approach to bracket prediction in the Ncaa men's basketball tournament based on a dual-proportion likelihood. In *Journal of Quantitative Analysis in Sports*. 11:53-67.
- Ibrahim, A., Kashef, R. and Corrigan, L. (2021). Predicting market movement direction for bitcoin: A comparison of time series modeling methods. In *Computers Electrical Engineering*. 89:106905.
- Joe, H. (1997). *Multivariate Methods and Dependence Concepts*. Chapman Hall, London.
- Klingenberg, B. and Agresti, A. (2006). Multivariate extensions of McNemar's test. In *Biometrics*. 62:921-928.
- Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering*. Wiley, Hoboken, New Jersey.
- Lachin, J. (1992). Power and sample size evaluation for the McNemar test with application to matched case-control studies. In *Statistics in Medicine*. 11(9):1239-1251.
- Leisenring, W., Alonzo, T. and Pepe, M.S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. In *Biometrics*. 56:345-351.
- Loeffelholz, B., Bednar, E. and Bauer, K. (2009). Predicting Mba games using neural networks. In *Journal of Quantitative Analysis in Sports*. 5(1):1-17.
- Lopez, M. and Matthews, G. (2015). Building and Ncaa men's basketball predictive model and quantifying its success. In *Journal of Quantitative Analysis in Sports*. 11(1):5-12.
- Lyles, R., Williamson, J., Lin, H. and Heilig, C. (2005). Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. In *Biometrics*. 61:287-294.
- Mantel, N. and Fleiss, J. (1975). The equivalence of the generalized McNemar tests for marginal homogeneity in  $2^3$  and  $3^2$  tables. In *Biometrics*. 31:731-735.
- McKinlay, S. (1975). A note on the chi-square test for pair-matched samples. In *Biometrics*. 31:731-735.

- McNemar, Q. (1947). A note on the sampling error of the difference between correlated proportions and percentages. In *Psychometrika*. 12: 153-157.
- Obuchowski, N. (1998). On the comparison of correlated proportions for clustered data. In *Statistics in Medicine*. 17:1495-1507.
- Ozmen, U. (2012). Foreign player quota, experience and efficiency of basketball players. In *Journal of Quantitative Analysis in Sports*. 8:1-18.
- Page, G., Fellingham, G. and Reese, C. (2007). Using Box-scores to determine a position's contribution to winning basketball games. In *Journal of Quantitative Analysis in Sports*. 3(4):1-16.
- Pembury Smith, M. and Ruxton, G. (2020). Effective use of the McNemar test. In *Behavioral Ecology and Sociobiology*. 74, 133.
- Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. Wiley, Chichester.
- Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Applications and Software*. Wiley, Chichester.
- Piette, J., Anand, S. and Zhang, K. (2010). Scoring and shooting abilities of Nba players. In *Journal of Quantitative Analysis in Sports*. 6(1):1-24.
- Rao, J. and Scott, A. (1992). A simple method for the analysis of clustered binary data. In *Biometrics*. 48(2):577-585.
- Ruiz, F. and Perez-Cruz, F. (2015). A generative model for predicting outcomes in college basketball. In *Journal of Quantitative Analysis in Sports*. 11(1):39-52.
- Shao, E., Liu, C., Wang, L., Song, D., Guo, L., Yao, X. and Hu, Y. (2021). Artificial intelligence-based detection of epimacular membrane from color fundus photographs. In *Scientific Reports*. 11(1):1-10.
- Shortridge, A., Goldsberry, K. and Adams, M. (2014). Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball. In *Journal of Quantitative Analysis in Sports*. 10:1-11.
- Stransky, J., Bassett, L., Bodnar, C., Anastasio, D., Burkey, D. and Cooper, M. (2021). A retrospective analysis on the impacts of an immersive digital environment on chemical engineering students' moral reasoning. In *Education for Chemical Engineers*. 35:22-28.

- Ury, H. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. In *Biometrics*. 3:643-650.
- West, B. (2006). A simple and flexible rating method for predicting success in the Ncaa tournament outcomes. In *Journal of Quantitative Analysis in Sports*. 2(3):1-16.
- Westfall, P., Troendle, J. and Pennello, G. (2010). Multiple McNemar tests. In *Biometrics*. 66(4):1185-1191.
- Wu, Y. (2018). Power calculation of adjusted McNemar's test based on clustered data of varying cluster size. In *Biometrical Journal*. 60(6):1190-1200.
- Yuan, L., Liu, A., Yeh, A. and Kaufman, A. (2015). A mixture-of-modelers approach to forecasting Ncaa tournament outcomes. In *Journal of Quantitative Analysis in Sports*. 11(1):13-27.
- Zuccolotto, P. and Manisera, M. (2020). *Basketball Data Science: With Applications in R*. Chapman Hall/CRC, -.



## WHEN IS 2 BETTER THAN 3 IN BASKETBALL?

**Phillip Sa-Ut Gjøen, Sofus August Hvattum, Eirik Malme Moltubak,  
Lars Magnus Hvattum<sup>1</sup>**

*Faculty of Logistics, Molde University College, Molde, Norway*

**Abstract** A recent trend in basketball is that teams are taking more shots outside of the three-point line and fewer shots inside. This is an advantage since the expected number of points scored, in general, is slightly higher for three-point shots. Through simulations, this paper shows that there are game situations where a strategy of taking fewer three-point attempts at the expense of more two-point attempts will improve the probability of winning the game.

**Keywords:** *Simulation, Coaching, Strategy, Sports.*

### 1. INTRODUCTION

The National Basketball Association (NBA) is continuously evolving over time. A recent trend involves teams attempting an increasing number of three-point field goals (Rocha da Silva and Rodrigues, 2021), based on analytics showing this to be a superior strategy in terms of maximizing the expected number of points scored per possession.

Skinner and Goldman (2017) pointed out, from a theoretical perspective, that it may be beneficial in certain situations to aim for two-point shots instead of three-point shots, even if the latter leads to higher expected points. Since two-point shots are converted more frequently, they lead to lower variance in the total score at the expense of potentially lower expected values.

In this paper we examine the following question using real-world data: Are there realistic situations that frequently appear in NBA games where teams would benefit from tilting their shot selection strategy in favor of taking more two-point attempts? We answer this question using simulations, while deriving simple guidelines that may be followed by basketball coaches to guide their teams towards increased sporting success.

For almost seventy years, researchers have proposed that coaches can use scientific methods to make improvements in the way that their teams perform

---

<sup>1</sup>Corresponding author: Lars Magnus Hvattum, hvattum@himolde.no, ORCID: 0000-0003-0490-9978

(Wright, 2009). An early example in basketball was the use of statistical models to evaluate players with adjusted plus-minus ratings (Winston, 2009), which has since evolved into ever more complex and powerful models (Engelmann, 2017), and has been adopted within a range of different sports (Hvattum, 2019).

Nikolaidis (2015) suggested that basketball teams can improve their decision-making processes significantly by choosing to employ statistical analysis of basketball data. In a recent review, Terner and Franks (2021) focused on research that models the performance of players and teams, while also discussing different sources of data and related software tools for data retrieval. Recent advances in this field involve using detailed tracking data (Bornn et al., 2017), but there is still much insight that can be gained also with simpler data sources, such as box scores.

An important concept in the analysis of basketball is the idea of possessions (Kubatko et al., 2007). A given possession begins when a team gains control of the ball, and lasts until the team no longer has control. Possessions can thus end after converting a shot, after missing a shot leading to a defensive rebound, or after a turnover. Since the end of one possession is followed by the beginning of a new possession by the other team, the two teams involved in a game always have approximately the same number of possessions.

The four-factor model of Oliver (2004) is a seminal work within basketball analytics. It proposes that the offensive rating of a team decomposes into four distinct qualities: the effective field goal percentage, the turnover percentage, the offensive rebound percentage, and the free throw attempt rate. Improving these areas of play, a team can improve its win percentage. Cecchin (2022) used structural equations modelling to validate the four-factor model, finding that the four factors are relevant in explaining teams' winning ability. When analyzing high-level European basketball, Charamis et al. (2022) found a slightly better model for win percentages, using a true shooting percentage instead of the effective field goal percentage and the free throw attempt rate.

Annis (2006) analyzed optimal end-game strategy, finding that intentionally fouling is a better strategy than playing tight defense to protect a small lead at the end of a game. McFarlane (2019) used logistic regression to find win probabilities and then created an end-of-game tactics metric to evaluate on-court decisions. One application of this is to find the time at which intentionally fouling becomes the optimal tactic for a given score differential. Christmann et al. (2018) used video-analysis to investigate offensive play types in the final two minutes of 115 close NBA games. Findings included that coaches should instruct their teams to

attempt transition play whenever possible, and that for set plays more complex play types are more effective.

The number of three-point and two-point attempts made has occasionally been studied in the scientific literature. Csataljay et al. (2009) analyzed games from the European Basketball Championship of 2007 and found that winning teams had a higher conversion rate for three-point attempts, while having fewer three-point attempts. Ibáñez et al. (2008) studied the Spanish Basketball League, finding no statistically significant differences between the best and the worst teams when it comes to successful, nor unsuccessful, two-point and three-point attempts.

This contrasts with analysis of modern era NBA games: Rocha da Silva and Rodrigues (2021) observed that between 2014 and 2019 three-point attempts and conversions had a positive effect on the performance of teams, while two-point conversions started to be a negative factor and then turned into a non-factor. Mandić et al. (2019) compared statistics from the NBA and the Euroleague between 2000 and 2017. They found that the number of three-point attempts in the NBA had almost doubled in the examined time period, while the number of three-point attempts in the Euroleague had increased by a much smaller magnitude. Fichman and O'Brien (2019) split the court into 11 zones, and used game theory to find optimal mixed strategies for which zones to use when making shots. They concluded that NBA is headed for a future with a higher number of three-point attempts, with an equilibrium analysis suggesting on average 62.1% two-point shots and 37.9% three-point shots.

The remainder of this paper is structured as follows. In Section 2 we describe the data used to find appropriate inputs to our simulations. Section 3 presents our simulation framework. Results and analyses are given in Section 4, followed by conclusions in Section 5.

## 2. DATA

The main source of data is <https://www.basketball-reference.com>. We extracted team statistics per 100 possessions for seven seasons of the NBA, from 2015/2016 to 2021/2022. These statistics are based on 82 games for each of 30 teams, except for the 2019/2020 and 2020/2021 seasons when fewer games were played due to an epidemic infectious disease. We thus focus on regular season games, and exclude the play-offs. The attributes extracted include the number of three-point field goal attempts (3PA), the number of two-point field goal attempts (2PA), the three-point field goal percentage (3P%), and the two-point field goal percentage (2P%). Table 1 summarizes the number of two-point



**Table 1: Descriptive statistics from seven recent seasons of the NBA, reporting the number of two-point attempts and the number of three-point attempts per 100 possessions for different teams. Data source: basketball-reference.com**

Season	2PA			3PA		
	Min.	Avg.	Max.	Min.	Avg.	Max.
2015/2016	53.5	62.7	70.2	16.4	24.9	31.5
2016/2017	46.7	60.2	68.5	22.1	27.8	40.1
2017/2018	42.8	58.3	66.0	23.3	29.6	43.2
2018/2019	42.6	56.8	63.8	25.2	31.8	46.0
2019/2020	43.3	54.2	61.4	28.0	33.7	43.4
2020/2021	45.6	53.9	62.1	27.7	34.7	43.5
2021/2022	47.1	53.6	61.2	29.3	35.6	41.4

attempts and three-point attempts for different teams, while Table 2 shows the corresponding conversion rates.

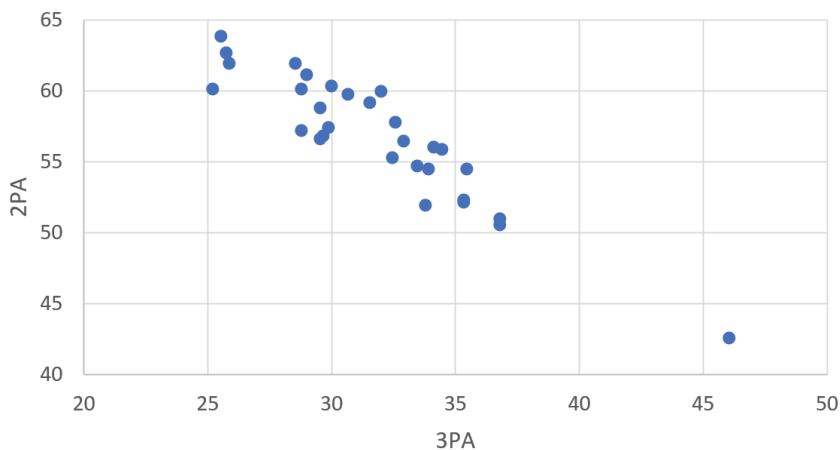
The team-based statistics show that there has been an evolution in shot strategies in the NBA over the span of these seven seasons. The number of two-point attempts has declined, while the number of three-point attempts has increased, in particular when considering the average across teams. While the conversion rates for three-point shots have been relatively stable across time, the conversion rates for two-point shots have improved.

In the following, we focus in particular on the 2018/2019 regular season, which was the last season prior to the playing schedules being interrupted due to pandemic-induced restrictions. Figure 1 shows the number of field goal attempts of each type per 100 possessions for each of the teams in the 2018/2019 regular season. Naturally, teams with many three-point attempts have, in general, fewer two-point attempts and vice versa. The outlier with the highest 3PA is the Houston Rockets, with 46 three-point attempts per 100 possessions. The field goal percentages per team are illustrated in Figure 2. The conversion percentages vary from 33% to 39% for three-point attempts and from 48% to 57% for two-point attempts.

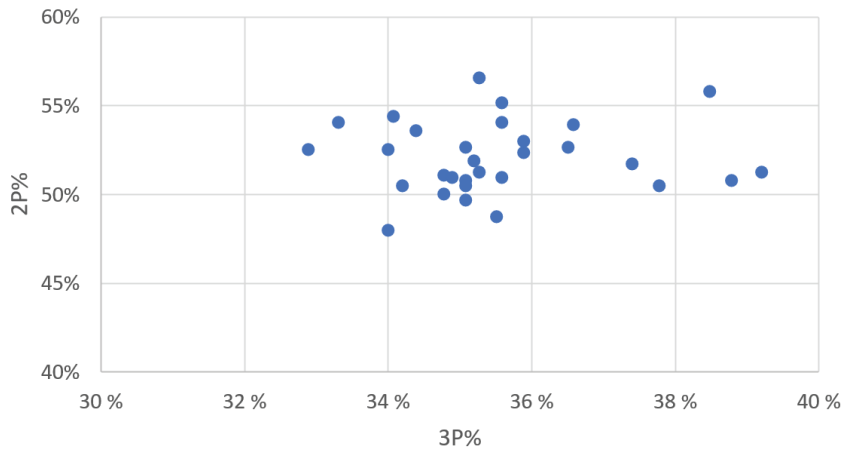
When looking at the number of attempts and the conversion rates, an important observation is that the conversion rates do not vary to a large degree with the number of attempts. This is illustrated for three-point field goals in Figure 3, and similar relationships were found for attempts and conversions of two-point field goals.

**Table 2: Descriptive statistics from seven recent seasons of the NBA, reporting the conversion rates for two-point attempts and three-point attempts for different teams. Data source: basketball-reference.com**

Season	2P%			3P%		
	Min.	Avg.	Max.	Min.	Avg.	Max.
2015/2016	45.4 %	49.2 %	52.8 %	31.7 %	35.3 %	41.6 %
2016/2017	47.3 %	50.4 %	55.7 %	32.7 %	35.7 %	39.1 %
2017/2018	47.8 %	51.1 %	56.0 %	33.4 %	36.2 %	39.1 %
2018/2019	47.9 %	52.0 %	56.5 %	32.9 %	35.6 %	39.2 %
2019/2020	48.9 %	52.4 %	56.7 %	33.3 %	35.8 %	38.0 %
2020/2021	47.6 %	53.1 %	56.5 %	33.6 %	36.6 %	41.1 %
2021/2022	49.7 %	53.3 %	57.5 %	32.3 %	35.4 %	37.9 %



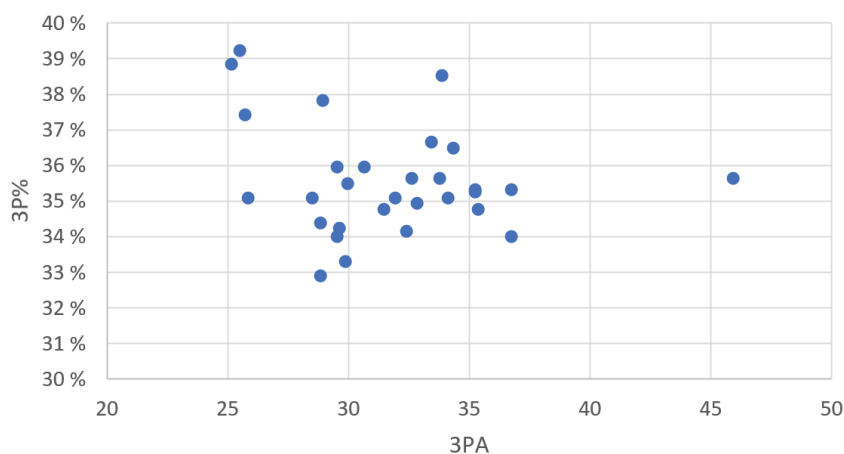
**Figure 1: Number of two-point and three-point attempts per 100 possessions for each of the 30 teams participating in the 2018/2019 season. Data source: basketball-reference.com**



**Figure 2: Field goal percentages for two-point and three-point attempts for each team in the 2018/2019 season. Data source: [basketball-reference.com](https://basketball-reference.com)**

To elaborate on these relationships, we ran simple linear regressions with the conversion rates as dependent variables and the number of attempts as independent variables, using team observations from all seven seasons in the data set. When 3P% is regressed on 3PA, the regression coefficient is very close to zero, but statistically significant with a p-value of 0.035. The coefficient implies that the conversion rate for three-point shots increases by 0.04 percentage points for each additional attempt per 100 possessions, which is very low. For 2P% regressed on 2PA, the regression coefficient of 2PA implies a decrease of 0.28 percentage points in the conversion rate per additional attempt per 100 possessions, and the coefficient is highly significant with a p-value that is essentially 0. However, including additional independent variables, such as free throw conversion rates and total points scored per 100 possessions, is associated with a reduction in the magnitude of the regression coefficient of 2PA.

We can expect that the number of free throw attempts depends on the shot selection strategy, since a player fouled within the three-point line is awarded two free throws, whereas a player fouled outside of the three-point line is awarded three free throws. Using the full data set with 280 team observations, we regressed the number of free throw attempts per 100 possessions on the number of two-point attempts and three-point attempts, respectively. We find that two-point attempts are not significant at explaining the number of free throw attempts, with



**Figure 3:** For three-point field goals, the relationship between attempts and conversion rates for each team in the 2018/2019 season. Data source: basketball-reference.com

a p-value of 0.31, whereas three-point attempts are significant with a p-value of 0.003. The regression coefficient of 3PA indicates that the number of free throw attempts decreases by 0.08 for each three-point shot attempted. Overall, there is no evidence of a strong relationship between free throw attempts and the number of two-point or three-point attempts.

### 3. EXPERIMENTAL SETUP

To compare the effect of different shot selection strategies, or in other words the effect of teams choosing to make more two-point attempts at the expense of three-point attempts, we use discrete-event simulation. The simulation takes as input the values of 3PA, 2PA, 3P%, and 2P% for each of two teams. In addition, it takes as input the current point difference and the number of remaining possessions per team. Considering the number of possessions remaining is a simplification, since in reality there is a game clock that determines how long the game lasts, and the number of possessions is unknown a priori. However, defining the remainder of a game through the number of remaining possessions per team makes the results easy to interpret.

The simulation then considers each remaining possession and, according to the given shot selection probabilities, randomly determines that the possession

**Table 3: Alternatives explored for shot selection strategies**

Style	2PA	3PA
Two-point focus	63.9%	25.5%
Balanced	57.1%	32.0%
Three-point focus	42.6%	46.0%

ends with a three-point shot, a two-point shot, or no shot. Then, if a shot is taken, according to the given shot conversion probabilities we draw whether the shot is successful, and then adjust the point difference. The simulation does not consider free throws.

When all possessions have been processed, the simulation terminates with a final point difference. However, if the final point difference is 0, extra time is needed to determine a winner. From <http://stats.inpredictable.com/>, we find that the average time per possession is slightly less than 15 seconds. Since overtime in the NBA lasts five minutes, we therefore use 10 possessions per team when simulating the overtime. Should the overtime also end with a draw, another overtime period is started.

Table 3 shows three alternative settings for the shot selection strategy of a team. For the analysis, input numbers are based primarily on the 2018/2019 regular season. The two-point focus strategy is based on the statistics of the team with most two-point attempts in that season, the San Antonio Spurs, while the three-point focus strategy is based on the team with the most three-point attempts, the Houston Rockets. The balanced strategy is based on the average of all the teams in the 2018/2019 season. However, since we want to analyze a situation where two teams have the same expected number of points per possession while following different shot selection strategies, the numbers given in the table are slightly adjusted, so that each strategy is made sure to produce the same expected number of points when executed by a team with an average conversion quality.

Three alternative settings for the quality of teams are reported in Table 4. Here, a good team corresponds to having the maximum conversion rates among all teams in the league for both types of shots considered. Correspondingly, an average team has the average conversion rates, and a bad team has the minimum conversion rates. With the given conversion rates, all three types of teams obtain a higher expected points total when using 3-point shots rather than 2-point shots, with an expected difference in the range of 0.029 to 0.046 points per shot.

**Table 4: Alternatives explored for team quality settings**

Quality	2P%	3P%
Good	56.5%	39.2%
Average	52.0%	36.0%
Bad	47.9%	32.9%

The experiments take into account a number of remaining possessions per team, ranging from 0 to 30, with a starting point difference between  $-10$  and  $10$ . A focal team, team 1, has a choice between two shot selection strategies: focusing on two-pointers or focusing on three-pointers, whereas the opposing team, team 2, has a fixed average strategy. The motivation behind this is to observe, from the perspective of team 1, what happens when going from a strategy favoring three-point shots to a strategy favoring two-point shots.

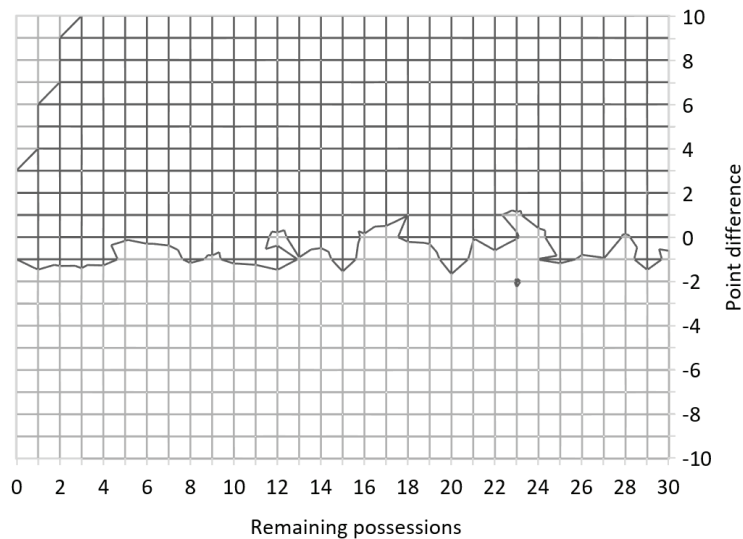
For the team quality we consider three alternatives: either both teams are average, and are thus expected to score the same number of points per possession independent of the selected shot strategy, or one of the teams is good and the other team is bad.

For each combination of remaining possessions and point difference we simulate 100,000 games with team 1 having a two-point focus and 100,000 games with team 1 having a three-point focus. We then calculate the difference in the number of wins for team 1, which then is used to conclude which shot selection strategy is to be preferred in a given situation.

#### 4. RESULTS AND ANALYSIS

We start by showing the results for two equally good teams playing against each other in Figure 4. The area of the figure with darker color shows game situations where a preference towards two-point shots leads to more wins than a strategy with more three-point shots. For this setting, it is clear that the two-point focus is beneficial as soon as a team is in the lead, whereas a three-point focus is best when a team is trailing.

Figure 5 shows the corresponding figure when team 1 is better than team 2. In this case the two-point focus strategy is beneficial in more situations: even if the team is trailing by a few points, going for two-point shots can be good. Since the other team is weaker, a less risky strategy is sufficient to maximize the winning chances.



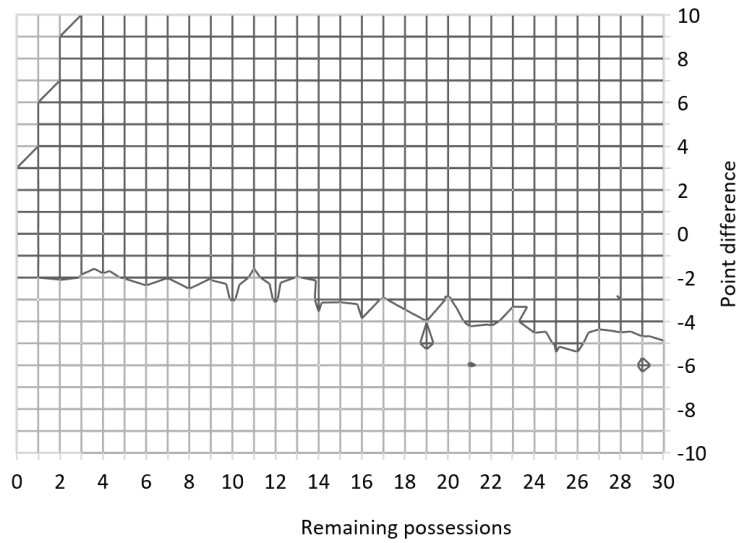
**Figure 4: Best strategy for shot selection when two equally good teams play against each other and the second team has an average shot selection strategy, with dark color indicating situations where two-point focus is beneficial**

Finally, Figure 6 illustrates the result for a bad team playing against a good team. In this case, if many possessions are left of the game, it may still be necessary to go for three-point shots when having a slight lead, as the more conservative two-point strategy is not sufficient to defeat the stronger opponent.

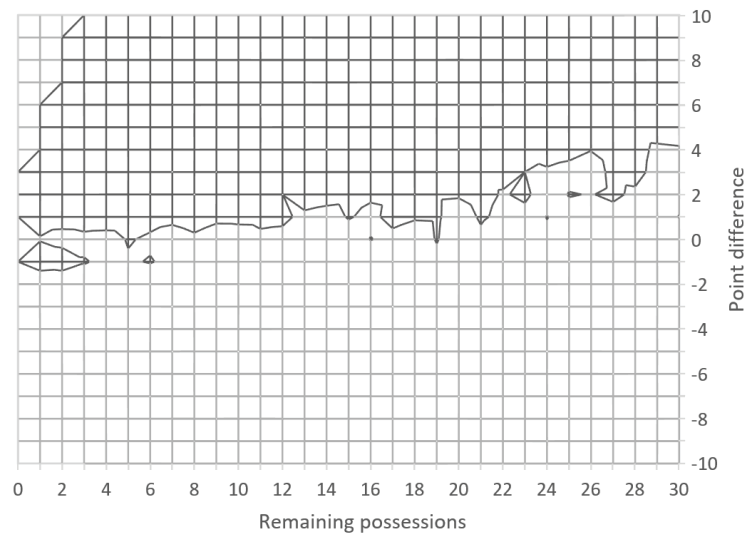
The figures conveniently demarcate the situations where a team may benefit from making more two-point attempts and fewer three-point attempts. However, they do not show whether the difference is large enough to warrant coaches to consider the effect. In each of the three situations analyzed, the magnitude of the differences in the number of wins when using either a two-point or a three-point focus is similar.

When taking two-point shots is better, this strategy leads to the team winning around 0.7 percentage points more games than when focusing on three-point shots. When taking three-point shots is better, the team also wins around 0.7 percentage points more of the games. This, however, is on average across all game states within the two regions of each figure.

Looking at the maximum numbers, there are certain states where the number of games won can change by up to 1.5 percentage points, and one example giv-



**Figure 5: Best strategy for shot selection when a good team plays against a bad team with an average shot distribution, with dark color indicating situations where two-point focus is beneficial**



**Figure 6: Best strategy for shot selection when a bad team plays against a good team with an average shot distribution, with dark color indicating situations where two-point focus is beneficial**



ing a change of 3 percentage points. The latter happens in an extreme situation where each team has a single possession left, the focal team is good and is trailing by three points. In this case, the team must first score on a three-point attempt and then the opponent must fail to score in their attack. The three-point strategy then leads to the team winning 6.9% of their games, compared to 3.8% for the two-point strategy. However, most of the cases with the large difference between strategies are less extreme, such as leading by four points with ten possessions remaining against an evenly matched opponent, where a two-point focus leads to 78.2% wins compared to 76.9% for a three-point focus.

## 5. CONCLUSIONS

Teams in the NBA use different strategies, leading to different distributions of two-point attempts and three-point attempts. In recent years, the number of three-point attempts has increased, based on observations that this leads to a higher expected number of points per possession.

From a theoretical point of view, it is clear that if two-point attempts and three-point attempts have a similar expected value, the difference in variance may lead to either two-point attempts or three-point attempts being better for maximizing a team's winning chances.

This paper has provided numerical experiments using simulations based on realistic shot selection strategies and conversion percentages. When two equally good teams face each other, a team that is in the lead benefits from increasing the number of two-point attempts at the expense of three-point attempts, while a team that is trailing should prefer to go for more three-point attempts. When one team is better than the other, a similar strategy is useful, but the team can be more conservative, and can prefer two-point attempts even when slightly behind, in particular if there is more time left of the game. On the other hand, a weaker team must be more willing to take risk by predominantly going for three-points also when slightly in the lead, assuming that there are many possessions left.

The interpretation of the results rests on several assumptions, thus suggesting some limitations of the analysis. First, free throws have been neglected. Taking into account free throws requires additional information about the probability of being fouled conditional on the selected shot strategy. Second, it is assumed that shot selection strategies do not influence conversion rates. This may be false if the shot selection strategies are very different, such as solely taking two-point attempts, as the defending team can adapt their strategies accordingly. However, the shot selection strategies compared in this study are strategies actually applied

by teams in the 2018/2019 NBA regular season, and the true change in conversion rates when modifying the shot selection accordingly could be relatively small. A third limitation is that the shot selection strategies are assumed to remain fixed throughout the remainder of the game in the simulations. In practice, a team can change strategies dynamically based on the change in point differences.

This study may be extended by considering each of the above limitations. As the shot selection strategies implemented in the NBA are still evolving, future research may investigate whether this evolution leads to different conclusions than when focusing on the strategies applied within the 2018/2019 regular season. In addition, there are some differences between the NBA and other basket leagues, such as the top leagues in Europe. Therefore, using data from other competitions may lead to slightly different results. For example, in Euroleague, the three-point shot line distance is shorter, the number of fouls is higher, and the number of possessions per game is lower (Mandić et al., 2019).

To conclude, this study may provide some balancing inputs to coaches when observing that strategies involving an increased number of three-point attempts become more successful: while three-point focused strategies may lead to better expected scores, certain game situations imply that two-point focused strategies improve the probabilities of winning a game. Our simulations suggest that such game situations are perhaps appearing more frequently than expected: It is not only in rare situations where a team is one point behind and has a single possession left that a two-point attempt may be best, but also in close games where a team is slightly ahead against an evenly matched opponent.

#### ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous reviewers for their insightful comments, which helped to improve the manuscript.

#### References

- Annis, D.H. (2006). Optimal end-game strategy in basketball. In *Journal of Quantitative Analysis in Sports*, 2 (2).
- Bornn, L., Cervone, D., Franks, A. and Miller, A. (2017). Studying basketball through the lens of player tracking data. In J. Albert, M. Glickman, T. Swartz, and R. Koning, eds., *Handbook of Statistical Methods and Analyses in Sports*, 245–269. Chapman and Hall/CRC, Boca Raton.

- Cecchin, A. (2022). Oliver's four-factor model: Validation through causality. In *Sports Science & Coaching*, 11. Forthcoming.
- Charamis, E., Marmarinos, C. and Ntzoufras, I. (2022). Estimating team possessions in high-level European basketball competition. In *Sports Science & Coaching*, 11. Forthcoming.
- Christmann, J., Akamphuber, M., Müllenbach, A.L. and Güllich, A. (2018). Crunch time in the NBA – The effectiveness of different play types in the endgame of close matches in professional basketball. In *International Journal of Sports Science & Coaching*, 13: 1090–1099.
- Csataljay, G., O'Donoghue, P., Hughes, M. and Dancs, H. (2009). Performance indicators that distinguish winning and losing teams in basketball. In *International Journal of Performance Analysis in Sport*, 9: 60–66.
- Engelmann, J. (2017). Possession-based player performance analysis in basketball (adjusted +/- and related concepts). In J. Albert, M. Glickman, T. Swartz and R. Koning, eds., *Handbook of Statistical Methods and Analyses in Sports*, 215–228. Chapman and Hall/CRC, Boca Raton.
- Fichman, M. and O'Brien, J.R. (2019). Optimal shot selection strategies for the NBA. In *Journal of Quantitative Analysis in Sports*, 15: 203–211.
- Hvattum, L.M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. In *International Journal of Computer Science in Sport*, 18: 1–23.
- Ibáñez, S.J., Sampaio, J., Feu, S., Lorenzo, A., Gómez, M.A. and Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. In *European Journal of Sport Science*, 8: 369–372.
- Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D. (2007). A starting point for analyzing basketball statistics. In *Journal of Quantitative Analysis in Sports*, 3 (3): 1.
- Mandić, R., Jakovljević, S., Erčulj, F. and Štrumbelj, E. (2019). Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017. In *PLoS One*, 14 (10): e0223524.
- McFarlane, P. (2019). Evaluating NBA end-of-game decision-making. In *Journal of Sports Analytics*, 5: 17–22.

- Nikolaidis, Y. (2015). Building a basketball game strategy through statistical analysis of data. In *Annals of Operations Research*, 227: 137–159.
- Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books Inc, Dulles, VA, USA.
- Rocha da Silva, J.V. and Rodrigues, P.C. (2021). The three eras of the NBA regular seasons: Historical trend and success factors. In *Journal of Sports Analytics*, 7: 263–275.
- Skinner, B. and Goldman, M. (2017). Optimal strategy in basketball. In J. Albert, M. Glickman, T. Swartz and R. Koning, eds., *Handbook of Statistical Methods and Analyses in Sports*, 245–260. Chapman and Hall/CRC, Boca Raton.
- Terner, Z. and Franks, A. (2021). Modeling player and team performance in basketball. In *Annual Review of Statistics and Its Application*, 8: 1–23.
- Winston, W.L. (2009). *Mathletics*. Princeton University Press, Princeton, New Jersey.
- Wright, M.B. (2009). 50 years of OR in sport. In *Journal of the Operational Research Society*, 60: S161–S168.



## USING (COPULA) REGRESSION AND MACHINE LEARNING TO MODEL AND PREDICT FOOTBALL RESULTS IN MAJOR EUROPEAN LEAGUES

**Hendrik van der Wurp, Andreas Groll**<sup>1</sup>

*Department of Statistics, TU Dortmund University, Dortmund, Germany*

**Abstract** *In this manuscript, we compare classical univariate regression approaches with copula models explicitly accounting for the dependency structure as well as with modern machine learning techniques in the context of modelling and predicting of football results in the major European leagues. Particularly, we want to present an extensive data set compiled from publicly available sources containing data and match results from the first men's football divisions from England, France, Germany, Italy, Spain (often referred to as the "big five"), the Netherlands and Turkey. We introduce several modelling approaches to predict upcoming matches and compare their predictive strengths. The gathered data set is presented in detail and made publicly available to motivate further work and modelling ideas.*

**Keywords:** *Count data regression, Football, Joint modelling, Regularisation, Application.*

### 1. INTRODUCTION

Generally, international football tournaments such as FIFA World Cups or the big confederation's championships (e.g. UEFA European Championship, CONCACAF Gold Cup, CONMEBOL Copa América) as well as international and national tournaments on the team-level are experiencing an ever increasing standing in terms of popularity and financial relevance. Also, modelling and predicting the results of sport matches and especially football matches has become a quite popular and present topic.

Even though no gold standard approach exists to model football results, a vast selection of methods and model classes has been proposed over the years. On the observed results of scored goals per team, Poisson regression approaches have been commonly used (e.g. by Lee, 1997, or Maher, 1982). These have been extended over the years to include several team-specific covariates in combination with regularisation techniques (e.g. by Groll and Abedieh, 2013 or Groll et al., 2015). The basic Poisson approaches can be extended by including dependency

---

<sup>1</sup>Corresponding author: Hendrik van der Wurp, email: vanderwurp@statistik.tu-dortmund.de

between the numbers of goals scored by competing teams, which Dixon and Coles (1997) investigated early. In particular, the bivariate Poisson approach was then proposed in detail by Karlis and Ntzoufras (2003). A different approach to dependency is the inclusion of copulas, which McHale and Scarf (2007) used to model the number of shots-on-target. Nikoloulopoulos and Karlis (2010) promoted copulas for the application to count data in general. More recently, van der Wurp et al. (2020) and van der Wurp and Groll (2021) extensively applied copulas within the GJRM (generalised joint regression modelling) framework by Marra and Radice (2019), and added football-specific regularisation into it.

A completely different approach is to dispense with the information of the numbers of goals and to model the nominal/ordinal outcome (win first team, tie, win second team) directly. The usage of ordinal or nominal regression approaches is rather straightforward as well (and e.g. discussed in Hvattum, 2017). Leitner et al. (2010) used national team abilities (depicted by Elo ratings) and bookmakers' odds to directly obtain winning probabilities in a binary (win / loss) setting. This was extended by Tutz and Schauburger (2014) with penalisation approaches for league football data and by Schauburger et al. (2017) to analyse on-field variables such as total running distance per team. A comparison of both score- and result-based approaches has been performed by Egidi and Torelli (2021).

Besides regression approaches, random forests (originally introduced by Breiman, 2001) are a very flexible and frequently used technique in the context of predicting sports results. Random forests were used e.g. by Groll et al. (2019) and Groll et al. (2021) to model FIFA World Cup and European championship data, respectively, and to predict the latest tournament. Also with the tree-based methods, principally both score- and result-based models can be used, see, e.g., Schauburger and Groll (2018).

Bayesian approaches (see, for example, Baio and Blangiardo, 2010) are also promising, but are omitted in this work. It will examine the predictive performance of the mentioned (and some other) approaches via suitable performance measures and will also investigate potential betting results. The probabilities gathered from several online bookmakers will be used as a natural benchmark. While copula regression and the proposed football-specific penalty structures by van der Wurp et al. (2020) and van der Wurp and Groll (2021) will receive special attention, a lot of different modelling approaches and covariate settings will be benchmarked against one another.

The underlying data set was gathered in July 2021 and contains all matches from the respective first men's divisions of England, France, Germany, Italy, Spain








(the “big five”), the Netherlands, and Turkey for ten seasons between 2010 and 2020. Our data set ends just before the start of the COVID-19 pandemic, as these extraordinary circumstances are deemed to be a research topic completely on its own (postponed or completely canceled games, games with less or no fans, etc.). A growing-window approach will be used to assess the approaches’ predictive potential, where the upcoming matchday is predicted using all prior matchdays and seasons.

We present this data set in detail in Section 2 with information about available covariates. Section 3 contains brief descriptions of all used model classes, covariate settings, underlying software packages, and provides an overview about the performance indicators used in our application. The corresponding results are presented and visualised in Section 4, before we conclude in Section 5.

## 2. DATA

The data set was freely available, gathered from different websites, and published (van der Wurp, 2022). As the analysis of market values by `transfermarkt.com` was started in 2010, we chose the season of 2010/2011 as a starting point and ended in the season of 2019/2020 with the start of the COVID-19 pandemic (see end of Section 2). The sample sizes and more information by country are given in Table 1.

**Table 1: Sample sizes per league. The season 2019/2020 was called off for the Ligue 1 and the Eredivisie, while postponed and later completed in the other leagues.**

League	matches	league size	observed teams	$\overline{\text{goals}}_{\text{home}}$	$\overline{\text{goals}}_{\text{away}}$
 Premier League	3800	20	36	1.55	1.19
 Ligue 1	3700	20	34	1.46	1.07
 Bundesliga	3060	18	28	1.65	1.30
 Serie A	3800	20	34	1.52	1.19
 Primera División	3800	20	33	1.59	1.13
 Eredivisie	2988	18	26	1.80	1.34
 Süper Lig	3060	18	34	1.54	1.20

The main information of each match (teams competing, date, day of week, matchday number, and the scored goals) is easily available data and was gathered from `kicker.de` in July 2021). Other covariates are:

- **Elo** rating of each team. Calculated and gathered from <http://clubelo.com/> (July 2021; Schiefler, 2015). It ranges from 1223 (FC Dordrecht in



2014) to 2106 (Barcelona in 2012) and can be interpreted via the differences in rating, denoted by  $d = \text{Elo}_{\text{home}} - \text{Elo}_{\text{away}}$ . The probability for the home team to win is then defined as  $\pi = P(\text{HomeWin}) = 1 / \left( (10^{\frac{-d}{400}}) + 1 \right)$

win (Schiefler, 2015). Equal Elo ratings will lead to a probability of 0.5. After each match, the team's Elo scores are adjusted by  $\Delta\text{Elo} = (R - \pi) \cdot 20$  with  $R$  corresponding to the results from each team's point of view (1 for a win, 0.5 for a tie and 0 for a loss). The factor of 20 is a weight index chosen by Schiefler (2015). With this scheme, unlikely results like an underdog's win will result in bigger Elo changes.

These (or similar) types of Elo rankings are commonly used in competitive sports. It was originally proposed by Arpad Emmerich Elo (1961) to rank the ability of chess players.

- **Market value (MV)** of a team. Determined and gathered from [transfermarkt.com](https://www.transfermarkt.com) (July 2021). Given in million euro and ranges from 2.8 (FC Dordrecht in 2014) to 1,300 (Manchester City in 2019/20). The market values of [transfermarkt.com](https://www.transfermarkt.com) are a community project, where each player's market value is discussed and determined by (known or rumoured) transfer fees and the player's standing in his team. The team's value is the simple sum of its current players. The values are updated twice a month to timely include transferred players. The earliest available data is from 2010-11-01, so missing values occur for the first matchdays of the season 2010/11. As the market values are growing over time, we are transforming the raw values to shares of the league's market value, using each matchday's sum as a total market value. Missing values are imputed as averages. With this approach, the dominance of single teams can be modelled over the years without a bias by inflation.
- **Bookmaker odds** averaged from multiple bookmaker companies. Collected from [oddsportal.com](https://www.oddsportal.com) (July 2021) and averaged over six different bookmakers in 2010 up to 12 bookmakers in 2019. The odds can be transformed to probabilities by inverting them to  $p_j = \frac{1}{\text{odds}_j}, j \in \{1, X, 2\}$ . As

these do not sum up to 1 (due to bookmakers' margins<sup>2</sup>), we adjust these by  $\tilde{p}_j = \frac{p_j}{p_1 + p_X + p_2}$  with  $p_1$  and  $p_2$  corresponding to wins of the first or second named team and  $p_X$  to a tie. With this, we implicitly assume an evenly distributed margin across these outcomes. An alternative, more complex normalisation approach, which is optimal against insider trading, was proposed by Shin (1991).

- **Promoted** status of a team. Indicates for each team, whether it has been promoted to the division immediately before the current season. This is used to include the “rookie status”.
- **Titleholder** from last season. Indicates for each team whether it is the league's current titleholder.
- **CupTitleholder** from last season. Indicates for each team whether it is the titleholder of the national cup (DFB-Pokal in Germany, FA CUP in England, Copa del Rey in Spain, Coppa Italia, Coupe de France, KNVB Cup in the Netherlands, Turkish Cup).
- **FormGoals3** is the number of goals scored by the corresponding team  $i$  in its last three matches. Easily calculated for matchdays 4 and later. For earlier matchdays the last seasons average of all teams  $\bar{g}$  is used.

– matchday 1:  $\text{FormGoals3} = \bar{g}$

– matchday 2:  $\text{FormGoals3} = \frac{1}{3}g_{\text{team } i, \text{ matchday1}} + \frac{2}{3}\bar{g}$

– matchday 3:  $\text{FormGoals3} = \frac{1}{3}g_{\text{team } i, \text{ matchday1}} + \frac{1}{3}g_{\text{team } i, \text{ matchday2}} + \frac{1}{3}\bar{g}$

In rare cases, when a result is missing in the last 3 matches, the average of the remaining 2 matches is used. Instead of 3, the last 5 (or 7, 10, ...) matches could be used. We settled on 3 to capture the most recent form of the teams, which in football can often change quite spontaneously.

Note that, of course, principally many more potential covariates could be collected and added to the data, such as e.g. the teams' *average ages* or the coaches' *job tenure*, or even so-called *hybrid* variables that are derived themselves by statistical model as done in Groll et al. (2019) and Groll et al. (2021). However, we

<sup>2</sup>The bookmakers' margins can be seen as the fee the bookmakers take for offering their bets. As a simplified example, fair betting odds for a (fair) coin toss would be 2. The offered odds need to be lower than that, maybe 1.9, so the bookmaker is running profits in the long run. For more details, see also the Betting Results paragraph in Section 3.4

abstain to do so here, as we want to present rather standard approaches that can be applied more or less directly by interested practitioners. For this purpose, we have restricted the set of potential covariates to a selection which we deem to be both highly informative and quite directly available.

### MISSING DATA AND ABNORMALITIES

As noted above, no market values were available before 2010-11-01. This affects 676 matches in total from all included leagues. The website [transfermarkt.com](http://transfermarkt.com) also does not provide data for teams that were dissolved or left professional and semi-professional divisions. This results in missing market values in the following cases:

- **Athlétic Club Arlésien** in the Ligue 1  was dissolved in 2016 and has missing market values in its only season of 2010/11.
- **Thonon Évian F.C.** in the Ligue 1  was relegated multiple times and left professional and semi-professional football, currently switching between France's 5th and 6th division. This leads to missing values in the four seasons of 2011/12, 2012/13, 2013/14, and 2014/15.
- **ACN Siena 1904** in the Serie A  was dissolved in 2014 and has missing market values in the seasons of 2011/12 and 2012/13. Although the team was re-established multiple times, it was never able to reach the higher divisions.
- **AC Cesena** in the Serie A  was dissolved in 2018 and has missing market values in the seasons of 2010/11, 2011/12, and 2014/15.
- **Kayseri Erciyesspor** in the Süper Lig  was dissolved in 2018 and has missing market values in the seasons of 2013/14 and 2014/15.
- **Orduspor** in the Süper Lig  was dissolved in 2019 and has missing market values in the seasons of 2011/12 and 2012/13.
- **Mersin Idman Yurdu** in the Süper Lig  was dissolved in 2019 and has missing market values in the seasons of 2011/12, 2012/13, 2014/15, and 2015/16.
- **Bucaspor** in the Süper Lig  was dissolved in 2020 and has missing market values in its only season of 2010/11.

- Gaziantepspor in the Süper Lig  was dissolved in 2020 and has missing market values in the seven seasons between 2010/11 and 2016/17.








In total, 2236 market values are missing, of which 1352 correspond to matches before 2010-11-01 and 884 to the teams mentioned above (after 2010-11-01).

For the bookmakers' odds a total of 346 entries is missing, belonging to 118 matches. In total 1706 matches include missing data, of which 676 are from the start of the season 2010/11. The other 1030 matches are spread throughout the leagues and seasons. Apart from these missing values of single covariates, due to the COVID-19 pandemic full matchdays were missing or performed under different circumstances.

### THE PANDEMIC

As noted before, we will omit games played during the COVID-19 pandemic. The dates on which each league was influenced is given in Table 2. As the leagues were handling the situation differently, e.g. in Ligue 1 the season was postponed and later cancelled while the Süper Lig had matches behind closed stadium doors and later postponed the season, we exclude all matches later than the given dates, which were those of the earliest decisions regarding each league. As single matches (e.g., matches in the Eredivisie in February) have been postponed due to different reasons and should have taken place later, those matches before that cut-off point are missing. The corresponding final sample sizes per league are found in Table 2 as well.

**Table 2: Start dates of matches under the COVID-19 pandemic influence. Date corresponds to the first decision, not the final one.**

League	decision	date	included matches	with missings
 Premier League	postponed	2020-03-13	3696	128
 Ligue 1	cancelled	2020-03-13	3687	325
 Bundesliga	postponed	2020-03-16	2966	116
 Serie A	postponed	2020-03-09	3668	296
 Primera División	postponed	2020-03-12	3674	119
 Eredivisie	cancelled	2020-03-12	2973	118
 Süper Lig	postponed	2020-03-12	2963	597

**Given the ever changing situation and decisions, we exclude all matches starting from 2020-03-01.** As the remaining missing data points are rather few compared to the full data set, we will not use any methods for data imputation and instead omit matches whenever a variable is used that is missing.

### 3. MODELS AND EVALUATION MEASURES

For all models the general notation includes the number of goals scored per team  $(y_1, y_2)$  and a covariate or design matrix  $\mathbf{X}$ , respectively, containing for each match a set of  $k$  different covariates as a single row  $\mathbf{x}_i = (1, x_1, \dots, x_k)^\top$ . The first column with entries of 1 corresponds to an intercept, which is included depending on the model.

#### 3.1. MODELLING THE NUMBER OF GOALS

All fitting procedures and evaluations were performed within R (R Core Team, 2020).

Most models will be used with two different model equation sets. First, each team's goals are modelled with the team's covariates, indicated by  $H$  and  $A$  for home and away teams, respectively, in the following pseudo model formulae:

$$\begin{aligned} y_H &\sim \text{elo}_H + \text{MV}_H + \tilde{p}_1 + \text{FormGoals3}_H + \text{Promoted}_H + \text{Title}_H + \text{CupTitle}_H, \\ y_A &\sim \text{elo}_A + \text{MV}_A + \tilde{p}_2 + \text{FormGoals3}_A + \text{Promoted}_A + \text{Title}_A + \text{CupTitle}_A. \end{aligned} \quad (1)$$

And for a second, more complex type of approaches, each team's goals are modelled by the covariates of both teams, including information about the opponents strength.

$$\begin{aligned} y_H &\sim \text{elo}_H + \text{elo}_A + \text{MV}_H + \text{MV}_A + \tilde{p}_1 + \tilde{p}_2 + \text{FormGoals3}_H + \text{FormGoals3}_A + \\ &\quad \text{Promoted}_H + \text{Promoted}_A + \text{Title}_H + \text{Title}_A + \text{CupTitle}_H + \text{CupTitle}_A \\ y_A &\sim \text{elo}_A + \text{elo}_H + \text{MV}_A + \text{MV}_H + \tilde{p}_2 + \tilde{p}_1 + \text{FormGoals3}_A + \text{FormGoals3}_H + \\ &\quad \text{Promoted}_A + \text{Promoted}_H + \text{Title}_A + \text{Title}_H + \text{CupTitle}_A + \text{CupTitle}_H \end{aligned} \quad (2)$$

#### POISSON REGRESSION

Poisson regression is typically performed via a generalised linear model (GLM) with an exponential link function and often used to model count data. The two margins are typically treated independently (conditional on the covariate information), so no dependency apart from the covariate level is included. For a general overview of these models, see, e.g., Groll and Schauburger (2019).

## REGULARISED POISSON REGRESSION

To achieve some form of sparsity, penalisation techniques such as the LASSO (Tibshirani, 1996) can be used. In this setting, the fitting procedure is able to shrink coefficients or to set them completely to zero. As is typical for LASSO, the penalty strength (commonly denoted as  $\lambda$ ) is determined via a cross validation approach, which is e.g. implemented in the `cv.glmnet` function from the `glmnet` R package (Friedman et al., 2010). The LASSO penalisation was used in the context of football, e.g. by Groll and Abedieh (2013) and Groll et al. (2015).

## COPULA REGRESSION

Copula regression applies dependency between (in this case) Poisson marginal regressions. The GJRM framework and R implementation by Marra and Radice (2019) is used, which was proposed to the application of football in van der Wurp et al. (2020). Detailed insights into the methodology can be found there and in the references therein. As the authors found the F (Frank) and FGM (Farlie-Gumbel-Morgenstern) copulae to be good choices for the application of FIFA World Cups, we concentrate on these dependency structures.

## REGULARISED COPULA REGRESSION

Moreover, van der Wurp et al. (2020) proposed a penalty to ensure equal coefficient estimates for the same covariates of both competing teams. Corresponding covariates in this case are e.g.  $\text{elo}_H$  and  $\text{elo}_A$  in Equations (1) or (2). The way a team's elo rating is influencing the goals scored by the team should be the same regardless of whether the team is first- or second-named, i.e. home or away team. It is important to note that for the models from Equation (2),  $\text{elo}_H$  in the first margin and  $\text{elo}_A$  in the second margin are **not** coinciding, but yielding the same interpretation in different margins. To clarify, they are not the same covariate, but are treated as identical in the penalisation scheme. The covariates' order in Equation (2) highlights this. However, it can be argued that their coefficients should coincide.

A second LASSO-type penalty proposed by van der Wurp and Groll (2021) introduces sparsity to the framework. We will use the two penalties both individually and combined to find the best approach. A fixed grid length of 100 is used for optimising the LASSO-penalty strength. Note that varying the construction of the grid (density or location) would yield slightly different results.

## RANDOM FORESTS

Multiple implementations of random forests exist in R. Groll et al. (2019) found the `cforest` from the `party` package by Hothorn et al. (2006) to be the best for the application of FIFA World Cups. Also, in the UEFA European Championship 2020 the `cforest` again yielded very promising results (Groll et al., 2021). We will follow these findings and use this implementation as a representative for random forests. For the general methodology about random forests see Breiman (2001), and Breiman et al. (1984) for the idea of classification and regression trees (CARTs) behind random forests.

## EXTREME GRADIENT BOOSTING

Instead of parallel ensemble methods like the random forest approach from above, one can also consider sequential ensembles such as *boosting*, a technique which stems from the machine learning community (Freund and Schapire, 1996) and was later adapted to estimate predictors for statistical models (Friedman, 2001; Friedman et al., 2000). Friedman (2001) introduced the idea of gradient tree boosting, with decision trees as learners. These are repeatedly fitted on the residuals of the previous fitting step and, hence, combined to a sequential ensemble. This technique was then further improved by Chen and Guestrin (2016) via introducing additional regularisation in the objective function. The regularisation terms make the single trees weak learners to avoid overfitting. In a certain boosting iteration, the next tree is additively incorporated into the ensemble after multiplication with a rather small learning rate, which makes the learners even weaker. The method is called *extreme gradient boosting* (XGBoost), and is known in the machine learning community for its high predictive power<sup>3</sup> The R package `xgboost` by Chen et al. (2021) contains the implementation of the algorithm.

For a brief summary of the methodology and an exemplary application to football, see e.g. Groll et al. (2021). Finally, note that an important aspect is that XGBoost involves several tuning parameters, such as e.g. the learning rate, the optimal number of boosting steps and several penalty parameters. For this purpose, we specified simple, discrete parameter grids and used multivariate 10-fold cross validation to determine optimal values for three key tuning parameters (namely the learning rate `eta`, the convergence criterion for splits `gamma`, and the max number of boosting iterations `nrounds`) on the training data (prior to

<sup>3</sup>It lately has been very successful in several prestigious machine learning prediction competitions, such as those launched by Kaggle (<https://www.kaggle.com>).

2014/2015). This is performed for each league individually and on the full training data set. The tuned parameters are kept constant after this.

### 3.2. MODELLING THE ORDINAL/NOMINAL OUTCOME

Beside modelling the number of goals per match  $(y_1, y_2)$  one can also model the three-way outcomes directly, which could be seen as a natural alternative as we are using multiple quality-of-prediction measures on this dimension and betting on these outcomes is rather popular. Hence, we will also model the match results `winHome` (with  $y_1 > y_2$ ), `draw` (with  $y_1 = y_2$ ) and `winGuest` (with  $y_1 < y_2$ ) and from now on will use the common short notation of bookmakers, i.e.  $1/X/2$ , for these three outcomes. As a draw is clearly positioned between the other two outcomes, ordinal approaches are deemed more suitable than nominal ones, as they can exploit this information. We use the `polr` function of the `MASS` package by Venables and Ripley (2002), which fits a cumulative proportional-odds logit model.

#### REMARKS

Model approaches from Equations (1) and (2) are used in comparison whenever possible. This includes (regularised) Poisson regression, all copula models, random forests and the XGBoost. The ordinal approach is modelling the one-dimensional outcome  $1/X/2$ , where all covariates from (one of the two parts from) Equation (2) are used.

For all models predicting independently both scores, the Skellam distribution as a difference between two Poisson distributed variables is used to calculate probabilities for the three-way outcomes. This affects Poisson regression, random forests and XGBoost.

### 3.3. PREDICTION APPROACH

To simulate a realistic prediction situation, we use all prior matches of a given league to predict the following matchday. For this, we declared the first 5 seasons from (2010 up to 2015) as “burn-in” training data. So, starting from the season of 2015/2016, this training data is used to predict the next matchday. Afterwards, the predicted matchday is added to the training data, continuing throughout all remaining matchdays and seasons.

For the global model, which does not differentiate between the leagues, we use the date instead of the matchday, as the latter is not consecutive anymore. Although this leads to smaller steps (dates vs. matchdays) and slightly changing



sizes of the test data in our prediction approach, we deem the differences to the league-specific approach to be negligible

The quality or goodness of the obtained predictions is observed on multiple levels and calculated with measures from the following Section 3.4.

### 3.4. GOODNESS OF PREDICTION MEASURES

This section will introduce measures of prediction quality. With these, we cover all interesting response levels, i.e. *goals*, *three-way outcomes*, and *betting results*. It should be noted that not all measures are applicable to all models. The ordinal model for example does not provide estimated goals, so no error measures on this level can be obtained.

#### RPS

The ranked probability score (RPS) observes the three-way outcomes. It takes the ordinal structure of *win*, *draw* and *loss* into account and is defined in this context as

$$\text{RPS}_i = \frac{1}{2} \sum_{r=1}^2 \left( \sum_{l=1}^r \hat{\pi}_{il} - \delta_{il} \right)^2$$

for each match  $i$  (see, e.g. Schauburger and Groll, 2018, for another application, and Gneiting and Raftery, 2007, for the original proposition). Here,  $\hat{\pi}_{il}$  are the estimated probabilities for the respective three-way outcomes  $l$  and  $\delta_{il}$  is the Kronecker's delta, containing the observed outcome. In general, the RPS is an error term on the probability-level and is to be minimised. Alternatively, the (multi-category extension of the) Brier score (Brier, 1950) could be used on the three-way outcomes. But as it does not account for the ordinal structure, we use the RPS instead.

#### MULTINOMIAL LIKELIHOOD

The multinomial likelihood (LH), which also operates on the probability-level, is defined as

$$\text{LH}_i = \hat{\pi}_{i1}^{\delta_{i1}} \hat{\pi}_{i2}^{\delta_{i2}} \hat{\pi}_{i3}^{\delta_{i3}},$$

which is essentially the predicted probability for the observed outcome (van der Wurp et al., 2020), and therefore is to be maximised.

## CLASSIFICATION RATE

The classification rate (CR) is maybe the simplest measure. Out of the three-way outcome, we classify the outcome with the highest predicted probability as the estimated outcome. For a single game  $i$ , this can be written via

$$CR_i = \mathbb{1} \left( \delta_i = \arg \max_{l \in \{1,2,3\}} (\hat{\pi}_{il}) \right).$$

The global classification rate is then averaged over all matches and is to be maximised.

## ERRORS IN GOALS

On the response-level of the goals scored, one can easily calculate the difference between the number of estimated and observed goals per team. For each match, we calculate the squared and absolute errors via

$$\begin{aligned} SE_i &= (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2, \\ AE_i &= |\hat{y}_1 - y_1| + |\hat{y}_2 - y_2|. \end{aligned}$$

## BETTING RESULTS

Last, as maybe the most popular benchmark measure, we will investigate each model's performance in regard to betting. For this, we use the bookmakers' odds from [oddsportal.com](http://oddsportal.com). It is important to note that these odds are averaged over a selection of bookmakers, so the results are not necessarily the same using a single or even a selection of bookmakers.

To create a betting strategy, we calculate the expected return of a given bet via

$$E[\text{return}_i] = \hat{\pi}_{il} \cdot \text{odds}_{il} - 1. \quad (3)$$

As soon as the expected return is positive, one should take that bet (a threshold value of 0 marks a fair bet). Larger thresholds than zero may be chosen.

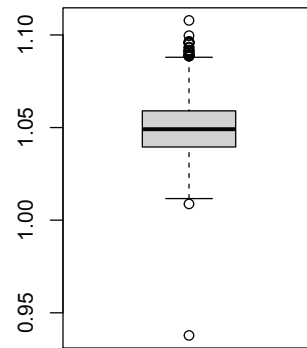
If multiple bets for a single match yield a positive expected return, we will simply take the one with the highest expected return, limiting us to a single bet per match. Other approaches, such as a variance-minimising strategy, taking the bet with (a positive expected return and) the highest probability of success are also possible.

We are using a stake of 1 fiscal unit for each bet, indicated by the  $-1$  in the expected return (3). Other strategies are possible as well, e.g. the Kelly criterion (Kelly, 1956), which gives weights and therefore different stake sizes to each bet. The outcome in terms of gains is then calculated via

$$\text{gains}_i = \begin{cases} -1, & \text{if bet failed} \\ \text{odds}_{il} - 1, & \text{if bet was successful} \end{cases}$$

and summed up over all matches of a given league. Making a profit (i.e. beating the bookmakers) is a very optimistic and challenging objective. Hence, achieving betting losses close to zero with rather simple models is already considered an achievement, especially considering bookmakers' costs and (presumably) taxes.

When transforming bookmakers odds to probabilities (see Section 2), the probabilities do not sum up to 1 because of margins. As bookmakers are offering smaller odds than a fair bet would be, the transformation yields higher probabilities. These sums average to 1.05. The downward outlier (see Figure 1) may be the result of the averaging process from [oddsportal.com](http://oddsportal.com) and is not further investigated. The distribution indicates the 5% winning margin (median) the bookmakers are collecting.










**Figure 1: Values of inverted and summed up odds. For fair bets, this would always sum up to exactly 1. The difference can be interpreted as bookmaker margin.**

#### 4. RESULTS

For all models and leagues, the resulting measures are averaged throughout all predicted matches. Exemplarily, the results for the simple independent Poisson model from Equation (1) are shown in Table 3. The quality of prediction differs between the national leagues. This is especially visible in the betting results, ranging from a profit of 38.41 stakes (fictional money unit) in the Premier League to a loss of 189.51 stakes in the Süper Lig. Relative to the “invested” stakes this corresponds to a winning rate (i.e.  $\text{bet.gains}/\text{bets}$ ) of 2.35% and a loss of 20.18%,

respectively. The ratio of matches that are bet on (i.e. bets/ $n$ ) also varies strongly and ranges from 92.36% for the Premier League to 70.50% for the Süper Lig. This should be taken with a grain of salt, as the leagues receive widely different attention in the national and international media and betting markets.

**Table 3: Results for the simple independent Poisson model from Equation (1).**

	RPS	LH	CR	SE	AE	$n$	bets	bet.gains
 Premier League	0.191	0.434	0.552	2.601	1.804	1768	1633	38.41
 Bundesliga	0.203	0.418	0.520	2.972	1.921	1414	1189	-117.80
 Primera División	0.191	0.435	0.537	2.583	1.770	1746	1353	-113.76
 Ligue 1	0.199	0.411	0.514	2.523	1.746	1784	1358	-76.70
 Serie A	0.185	0.442	0.577	2.534	1.755	1744	1539	-89.17
 Eredivisie	0.189	0.446	0.583	2.961	1.912	1442	1061	-83.06
 Süper Lig	0.200	0.405	0.527	2.653	1.808	1332	939	-189.51

First, to be able to compare our big selection of models, we average the measures throughout all leagues. We are using a weighted average by sample sizes for the measures of RPS, LH, CR, SE, and AE and a simple sum for the number of matches  $n$ , the number of bets and the bet gains. The results for all models can be found in Table 4. Goodness-of-prediction results, exemplarily in terms of RPS and betting returns, for each league can be found in the appendix, Tables 9 and 10.

The RPS is, ever so slightly, improving with the copula models becoming more complex. Both the equal and the LASSO penalty are improving the results. Regarding the average multinomial likelihood the BIC models with lasso penalisation are performing worse than their AIC counterparts. We found no noteworthy differences between models using both marginal covariates in both marginal regressions and their simpler counterparts (see Equation (2) in Section 3.1 compared to Equation (1)). The classification rate CR has little to no variation in any direction. Sadly, no model was able to end with a net gain in betting from thousands of matches and bets. But some models are getting close to break-even. The simple copula models with all available covariates are achieving losses of less than 2.5% of stakes from more than 8600 bets. As discussed and shown above in Section 3.4, the calculated margin of bookmakers can be assumed to be about 5%, as they have expenses to cover. A selection of our models is solidly beating that threshold and might create frowning reactions with bookmaker companies. The equalisation penalty from van der Wurp et al. (2020) is impairing the models with and without LASSO penalisation. The gain in interpretability (see van der Wurp and Groll, 2021 and the aforementioned reference from 2020) comes at a cost of

**Table 4: Results for all modelling approaches. Calculated separated by leagues, but then aggregated. Cell colors best (green) to worst (red) for visualisation. See digital version.**

Model	Eq	Cop.	regul.	RPS	LH	CR	SE	AE	bets	gainratio
pois	1	-	-	0.1938	0.428	0.545	2.674	1.811	9072	-0.0696
pois	2	-	-	0.1941	0.428	0.542	2.683	1.812	9506	-0.0430
pois	1	-	LASSO	0.1938	0.425	0.544	2.672	1.808	8915	-0.0693
pois	2	-	LASSO	0.1939	0.425	0.543	2.676	1.809	9352	-0.0524
RF	1	-	-	0.1975	0.427	0.536	2.753	1.840	10365	-0.0718
RF	2	-	-	0.1961	0.427	0.539	2.721	1.827	10188	-0.0617
XGboost	1	-	-	0.1967	0.412	0.543	2.732	1.818	10188	-0.0765
XGboost	2	-	-	0.1970	0.412	0.542	2.733	1.819	10312	-0.0609
Cop	1	F	-	0.1937	0.429	0.544	2.676	1.812	7874	-0.0549
Cop	1	FGM	-	0.1937	0.429	0.544	2.676	1.812	7936	-0.0573
Cop	2	F	-	0.1940	0.429	0.542	2.683	1.812	8696	-0.0250
Cop	2	FGM	-	0.1940	0.429	0.542	2.683	1.812	8753	-0.0243
Cop	1	F	equal	0.1937	0.429	0.543	2.674	1.808	7200	-0.0898
Cop	1	FGM	equal	0.1938	0.429	0.543	2.674	1.808	7276	-0.0897
Cop	2	F	equal	0.1938	0.429	0.542	2.675	1.810	8109	-0.0517
Cop	2	FGM	equal	0.1938	0.429	0.542	2.674	1.810	8164	-0.0497
Cop AIC	1	F	LASSO	0.1937	0.428	0.544	2.676	1.810	7578	-0.0670
Cop BIC	1	F	LASSO	0.1939	0.425	0.544	2.681	1.809	8016	-0.0800
Cop AIC	1	FGM	LASSO	0.1937	0.428	0.543	2.676	1.810	7669	-0.0733
Cop BIC	1	FGM	LASSO	0.1940	0.425	0.544	2.682	1.810	8035	-0.0921
Cop AIC	2	F	LASSO	0.1939	0.428	0.543	2.684	1.811	8150	-0.0363
Cop BIC	2	F	LASSO	0.1944	0.423	0.544	2.694	1.814	8315	-0.0715
Cop AIC	2	FGM	LASSO	0.1939	0.428	0.542	2.684	1.812	8259	-0.0408
Cop BIC	2	FGM	LASSO	0.1943	0.423	0.544	2.693	1.813	8456	-0.0717
Cop AIC	1	F	both	0.1937	0.429	0.543	2.679	1.808	6028	-0.0858
Cop BIC	1	F	both	0.1936	0.429	0.543	2.679	1.808	5729	-0.0807
Cop AIC	1	FGM	both	0.1935	0.429	0.543	2.670	1.806	5864	-0.0901
Cop BIC	1	FGM	both	0.1935	0.429	0.543	2.669	1.806	5560	-0.0960
Cop AIC	2	F	both	0.1938	0.428	0.543	2.673	1.808	6505	-0.1004
Cop BIC	2	F	both	0.1938	0.428	0.543	2.675	1.808	5816	-0.0963
Cop AIC	2	FGM	both	0.1936	0.429	0.543	2.670	1.807	6257	-0.0899
Cop BIC	2	FGM	both	0.1935	0.428	0.543	2.673	1.807	5729	-0.1037
ordinal	-	-	-	0.1944	0.430	0.542	-	-	9419	-0.0433

prediction quality.

It should be noted that the mentioned measures are operating on the three-way-outcome dimension, while most model fitting procedures are using the likelihood on the number of goals. So errors on goals (SE and AE) might be a fairer measurement with regard to the models' original purpose apart from football modelling. With the exception of BIC models being constantly worse than their AIC counterparts, more sophisticated models in terms of penalisation are achieving better prediction performances. The combined models with equalisa-

tion and LASSO penalties are yielding the best results, albeit quite close to the LASSO-penalised Poisson model.

To summarise, it is not possible to declare a clear winning model. Depending on the context and the user's aims and scope, we deem multiple models to be suitable. For pure interpretability very simplistic models such as the ordinal or the simple Poisson model might be favoured. The equalisation approach allows for a better insight into coefficients, as they are cleaned of home- and away-team-specific differences in covariate effects. The best model – if the objective is to beat bookmakers – is, in this case, neither the most complex nor the simplest approach. In the following, we will present selected models in detail to highlight certain advantages and disadvantages.

The results by league are rather interesting, see Tables 9 and 10 in the appendix. Regarding the RPS our predictions for the French Ligue 1 and the Turkish Süper Lig are considerably worse than for the other leagues. The fictional betting returns show a similar pattern for the Süper Lig - matches in this league seem to be harder to predict than those of other leagues. Especially for the English Premier League and the German Bundesliga the models seem to perform quite well. As the investigated leagues receive quite different amounts of international attention, some difference in data quality can be assumed particularly for bookmaker odds and market values, the latter variable originating from a community project.

### SELECTED MODELS IN DETAIL

We begin with examining the clear winner model regarding the betting outcome, which is the copula model with all available covariates and no penalisation whatsoever. As the differences between FGM and F copula are negligible we will show examples from both. Some resulting coefficients, exemplarily for the Premier League, can be found in Table 5.

These coefficients (and especially the differences between the two margins) are rather hard to interpret. While each respective team's market value has a positive influence on the team itself, the opponent's market value is behaving quite differently. For home teams, the market value of their opponents has a positive impact and for away teams, the respective market value of their opponents has a negative influence. Due to high levels of multicollinearity, think for example of elo, market value and bookmaker probabilities  $p$ , the exact values cannot be taken at face value. But rather big differences between the first and second margin are still hard to justify.

Models with higher value regarding interpretability may be desired, even if

**Table 5: Estimated coefficients for the copula FGM model with all covariates and no penalisation, exemplarily for the Premier League**

	$\beta^{(1)}$	SE( $\beta^{(1)}$ )	$\beta^{(2)}$	SE( $\beta^{(2)}$ )
(Intercept)	-0.9255	0.6616	-0.1517	0.7508
elo Team	-0.0001	0.0004	-0.0006	0.0004
elo Opponent	0.0004	0.0003	0.0007	0.0004
MV Team	0.3625	0.9023	1.2796	1.0520
MV Opponent	1.6485	1.0235	-2.0060	1.1524
$p$ Team	1.6295	0.3602	1.4432	0.3845
$p$ Opponent	-0.3531	0.3756	-0.4858	0.4138
FormGoals3 Team	-0.0205	0.0204	-0.0007	0.0230
FormGoals3 Opponent	-0.0081	0.0217	-0.0227	0.0249
Promoted Team	0.0443	0.0464	-0.1448	0.0549
Promoted Opponent	0.0516	0.0397	-0.0139	0.0453
Title Team	-0.0853	0.0598	-0.0347	0.0668
Title Opponent	-0.0477	0.0780	0.1858	0.0865
Title Cup Team	-0.0303	0.0602	-0.0657	0.0686
Title Cup Opponent	0.0185	0.0792	-0.0470	0.0969

they offer a slightly worse performance in specific measures or even overall. The results for the separate leagues (see Table 8 in the appendix) are varying strongly between the leagues and each league's margins. This could be for two reasons: A) The covariates' influence is immensely different in each league and the leagues should therefore be fitted independently. We will discuss this in Section 4.1 in more detail. Or B) a lot of noise and artefacts are included in the models. Therefore, some form of sparsity should be incorporated. We will take a closer look at other well-performing models from Table 4, i.e. applying the LASSO-type penalty and a second model using both the LASSO and the presented equalisation penalty.

### SPARSER MODELS

The (LASSO-) penalised model via BIC with an F copula from Table 4 has a slightly worse performance regarding betting and no noteworthy changes in the other measures. The resulting coefficients can be found in Table 6.

With eight coefficients shrunk to zero, the model is slightly sparser and easier to interpret, while maintaining virtually the same quality of prediction. Some oddness remains: Playing against the current titleholder has a positive impact on

**Table 6: Estimated coefficients for the LASSO-penalised copula F model (left) and with both penalties combined (right) with all covariates, exemplarily for the Premier League. For both models the optimal tuning parameters were selected via BIC.**

	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(1)}$	$\beta^{(2)}$
(Intercept)	-0.6960	-0.2231	-0.3593	-0.3637
elo Team	0.0005	0.0002		
elo Opponent	0.0001	-0.0002		
MV Team	0.0488	0.3249		
MV Opponent	0.6337	-0.7649		
$p$ Team	0.8067	1.2049	1.6492	1.6498
$p$ Opponent	-0.7372	-0.1131		
FormGoals3 Team		0.0019		
FormGoals3 Opponent		-0.0084		
Promoted Team	0.0190	-0.0535		
Promoted Opponent	0.0153			
Title Team	-0.0702			
Title Opponent		0.0389		
Title Cup Team				
Title Cup Opponent		-0.0081		

the away team, but no influence at all on the home team. The opponent's market value even changes its sign completely if a team is playing at home or away. This can be rationalised with strong interdependencies and collinearities or with missed features such as psychological factors and others.

To (partly) tackle this issue, we will take a look at the model with the combined penalties (BIC tuning and F copula again) in Table 4. The results can also be found in Table 6. With only five coefficients estimated different from zero (including the copula parameter, which, interestingly, was estimated to be virtually zero), the resulting model is extremely sparse and easy to interpret. Here,  $\beta^{(1)}$  and  $\beta^{(2)}$  are virtually equal, allowing straightforward interpretations. The predicted probabilities by bookmakers  $p$  – which can be interpreted as a substitute variable for team strength – are estimated yielding a positive influence on each respective team. Note here that the intercept was only penalised by the equalisation penalty and not by the LASSO-type approach, as is common for the LASSO framework.



#### 4.1. DIFFERENCES BETWEEN LEAGUES

In this section, we investigate whether the leagues are different regarding their assumed underlying model. Instead of comparing or testing the models' coefficients, we compare the quality of prediction in the ever updating models when differentiating between the national leagues and when treating them as one global training data set. Instead of predicting the next matchday (as done before), we are using the dates of matches. This results in 1793 unique dates of which the first 913 are solely used as training data and the other 880 are predicted using all matches before the given date. The results in comparison to Table 4 from before are shown in Table 11. Unsurprisingly, the results are not wildly different. Instead, the results seem to be more homogeneous than before. Especially, the betting results are clearly more consistent between models.

The estimated coefficients for a selected copula regression model can be found in Table 7. As interpretability is limited with wildly different marginal coefficients, the equalisation penalty is applied again and the resulting coefficients are compared. The resulting model contains four covariates for each margin. The bookmakers'  $p$  was consistently chosen in both settings. Interestingly, the estimated copula parameter  $\theta$  was again estimated to be virtually zero in terms of Kendall's  $\tau$  (0.0297 and  $< 0.0001$  in absolute value, respectively for the models from Table 7), indicating no correlation structure whatsoever.

#### 5. CONCLUSIONS

In this work, we presented an extensive data set of football matches in European leagues and the application of different modelling approaches to it. Comparing methodologies, we found regularised copula regression approaches to yield good results. The very flexible machine learning techniques of Random Forests and XGBoost are very sensible to tuning - their rather mediocre results in this application can almost certainly be improved via extensive tuning. The (copula) regression approaches yield models that are both easy to interpret and to use. However, the gain compared to simple approaches such as standard independent Poisson modelling is rather small.

We found a set of covariates that are more important than others. Unsurprisingly, especially the bookmakers' probabilities (converted from odds) are deemed to be full of information and can be a solid predictor on their own. Differences between the investigated seven European leagues were found considering relevant covariates. The common ground was found to be the previously mentioned bookmakers' odds. The influence of other coefficients varies greatly in different coun-

**Table 7: Estimated coefficients for the LASSO-penalised copula F model (left) and with both penalties combined (right) with all covariates for all leagues combined in comparison to Table 6. For both models the optimal tuning parameters were selected via BIC.**

	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(1)}$	$\beta^{(2)}$
(Intercept)	-0.5769	-0.1050	-0.1143	-0.1104
elo Team	-0.0005	-0.0003	-0.0002	-0.0002
elo Opponent	0.0003	0.0000		
MV Team	0.7206	0.4752	0.7013	0.6972
MV Opponent	-0.3280			
$p$ Team	2.2575	1.9893	1.7061	1.7082
$p$ Opponent	0.6245	0.1176		
FormGoals3 Team	0.0204	0.0094		
FormGoals3 Opponent	0.0339	0.0060	0.0178	0.0169
Promoted Team	-0.0130	-0.0579		
Promoted Opponent	-0.0221	-0.0168		
Title Team		-0.0385		
Title Opponent	-0.0437	-0.1077		
Title Cup Team	0.0558	-0.0123		
Title Cup Opponent	0.0285	-0.0165		

tries in both strength and sign. As these can be interpreted as correction factors onto the immense importance of bookmakers' odds, the variation can be caused by the leagues themselves or different prediction strategies by the bookmakers.

Principally, one reason for all regarded modelling approaches yielding rather similar results could be that they all base on the highly informative bookmakers' odds, as described above. Hence, the specific type of modelling (linear vs. non-linear, interactions, dependence structure, etc.) here seems to play a minor role. We believe that extending the regarded set of covariates by additional features which cover new types of information, such as e.g. the "hybrid" features regarded in Groll et al. (2021, 2019) for the modelling of national team tournaments could on the one hand side increase the overall predictive performance of the models, on the other hand manifest more distinctive results across model classes. Unfortunately, the calculation of these features is rather extensive and went beyond the scope of this work. Besides, as mentioned above, the machine learning approaches are subject to complex tuning. Hence, they typically need a large training data set to utilize their full potential.

The data indicate that bookmakers are calculating with a betting margin of about 5%. While some models were able to beat this margin, we can not claim to have beaten the bookmakers, as other models ran significant losses. There are obvious limitations due to the available data. Our data set was completely compiled from publicly available sources and from a fixed point in time. Bookmakers are able to shift existing odds depending on betting behaviour of customers or depending on external events, such as a core player getting injured before a match. A public-data driven approach such as this cannot be that flexible.

While this work is focussed around national leagues, all models can principally be applied to different tournaments as well, such as FIFA World Cups, UEFA European Championships, or the UEFA Champions League and comparable tournaments on the club-level on other continents. However, some additional aspects need to be considered. For one the existing sample sizes are considerably smaller, causing issues for complex machine learning approaches. Also each tournament's specific structure (how groups are built in group stages, tournament schedule, potential extra time and penalty shoot-outs etc.) needs to be taken into account. See, for example, thoughts by Egidi and Torelli (2021), van der Wurp et al. (2020), and van der Wurp and Groll (2021).

All in all, our aim was mainly to create an interesting data set and motivate different statistical and machine learning modelling approaches to it, rather than finding the actual/virtual/definite best prediction approach on the regarded data, e.g. in terms of betting profits. The manuscript shall give an overview of their general predictive potential in this field of application as well as other aspects such as interpretability, which might also be relevant for the practitioner. The underlying data set is publicly available in an R package `EUfootball` (van der Wurp, 2022). The reader is both invited to create their own modelling ideas for the underlying football data and to apply the here presented approaches to other fields and applications. Also, we hope that this work inspires other researches to use and extend our data set, and to build upon and further improve the modelling strategies presented here.

## References

- Baio, G. and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. In *Journal of Applied Statistics*, 37 (2): 253–264.
- Breiman, L. (2001). Random forests. In *Machine Learning*, 45: 5–32.

- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, J.C. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, CA.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. In *Monthly Weather Review*, 78: 1–3.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y. and Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. URL <https://CRAN.R-project.org/package=xgboost>. R package version 1.3.2.1.
- Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46 (2): 265–280.
- Egidi, L. and Torelli, N. (2021). Comparing goal-based and result-based approaches in modelling football outcomes. In *Social Indicators Research*, 156 (2): 801–813.
- Elo, A.E. (1961). New Uscf rating system. In *Chess Life*, 16: 160–161.
- Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156. Morgan Kaufmann, San Francisco, CA.
- Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. In *Annals of Statistics*, 29: 337–407.
- Friedman, J.H., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. In *Annals of Statistics*, 28: 337–407.
- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software*, 33 (1): 1.
- Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. In *Journal of the American Statistical Association*, 102 (477): 359–378.









- Groll, A. and Abedieh, J. (2013). Spain retains its title and sets a new record - Generalized linear mixed models on European football championships. In *Journal of Quantitative Analysis in Sports*, 9 (1): 51–66.
- Groll, A., Hvattum, L.M., Ley, C., Popp, F., Schauburger, G., Van Eetvelde, H. and Zeileis, A. (2021). Hybrid machine learning forecasts for the Uefa Euro 2020. In *arXiv preprint arXiv:2106.05799*.
- Groll, A., Ley, C., Schauburger, G. and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. In *Journal of Quantitative Analysis in Sports*, 15: 271–287.
- Groll, A. and Schauburger, G. (2019). Prediction of soccer matches. In *Wiley StatsRef: Statistics Reference Online*, 1–7.
- Groll, A., Schauburger, G. and Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. In *Journal of Quantitative Analysis in Sports*, 11 (2): 97–115.
- Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M.J. (2006). Survival ensembles. In *Biostatistics*, 7: 355–373.
- Hvattum, L.M. (2017). Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. In *International Journal of Computer Science in Sport*, 16 (1): 50–64.
- Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. In *The Statistician*, 52: 381–393.
- Kelly, J.L. (1956). A new interpretation of information rate. In *Bell System Technical Journal*, 35 (4): 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x. URL <http://dx.doi.org/10.1002/j.1538-7305.1956.tb03809.x>.
- Lee, A.J. (1997). Modeling scores in the Premier League: Is Manchester United really the best? In *Chance*, 10: 15–19.
- Leitner, C., Zeileis, A. and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. In *International Journal of Forecasting*, 26 (3): 471–481.

- Maher, M.J. (1982). Modelling association football scores. In *Statistica Neerlandica*, 36: 109–118.
- Marra, G. and Radice, R. (2019). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2.
- McHale, I. and Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. In *Statistica Neerlandica*, 61 (4): 432–445. doi:10.1111/j.1467-9574.2007.00368.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2007.00368.x>.
- Nikoloulopoulos, A.K. and Karlis, D. (2010). Regression in a copula model for bivariate count data. In *Journal of Applied Statistics*, 37: 1555–1568.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. In *Statistical Modelling*, 18 (5–6): 1–23.
- Schauberger, G., Groll, A. and Tutz, G. (2017). Analysis of the importance of on-field covariates in the German Bundesliga. In *Journal of Applied Statistics*, 45 (9): 1561–1578.
- Schiefler, L. (2015). *Football Club Elo Ratings*. <http://clubelo.com/> [Accessed: July 2021].
- Shin, H.S. (1991). Optimal betting odds against insider traders. In *The Economic Journal*, 101 (408): 1179–1185.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. In *Journal of the Royal Statistical Society*, B 58: 267–288.
- Tutz, G. and Schauburger, G. (2014). Extended ordered paired comparison models with application to football data from German Bundesliga. In *Advances in Statistical Analysis*, 99 (2): 209–227. doi:10.1007/s10182-014-0237-1. URL <http://dx.doi.org/10.1007/s10182-014-0237-1>.
- van der Wurp, H. (2022). *EUfootball: Football Match Data of European Leagues*. URL <https://CRAN.R-project.org/package=EUfootball>. R package version 0.0.1.







- van der Wurp, H. and Groll, A. (2021). Introducing Lasso-type penalisation to generalised joint regression modelling for count data. In *Advances in Statistical Analysis*. URL <https://doi.org/10.1007/s10182-021-00425-5>.
- van der Wurp, H., Groll, A., Kneib, T., Marra, G. and Radice, R. (2020). Generalised joint regression for count data: A penalty extension for competitive settings. In *Statistics and Computing*, 30 (5): 1419–1432.
- Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edn.

## APPENDIX

**Table 8: Estimated coefficient for the copula FGM model with all covariates and no penalisation for all leagues; left columns: home team; right columns: away team**








								
(Intercept)	-0.99	0.25	-0.48	-0.49	-1.73	-1.24	-0.15	-0.18
elo Team	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
elo Opponent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00
MV Team	0.43	1.52	0.81	-0.18	-0.26	-0.39	0.63	1.10
MV Opponent	1.70	-1.73	-1.07	-0.38	-1.56	0.87	0.19	-0.46
$\rho$ Team	1.55	1.49	1.87	2.27	1.98	2.12	2.67	1.65
$\rho$ Opponent	-0.38	-0.42	0.59	0.27	-0.24	-0.08	0.79	0.07
FormGoals3 Team	-0.02	0.01	0.03	0.01	0.01	0.00	0.03	-0.01
FormGoals3 Opp.	-0.01	-0.02	0.02	-0.03	0.04	0.02	0.04	-0.00
Promoted Team	0.04	-0.17	-0.06	-0.11	0.05	-0.01	-0.01	0.04
Promoted Opp.	0.06	-0.03	-0.06	-0.00	0.06	0.11	-0.04	-0.01
Title Team	-0.07	-0.03	-0.06	0.01	-0.01	-0.12	0.05	-0.07
Title Opponent	-0.05	0.17	-0.25	-0.16	-0.18	-0.53	-0.02	0.10
Title Cup Team	-0.01	-0.06	0.10	0.03	0.12	-0.05	0.05	0.13
Title Cup Opponent	0.01	-0.07	0.18	-0.04	0.12	0.09	0.08	0.07

---








						
(Intercept)	-1.57	-1.12	-1.37	1.31	-0.59	-0.99
elo Team	-0.00	0.00	-0.00	-0.00	-0.00	-0.00
elo Opponent	0.00	-0.00	0.00	-0.00	0.00	0.00
MV Team	-0.38	0.48	0.50	0.81	1.25	-0.33
MV Opponent	-0.84	-0.19	-1.57	0.33	-1.17	0.35
$\rho$ Team	2.99	2.60	2.28	0.22	2.45	3.37
$\rho$ Opponent	1.43	0.89	0.97	-1.40	1.11	1.33
FormGoals3 Team	0.03	0.02	-0.01	-0.00	0.03	0.00
FormGoals3 Opp.	0.04	-0.04	0.05	0.03	-0.01	0.02
Promoted Team	-0.01	-0.01	-0.03	-0.17	-0.01	-0.02
Promoted Opp.	-0.00	-0.02	0.01	-0.04	-0.09	-0.06
Title Team	0.11	-0.05	-0.10	0.06	0.08	-0.01
Title Opponent	0.15	-0.12	-0.06	-0.27	0.07	-0.20
Title Cup Team	0.09	-0.12	0.07	-0.04	-0.02	-0.01
Title Cup Opponent	-0.16	0.08	0.02	-0.14	0.05	-0.06



**Table 9: RPS (ranked probability score) results for all models and leagues. Cell colors best (green) to worst (red) for visualisation. See digital version.**

Model	Eq	Cop	regul.							
pois	1	-	-	0.191	0.191	0.185	0.191	0.199	0.189	0.200
pois	2	-	-	0.192	0.192	0.185	0.191	0.199	0.189	0.202
pois	1	-	LASSO	0.192	0.192	0.186	0.191	0.199	0.189	0.201
pois	2	-	LASSO	0.192	0.192	0.185	0.191	0.199	0.189	0.201
RF	1	-	-	0.195	0.195	0.189	0.194	0.202	0.192	0.207
RF	2	-	-	0.194	0.194	0.188	0.192	0.202	0.191	0.203
XGboost	1	-	-	0.196	0.196	0.189	0.194	0.202	0.191	0.203
XGboost	2	-	-	0.196	0.196	0.190	0.194	0.202	0.190	0.204
Cop	1	F	-	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	1	FGM	-	0.191	0.191	0.185	0.191	0.199	0.189	0.201
Cop	2	F	-	0.192	0.192	0.185	0.191	0.199	0.190	0.202
Cop	2	FGM	-	0.192	0.192	0.185	0.191	0.199	0.190	0.202
Cop	1	F	equal	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	1	FGM	equal	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	2	F	equal	0.192	0.192	0.184	0.191	0.199	0.189	0.201
Cop	2	FGM	equal	0.192	0.192	0.184	0.191	0.199	0.189	0.201
Cop AIC	1	F	LASSO	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	F	LASSO	0.191	0.191	0.185	0.192	0.199	0.190	0.200
Cop AIC	1	FGM	LASSO	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	FGM	LASSO	0.192	0.192	0.185	0.192	0.199	0.190	0.200
Cop AIC	2	F	LASSO	0.191	0.191	0.185	0.191	0.199	0.190	0.201
Cop BIC	2	F	LASSO	0.192	0.192	0.187	0.192	0.199	0.190	0.201
Cop AIC	2	FGM	LASSO	0.192	0.192	0.185	0.191	0.199	0.190	0.201
Cop BIC	2	FGM	LASSO	0.192	0.192	0.187	0.192	0.199	0.190	0.201
Cop AIC	1	F	both	0.191	0.191	0.185	0.192	0.199	0.189	0.200
Cop BIC	1	F	both	0.191	0.191	0.185	0.192	0.199	0.189	0.200
Cop AIC	1	FGM	both	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	FGM	both	0.191	0.191	0.185	0.191	0.198	0.189	0.200
Cop AIC	2	F	both	0.191	0.191	0.185	0.192	0.199	0.190	0.200
Cop BIC	2	F	both	0.192	0.192	0.185	0.192	0.199	0.190	0.200
Cop AIC	2	FGM	both	0.191	0.191	0.185	0.191	0.199	0.190	0.200
Cop BIC	2	FGM	both	0.191	0.191	0.185	0.191	0.199	0.189	0.200
ordinal	-	-	-	0.192	0.192	0.185	0.192	0.200	0.190	0.201

**Table 10: Fictional betting results (in gain ratio, gains per betted unit of currency) for all models and leagues. Cell colors best (green) to worst (red) for visualisation. See digital version.**

Model	Eq	Cop	regul.							
pois	1	-	-	0.024	0.024	-0.058	-0.084	-0.056	-0.078	-0.202
pois	2	-	-	0.017	0.017	0.003	-0.032	-0.077	-0.052	-0.138
pois	1	-	LASSO	0.017	0.017	-0.097	-0.088	-0.075	-0.135	-0.093
pois	2	-	LASSO	0.015	0.015	-0.019	-0.067	-0.048	-0.105	-0.130
RF	1	-	-	-0.004	-0.004	-0.046	-0.087	-0.046	-0.111	-0.190
RF	2	-	-	-0.039	-0.039	-0.116	0.001	-0.069	-0.077	-0.097
XGboost	1	-	-	-0.023	-0.023	-0.136	-0.067	-0.062	-0.102	-0.125
XGboost	2	-	-	0.001	0.001	-0.104	-0.067	-0.050	-0.092	-0.092
Cop	1	F	-	0.068	0.068	-0.049	-0.073	-0.043	-0.078	-0.215
Cop	1	FGM	-	0.054	0.054	-0.034	-0.070	-0.044	-0.081	-0.236
Cop	2	F	-	0.035	0.035	-0.001	0.027	-0.037	-0.062	-0.134
Cop	2	FGM	-	0.037	0.037	0.012	0.012	-0.040	-0.062	-0.127
Cop	1	F	equal	-0.016	-0.016	-0.083	-0.098	-0.086	-0.141	-0.212
Cop	1	FGM	equal	-0.025	-0.025	-0.082	-0.102	-0.077	-0.142	-0.198
Cop	2	F	equal	-0.003	-0.003	0.009	-0.071	-0.060	-0.072	-0.147
Cop	2	FGM	equal	-0.007	-0.007	0.023	-0.068	-0.064	-0.070	-0.150
Cop AIC	1	F	LASSO	0.037	0.037	-0.080	-0.068	-0.066	-0.113	-0.168
Cop BIC	1	F	LASSO	-0.007	-0.007	-0.059	-0.140	-0.045	-0.119	-0.153
Cop AIC	1	FGM	LASSO	0.017	0.017	-0.091	-0.088	-0.055	-0.111	-0.168
Cop BIC	1	FGM	LASSO	-0.006	-0.006	-0.099	-0.143	-0.097	-0.125	-0.141
Cop AIC	2	F	LASSO	0.035	0.035	0.047	-0.014	-0.091	-0.065	-0.197
Cop BIC	2	F	LASSO	-0.021	-0.021	-0.121	-0.070	-0.039	-0.115	-0.112
Cop AIC	2	FGM	LASSO	0.042	0.042	0.019	-0.025	-0.078	-0.061	-0.208
Cop BIC	2	FGM	LASSO	-0.024	-0.024	-0.136	-0.070	-0.047	-0.100	-0.123
Cop AIC	1	F	both	0.022	0.022	-0.045	-0.086	-0.161	-0.167	-0.340
Cop BIC	1	F	both	0.032	0.032	-0.045	-0.100	-0.125	-0.180	-0.337
Cop AIC	1	FGM	both	0.024	0.024	-0.070	-0.119	-0.146	-0.156	-0.357
Cop BIC	1	FGM	both	0.018	0.018	-0.070	-0.120	-0.191	-0.152	-0.357
Cop AIC	2	F	both	-0.022	-0.022	0.005	-0.154	-0.178	-0.155	-0.242
Cop BIC	2	F	both	0.025	0.025	-0.013	-0.149	-0.122	-0.233	-0.293
Cop AIC	2	FGM	both	-0.003	-0.003	-0.031	-0.136	-0.135	-0.134	-0.310
Cop BIC	2	FGM	both	0.011	0.011	-0.103	-0.125	-0.235	-0.135	-0.349
ordinal		-	-	0.024	0.024	0.006	-0.050	-0.083	-0.017	-0.152

**Table 11: Results for all modelling approaches. Calculated by combining all leagues to one large data set. Cell colors best (green) to worst (red) for visualisation. See digital version.**

Model	Eq	Copula	regul.	RPS	LH	CR	SE	AE	bets	gainratio
pois	1	-	-	0.1935	0.429	0.544	2.668	1.810	8316	-0.0734
pois	2	-	-	0.1936	0.428	0.544	2.666	1.808	9017	-0.0468
pois	1	-	LASSO	0.1935	0.428	0.544	2.667	1.809	8207	-0.0790
pois	2	-	LASSO	0.1935	0.428	0.544	2.664	1.808	8833	-0.0415
RF	1	-	-	0.1969	0.427	0.536	2.736	1.835	10318	-0.0612
RF	2	-	-	0.1953	0.427	0.539	2.694	1.822	9982	-0.0503
XGboost	1	-	-	0.1943	0.424	0.544	2.680	1.812	9691	-0.0690
XGboost	2	-	-	0.1944	0.424	0.543	2.674	1.811	9855	-0.0729
Cop	1	F	-	0.1933	0.430	0.545	2.668	1.810	6456	-0.0659
Cop	1	FGM	-	0.1932	0.430	0.545	2.668	1.810	6509	-0.0663
Cop	2	F	-	0.1934	0.429	0.544	2.666	1.808	7622	-0.0432
Cop	2	FGM	-	0.1934	0.429	0.544	2.666	1.808	7675	-0.0456
Cop	1	F	equal	0.1935	0.429	0.543	2.670	1.808	6004	-0.0639
Cop	1	FGM	equal	0.1935	0.429	0.543	2.670	1.808	6087	-0.0635
Cop	2	F	equal	0.1935	0.429	0.544	2.668	1.808	7551	-0.0565
Cop	2	FGM	equal	0.1935	0.429	0.544	2.668	1.808	7605	-0.0513
Cop AIC	1	F	LASSO	0.1933	0.430	0.545	2.668	1.810	6315	-0.0675
Cop BIC	1	F	LASSO	0.1934	0.429	0.545	2.669	1.809	6286	-0.0677
Cop AIC	1	FGM	LASSO	0.1933	0.430	0.545	2.668	1.810	6385	-0.0688
Cop BIC	1	FGM	LASSO	0.1934	0.429	0.545	2.669	1.809	6374	-0.0683
Cop AIC	2	F	LASSO	0.1935	0.429	0.544	2.667	1.808	7527	-0.0490
Cop BIC	2	F	LASSO	0.1934	0.429	0.544	2.669	1.810	7461	-0.0510
Cop AIC	2	FGM	LASSO	0.1935	0.429	0.544	2.665	1.807	7565	-0.0521
Cop BIC	2	FGM	LASSO	0.1935	0.429	0.544	2.667	1.809	7551	-0.0623
Cop AIC	1	F	both	0.1939	0.429	0.543	2.693	1.812	6255	-0.0687
Cop BIC	1	F	both	0.1939	0.429	0.543	2.692	1.812	6050	-0.0651
Cop AIC	1	FGM	both	0.1934	0.429	0.544	2.670	1.807	5773	-0.0660
Cop BIC	1	FGM	both	0.1934	0.429	0.544	2.670	1.807	5565	-0.0617
Cop AIC	2	F	both	0.1948	0.428	0.543	2.721	1.815	7103	-0.0859
Cop BIC	2	F	both	0.1947	0.428	0.544	2.722	1.816	5775	-0.0774
Cop AIC	2	FGM	both	0.1935	0.429	0.544	2.668	1.806	6786	-0.0930
Cop BIC	2	FGM	both	0.1934	0.429	0.545	2.670	1.806	5321	-0.0626
ordinal	-	-	-	0.1936	0.430	0.544	-	-	7937	-0.0506

## **welo: AN R PACKAGE FOR WEIGHTED AND STANDARD ELO RATES**

**Vincenzo Candila**<sup>1</sup>

*Department of Economics and Statistics, University of Salerno, Fisciano, Italy*

**Abstract** *This paper describes the characteristics of the `welo` package, dedicated to calculating the weighted and unweighted (or standard) Elo rates in tennis. The Elo rates are one of the most accurate proxies of the strength of players/teams. In the standard version, the Elo rates are dynamically obtained using the outcome of the two players. In the recent paper of Angelini et al. (2022), the weighted version of the Elo rates (labeled as WElo) has been proposed in order to take into account not only the outcome of the matches but also the scoreline. The present work illustrates the main features of the R package, which allows the user to easily and quickly obtain the WElo and Elo rates, as well as the predicted probabilities of winning.*

**Keywords:** *Elo rates, Weighted Elo rates, R, Tennis, Betting.*

### **1. INTRODUCTION**

The attention of the literature on sport's outcome forecasting has largely increased over the last few years. Many contributions focus on soccer (see, for instance, Angelini and De Angelis, 2017; Koopman and Lit, 2015; Mattera, 2021, among others) and tennis (see Arcagni et al., 2022; Lisi and Zanella, 2017, and references therein). Recently, Kovalchik (2020) has improved the Elo rates for tennis by taking into account, for the first time, the margin of victory. The Elo rates were proposed by the physics professor Arpad Elo in 1978 (Elo, 1978) for the rating of chess players. Since then, the Elo rates have been applied in a variety of sports: rugby (Carbone et al., 2016), soccer (Hvattum and Arntzen, 2010; Leitner et al., 2010), American football (Ryall and Bedford, 2010), and tennis (Kovalchik, 2016; Kovalchik and Reid, 2019). Angelini et al. (2022) have further extended the Elo-based models in tennis by weighting the Elo rates according to the number of games or sets won by each player. If the standard Elo rates take only into account the outcome of the match (that is, if a player has won or lost), the recently proposed Weighted Elo (WElo) rates of Angelini et al. (2022) instead are also based

---

<sup>1</sup>vcandila@unisa.it

on the final scoreline. This additional feature has provided a large benefit in using the WElo rates to calculate the probability of winning, compared to a set of the competing models<sup>2</sup>. The present paper aims at illustrating, in detail, all the tools of the `welo` package, which currently is available on the Comprehensive R Archive Network (CRAN).

There are several R packages on the CRAN and GitHub<sup>3</sup> repositories dealing with the Elo rating systems. But none of the available packages is suitable for calculating the WElo rates as the `welo` package. Moreover, the `welo` package allows the user to directly download tennis data using the <http://www.tennis-data.co.uk/> site, which is weekly updated. The `welo` package can also easily plot the WElo and Elo rates, and it is flexible to include specific and user-based weights to some match conditions (for instance, if the match is Grand Slam match or it is played on a given surface). Another feature is the setting of the scale factor (more details will be provided in the next section), which is used to define how much the rate changes after the end of the match. Finally, the `welo` package also calculates the profits and losses deriving from a set of betting strategies. In what follows, we describe the main features of existing R packages dealing with the Elo rates. These information are then synthesized in Table 1.

Package `e1o` (Heinzen, 2022) is on CRAN since 2017. It allows the calculation of the Elo rates both for team and non-team sports via its function `e1o.run`, which is very flexible because it only requires the indication of the points of the two contenders. However, it does not include the possibility of taking into account the past scoreline to predict future winning probabilities. It neither allows to weight differently specific matches (for instance, the tennis matches played on a particular surface).

Package `EloRating` (Neumann and Kulik, 2020) is devoted to quantify animal dominance hierarchies. However, the main function providing the Elo rates, labeled `fastelo`, could also be used for non-animals data. In particular, it is sufficient to include as inputs in `fastelo` the names of the winners and losers, match by match. On the other side, `EloRating` package can not consider the scoreline of the last matches or specific match conditions. Moreover, it does not allow for a dynamic choice for the scale factor.

Package `EloOptimized` (Feldblum et al., 2021) has the maximum likelihood

---

<sup>2</sup>The set of competing models used in Angelini et al. (2022) is: the standard Elo model, the Bradley-Terry type model (McHale and Morton, 2011), the logit and probit regressions of Klaassen and Magnus (2003) and Del Corral and Prieto-Rodriguez (2010), respectively.

<sup>3</sup>GitHub hosts freely R packages, many of which are under development before being published on the CRAN.

estimation of the scale factor as the main feature. In particular, such a scale factor is not fixed by the user (even though there is also this possibility), but it is estimated by maximizing the likelihood of the sigmoid probability function as defined by Foerster et al. (2016). In addition, following the same maximum likelihood procedure, `EloOptimized` can also estimate the initial Elo rates.

Package `EloChoice` (Neumann, 2019) calculates the Elo rates through the `elochoice` function. However, the scale factor is fixed and there are no possibilities of setting different weights according to specific match conditions.

Package `comperank` (Chasnovski, 2020a) offers a variety of ranking and rating based on competition methods. Among these methods, the user can obtain the Elo rates via the `elo` function. One of the advantages of the `elo` function is the possibility of having ties. But, on the other side, the `comperank` package requires a specific format of the matches' data, making use of the `as_longcr` function of the `comperes` package (Chasnovski, 2020b). Also for this package, the resulting Elo rates consider a fixed scale factor and do not take into account the match conditions.

There are at least three packages dealing with the Elo rates on GitHub: `elomov`, `mELO` and `bwsTools`. The package `elomov` implements the Elo rates with the margin of victory option, as recently proposed by Kovalchik (2020). At the time of this writing, the whole installation of the `elomov` package via GitHub does not work. However, it is possible to manually install the functions of the package. The package `ELO`, using the `LO` function, also admits ties, but the resulting Elo rates are based on a fixed scale factor. Finally, the package `wTools` allows for the calculation of the Elo rates via `elo` function. The `bwsTools` package does not allow for a time-varying scale factor or for a different rate according to specific match conditions.

Finally, none of the previously cited packages implements betting functions.

The rest of the paper is as follows. Section 2 illustrates how to compute the `WElo` and Elo rates. Section 3 presents the details of the `welo` package for computing the `WElo` and Elo rates. Section 4 is devoted to the betting application via the `welo` package. Conclusions follow.

## 2. WEIGHTED AND STANDARD ELO RATES

Throughout all the work, we use the same notation of Angelini et al. (2022). Therefore,  $i$  and  $j$  will indicate two opponents in a tennis match and  $E_i(t)$  and  $E_j(t)$  their Elo ratings for the match at time  $t$ . Then, the probability that player  $i$

**Table 1: R packages**

Name	Repository	Weighted rates	Scale factor
welo	CRAN	Yes	Varying or fixed
elo	CRAN	No	Fixed
EloRating	CRAN	No	Fixed
EloOptimized	CRAN	No	Estimated
EloChoice	CRAN	No	Fixed
comparank	CRAN	No	Fixed
elomov	GitHub	Yes	Fixed
mELO	GitHub	No	Fixed
bwsTools	GitHub	No	Fixed

wins against player  $j$  in match  $t$  is:

$$\hat{p}_{i,j}(t) = \frac{1}{1 + 10^{(E_j(t) - E_i(t))/400}}. \quad (1)$$

The formula updating the Elo ratings for player  $i$  is:

$$E_i(t+1) = E_i(t) + K_i(t) [W_i(t) - \hat{p}_{i,j}(t)], \quad (2)$$

where  $W_i(t)$  represents an indicator function, which is one if player  $i$  wins match  $t$  and zero otherwise, and  $K_i(t)$ , as mentioned above, is a scale factor determining how much the Elo rate changes after match  $t$ . Such a scale factor is crucial in making effective the differences between the rates across players. It could be fixed to a given value (as many existing packages do). It could be estimated (as the `EloOptimized` package does). Or, as the `welo` package does, it could be fixed, time-varying or even time-varying and, jointly, larger for some specific tournaments or surfaces.

The WElo rates, contrary to what happens for the Elo standard rates, allow for the consideration of the scoreline of the matches in the updating formula. More in detail, Eq. (2) incorporates an additional function  $f(\cdot)$ , depending on the number of games  $G_{i,j}(t)$  or number of sets  $S_{i,t}(t)$  won by players  $i$  and  $j$  during match  $t$ . When the WElo rates depend on the number of games  $G_{i,j}(t)$ , the rates (for player  $i$ ) are defined as:

$$E_i^*(t+1) = E_i^*(t) + K_i(t) [W_i(t) - \hat{p}_{i,j}^*(t)] f(G_{i,j}(t)), \quad (3)$$

where  $\hat{p}_{i,j}^*(t)$  is estimated using Eq. (1) but with  $E_i(t)$  and  $E_j(t)$  replaced by the corresponding WElo rates, labeled as  $E_i^*(t)$  and  $E_j^*(t)$ , respectively. In Eq. (3),  $f(G_{i,j}(t))$  is a function whose values depend on the games played in the previous match. In particular,  $f(G_{i,j}(t))$  is defined as:

$$f(G_{i,j}(t)) = \begin{cases} \frac{NG_i(t)}{NG_i(t)+NG_j(t)} & \text{if player } i \text{ has won match } t; \\ \frac{NG_j(t)}{NG_i(t)+NG_j(t)} & \text{if player } i \text{ has lost match } t, \end{cases} \quad (4)$$

where  $NG_i(t)$  and  $NG_j(t)$  represent the number of games won by player  $i$  and player  $j$  in match  $t$ , respectively.

When the WElo rates depend on the number of sets,  $f(S_{i,t}(t))$  is obtained as:

$$f(S_{i,j}(t)) = \begin{cases} \frac{NS_i(t)}{NS_i(t)+NS_j(t)} & \text{if player } i \text{ has won match } t; \\ \frac{NS_j(t)}{NS_i(t)+NS_j(t)} & \text{if player } i \text{ has lost match } t, \end{cases} \quad (5)$$

where  $NS_i(t)$  and  $NS_j(t)$  represent this time the sets won by player  $i$  and player  $j$  in match  $t$ , respectively. Then,  $f(S_{i,t}(t))$  replaces  $f(G_{i,j}(t))$  in Eq. (3).

### 3. WELO AND ELO RATES THROUGH THE welo PACKAGE

For ease of replicability, the interested user can reproduce all the following codes, once that the welo package has been installed from CRAN and loaded, that is:

```
R> install.packages("welo") # only the first time
R> library(welo)
```

The first step for using the welo package is the collection of tennis matches. By means of the tennis\_data function, this step is immediately achieved:

```
R> db<-tennis_data("2021", "ATP")
```

By the previous code, the db object includes all the matches played in 2021 for the Association of Tennis Professionals (ATP). If we are interested in female matches, then we can replace “ATP” by “WTA”, where WTA stands for *Women Tennis Association*.

The second step for obtaining the WElo and Elo rates is cleaning the data. This operation is extremely delicate and is performed accurately through the clean function:

```
R> db_cleaned<-clean(db)
Number of matches (before cleaning) 2489
```



```

Number of matches (after cleaning) 1771
Number of players (before cleaning) 307
Number of players (after cleaning) 121

```

After running the `clean` function, some information automatically appear: the number of matches and players before and after the cleaning. More in detail, the `clean` function executes the following steps:

1. Remove all the uncompleted matches;
2. Remove all the NAs from B365 odds;
3. Remove all the NAs from the variable “ranking”, if any;
4. Remove all the NAs from the variable “games”, if any;
5. Remove all the NAs from the variable “sets”, if any;
6. Remove all the matches where the odds provided by the professional bookmaker Bet365 are equal, if any;
7. Define players  $i$  and  $j$  and their outcomes ( $Y_i$  and  $Y_j$ );
8. Remove all the matches of players who played less than the parameter of the `clean` function defined as `MNM`. By default, `MNM = 10`, which means that all the players playing less than 10 matches in `db` are excluded;
9. Remove all the matches of players with rank greater than the `MRANK` parameter. By default, `MRANK = 500`, which means that all the matches involving players rank above position 500 are excluded;
10. Sort the matches by date.

Changing the optional parameters of the `clean` function will return different cleaned datasets. For instance, if the interest is in the top-100 players playing at least one match, then the code will be:

```

R> db_cleaned_top_100<-clean(db, MNM=1, MRANK=100)
Number of matches (before cleaning) 2489
Number of matches (after cleaning) 1386
Number of players (before cleaning) 307
Number of players (after cleaning) 116

```

Finally, the `clean` function configures the dataset to be ready for the core function of the `welo` package, that is `welofit`. This is done by adding the columns of  $NG_i$ ,  $NG_j$ ,  $NS_i$ ,  $NS_j$ ,  $f(G_{i,j}(t))$  and  $f(S_{i,j}(t))$  to the cleaned `db`.

As mentioned above, the most important function of the `welo` package is the `welofit` function, which is very flexible and has several options. By default, it calculates the WElo and Elo rates with the following code:

```
R> res<-welofit(db_clean)
      Brier Log-Loss
WElo 0.2274  0.6451
Elo  0.2325  0.6581
```

As for the `clean` function, also the `welofit` function automatically synthesizes some information in the console after the execution. In this case, the user can quickly verify if the WElo performs better or worse than the standard Elo rates, according to the Brier (Brier, 1950) and Log-Loss (used by Kovalchik, 2016, among others) loss functions. These two loss functions map the distance between the predicted probability and the actual outcome of all the matches. The smaller the loss function is, the better that model is. By default, the WElo and Elo rates are calculated using the time-varying scale factor reported in Kovalchik (2016), that is:

$$K_i(t) = \frac{250}{(N_i(t) + 5)^{0.4}}, \quad (6)$$

where  $N_i(t)$  represents the number of matches of player  $i$  at time  $t$ . This configuration increases the variation of the Elo and WElo ratings if player  $i$  has played few matches and vice versa.

Finally, the default setting of the `welo` function considers the scores of the games (see Eq. (4)) for the WElo rates, the starting points fixed to 1500, while the standard errors are not estimated.

Let us now focus on the resulting object of the `welo` function, which, in this case, has been called `res`. This object is a ‘welo’ object, which is a list containing the following components:

```
R> class(res)
[1] "welo"
R> names(res)
[1] "results" "matches" "period" "loss" "highest_welo"
[6] "highest_elo" "dataset"
```

The previous components are:

1. results: The data.frame including a variety of variables, among which there are the estimated WElo and Elo rates, before and after the match  $t$ , for players  $i$  and  $j$ , the probability of winning the match for player  $i$  (labeled as WElo\_pi\_hat and Elo\_pi\_hat, for the probabilities obtained from the WElo and Elo models, respectively).
2. matches: The number of matches analyzed.
3. period: The sample period considered.
4. loss: The Brier score and log-loss averages.
5. highest\_welo: The player with the highest WElo rate and the correspondent date.
6. highest\_elo: The player with the highest Elo rate and the correspondent date.
7. dataset: The dataset used for the estimation of the WElo and Elo rates.

The `welo` function allows for a variety of options. Firstly, the WElo rates can be calculated using the sets instead of the games. This is can be easily achieved through:

```
R> res_s<-welofit(db_clean ,W="SETS" )
      Brier Log-Loss
WElo 0.2301  0.6521
Elo   0.2325  0.6581
```

Unsurprisingly, the (smaller) information content included in the sets, with respect to the games, worsens the WElo performance.

Moreover, the user can change the starting values of the WElo and Elo rates setting the parameter `SP` to another option. For instance, if the user wants the starting values equal to 1000 (instead of 1500, which is the default value), it is sufficient to run the following code:

```
R> res_1000<-welofit(db_clean ,SP=1000)
      Brier Log-Loss
WElo 0.2274  0.6451
Elo   0.2325  0.6581
```

Unexpectedly, it can be noted that the performances of the WElo and Elo models, when the starting values are set to 1000, are the same as the case with the starting points equal to 1500.

Another interesting feature of the `welo` function is flexibility of the scale factor  $K_i(t)$  (and  $K_j(t)$ ). By default, the scale factor is time-varying, according to Eq. (6). But such a parameter could be easily changed to be constant. For instance, if the user wants a constant scale factor of 100, the code will be:

```
R> res_K_100<-welofit(db_clean ,K=100)
      Brier Log-Loss
WElo 0.2274  0.6450
Elo  0.2340  0.6619
```

In this case, the better performance of the WElo model appears even more evident. Another possibility is to set  $K$  such that more weight is given to specific tournaments or match surfaces. Currently, four options are available: “Grand\_Slam”, “Surface\_Hard”, “Surface\_Clay” and “Surface\_Grass”. Each of the previous options increases the time-varying scale factor in (6) by 1.1 if the match is a Grand Slam match, is played on hard, clay, or grass, respectively. For instance, if the user wants to calculate the WElo and Elo rates giving more emphasis on the Grand Slam matches, then the code will be:

```
R> res_gs<-welofit(db_clean ,K="Grand_Slam")
      Brier Log-Loss
WElo 0.2272  0.6447
Elo  0.2325  0.6584
```

Another peculiar feature of the `welofit` function is the calculation of the standard errors for the WElo and Elo rates, according to the procedure suggested by Angelini et al. (2022). The code will be:

```
R> res_ci<-welofit(db_clean ,CI=TRUE)
      Brier Log-Loss
WElo 0.2274  0.6451
Elo  0.2325  0.6581
```

The resulting Brier and Log-Loss averages are exactly the same of `res`. This is because the setting parameters are unchanged. But, this time, the “results” component of `res_ci` includes also the lower (labeled with the suffix “\_lb”) and upper (labeled with the suffix “\_ub”) bootstrap confidence intervals. The confidence intervals are obtained according to the procedure illustrated in Angelini et al. 2022 (see their Section 2.1). By default, the bootstrap confidence intervals are obtained

**Table 2: Grand Slam 2021 finals, WElo rates and standard errors**

Grand Slam	Players	WElo	$\hat{p}_{i,j}(t)$	LB	UB
Australian Open	i) <b>Djokovic N.</b>	1704.187	0.512	1655.157	1750.961
	j) Medvedev D.	1696.004		1649.230	1745.034
Roland Garros	i) <b>Djokovic N.</b>	1847.900	0.487	1816.258	1881.268
	j) Tsitsipas S.	1857.125		1829.618	1883.210
Wimbledon	i) <b>Djokovic N.</b>	1894.685	0.632	1856.548	1916.848
	j) Berrettini M.	1800.394		1778.231	1838.531
US Open	i) Djokovic N.	1943.422	0.649	1906.715	1963.316
	j) <b>Medvedev D.</b>	1837.013		1818.781	1870.651

Note: Winning player is in **Bold**.

using a significance level  $\alpha = 0.05$  and a number of bootstrap replicates  $B = 1000$ . The WElo rates calculated before each Grand Slam 2021 final, together with the bootstrap standard errors and the probability that player  $i$  wins over player  $j$  (that is,  $\hat{p}_{i,j}(t)$ ) are reported in Table 2.

One of the most interesting features of the `welo` package is the possibility of plotting the WElo and Elo rates in nice graphs. The plot can be obtained by the `welo_plot` function, whose only input required is the (character) vector of players. Being in a `ggplot2` environment, the user can complete the plot by adding font size details via the `ggplot2::theme()` option. Suppose that the user wants the plot of the WElo rates for the following players: Nadal, Djokovic, Berrettini, and Sinner. Moreover, suppose that the user considers the rates from the `res` object previously obtained. Then, the code will be:

```
R> require(ggplot2)
R> players<-c("Nadal R.", "Djokovic N.",
"Berrettini M.", "Sinner J.")
R> welo_plot(res, players)+
ggplot2::theme(text = element_text(size = 20))
```

The output of the previous lines is in Figure 1(a). Figure 1(a) has some interesting peculiarities. First, at the end of 2021, Djokovic was largely the player with the highest WElo rate. Second, there is evidence of periods where some players did not play. These periods are highlighted in the plot with a horizontal line. For instance, during the second half of the 2021 season, Nadal played

only two matches (in August, at the Washington City Open) after the defeat at the Roland Garros in June. This is the reason why Nadal's orange line is horizontal from mid-June to the end of 2021. By default, the WElo rates are considered. Changing the optional parameter `rates` of `welo_plot` from "WElo" to "Elo", the standard Elo rates depicted in Figure 1(b) are obtained. It is worth noting that the patterns of the WElo and Elo rates are very similar, even though the former are always smaller than the latter.

```
R> welo_plot(res, players, rates="Elo")+  
ggplot2::theme(text = element_text(size = 20))
```

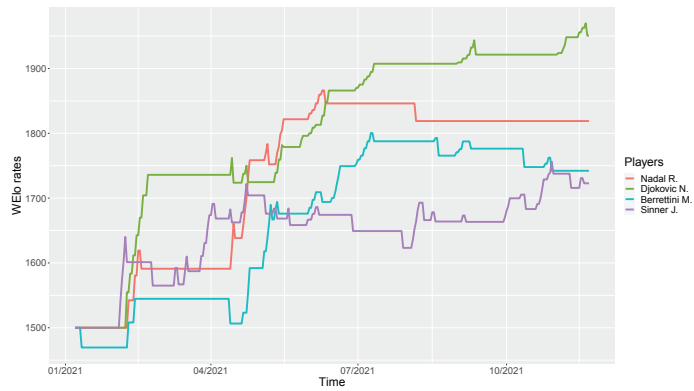
The `welo` package also provides a function to plot players' official (ATP or WTA) rank. The following code will plot (in Figure 1(c)) the official ranks of the four players already used in Figures 1(a) and 1(b):

```
R> rank_plot(res, players)+ ggplot2::theme(text =  
element_text(size = 20))
```

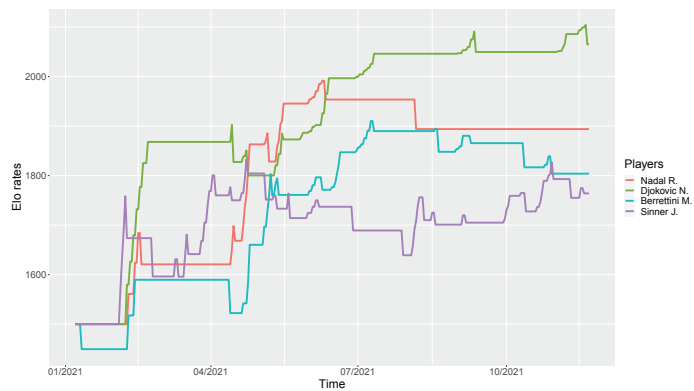
Some considerations arise looking at Figure 1. First, a necessary burn-in period is required to make the WElo and Elo rates reliable. This problem could be easily solved by enlarging the sample period. Second, even though at the end of sample, the WElo and Elo rates have the same order of the official ATP rank, the best player during Spring 2021 is Rafael Nadal for the WElo and Elo rates (in place of the official number one of the ATP rank, Novak Djokovic). Third, at the end of the 2021 season, mainly for the WElo rates, the young Italian player Jannik Sinner is pretty close to the other Italian tennis top player, Matteo Berrettini. This closeness between the two Italian players is not overall captured by the official ATP rank.

#### 4. BETTING WITH THE `welo` PACKAGE

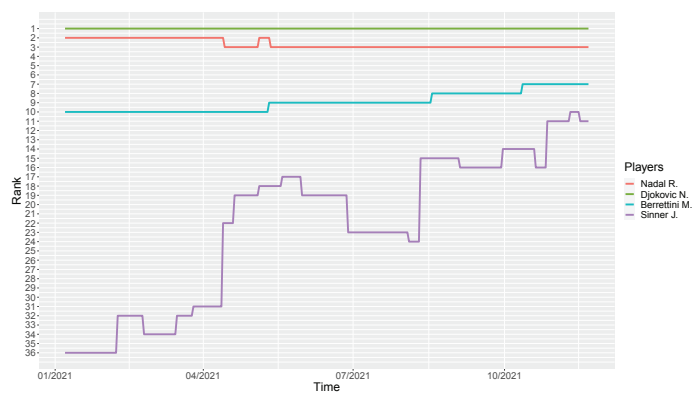
In the sports literature, one of the main aims of a forecasting model is to verify its performance from an economic point of view. This can be easily achieved through the `welo` package, thanks to the `betting` function. Such a function represents a novelty in the context of R packages dealing with Elo rates because none of the existing packages has similar functions at the time of this writing. The `betting` function requires four inputs to work: `x`, `r`, `q` and `model`. The first input `x` is a 'welo' object from the `welofit` function. The second and third inputs `r` and `q` are two thresholds that identify the matches on which place an amount of \$1. More in detail, as suggested by Angelini et al. (2022), the bets are placed on the matches



(a) WElo rates



(b) Elo rates



(c) ATP Rank

**Figure 1: Plots of the weLo package**

satisfying the following conditions:

$$\frac{\hat{P}_{i,j}(t)}{q_{i,j}(t)} > r \quad \text{and} \quad q_{i,j}(t) > q, \quad (7)$$

where  $\hat{P}_{i,j}(t)$  are the two probabilities as resulting from the Elo and WElo models for the match between  $i$  and  $j$  at time  $t$ , that is,  $\hat{P}_{i,j}(t) = \{\hat{p}_{i,j}(t), \hat{p}_{i,j}^*(t)\}$ , and  $q_{i,j}(t)$  is the inverse of the published odds for the same match, also named implied probability. For coverage reasons, the `welo` package considers the implied probabilities  $q_{i,j}(t)$  offered by the professional bookmaker Bet365. The user can decide the model (WElo or Elo) originating the probabilities of winning via the fourth input of the `betting` function, that is setting `model = "WELO"` or `model = "ELO"`.

In line with McHale and Morton (2011) and Dixon and Coles (1997), the threshold  $r$  is used to discriminate among matches on which place a bet or not. For instance, if  $r = 1$ , only matches whose predicted probabilities are greater than the implied probabilities are worthy of a bet. When  $r$  increases, fewer matches will be selected. The `betting` function allows also for the inclusion of a set of values for  $r$ . As concerns the threshold  $q$ , such a value is needed to exclude heavy underdogs. For instance, when  $q = 0.30$ , then all the matches whose Bet365 implied probabilities are smaller than 0.30 will be excluded. Bearing this in mind, a general configuration of the `betting` function could be:

```
R> res_bet_welo<-betting(res , r=seq(1 , 1.3 , 0.05) ,
q=0.3 , model="WELO")
```

	r	# Bets	ROI(%)	LCI	UCI
[1, ]	1.00	1002	10.045908	3.4753003	16.22180
[2, ]	1.05	823	8.883354	1.4252170	16.11634
[3, ]	1.10	655	10.175573	1.8184030	18.80244
[4, ]	1.15	533	10.553471	1.4241092	20.09728
[5, ]	1.20	438	11.808219	1.1060541	22.28150
[6, ]	1.25	338	13.239645	0.6448331	25.11509
[7, ]	1.30	265	16.584906	1.6015294	31.78393

The predicted probabilities included in the ‘welo’ object labeled `res` are considered in the previous command. The resulting output of the `betting` function is a matrix that includes five columns. The first column reports the values of the threshold  $r$ . Hence, among the 1771 matches of the full dataset, the betting rule suggests of betting on 1002 matches, when  $r = 1$  and  $q = 0.3$ . The second column



includes the number of bets (for each correspondent threshold  $r$ ). As mentioned above, the higher the threshold is, the smaller the number of matches to bet on is. The third column reports the returns-on-investment (ROI), in percentage. The last two columns show the lower (LCI) and upper (UCI) bootstrap confidence intervals, computed using the default number of bootstrap replicates ( $R = 2000$ ) and the default significance level ( $\alpha = 0.1$ ). The user can easily change those two settings. In what follows, there is the code and the output for the Elo probabilities:

```
R> res_bet_elo<-betting(res,r=seq(1,1.3,0.05),
q=0.3,model="ELO")
      r # Bets   ROI(%)    LCI    UCI
[1,] 1.00  1096  6.914234  1.4332144 12.67208
[2,] 1.05   893  7.767077  1.0466995 14.56551
[3,] 1.10   713  8.830295  1.2234941 16.32349
[4,] 1.15   607  9.059308  0.2738958 17.10543
[5,] 1.20   506 11.373518  2.0207622 21.47579
[6,] 1.25   409  9.312958 -1.1881413 20.27323
[7,] 1.30   329 11.556231 -0.7741607 23.21550
```

Interestingly, it can be noted that the ROI(%) of the WElo probabilities are higher than the corresponding Elo probabilities, independently of the threshold  $r$  adopted. Moreover, all the ROI(%) of the WElo model are statistically significant, while the same does not happen for the ROI(%) of the Elo model.

Finally, the `betting` function has some optional parameters which could be set: `bets`, `R`, `alpha`, `start_oos`, and `end_oos`. The parameter `bets` identifies the type of bet used. By default, it is “Best\_odds”, which means that the bets are placed using the best odds available among all the bookmakers. Alternative choices for `bets` are: “Avg\_odds” and “B365\_odds”. “Avg\_odds” are the average odds among all the odds published by the professional bookmakers for the match under consideration and “B365\_odds” are the Bet365 odds. The parameter `R` represents the number of bootstrap replicates to calculate the confidence intervals of the ROI(%). Its default value is 2000. The parameter `alpha` is the significance level for the bootstrap confidence intervals. By default,  $\alpha = 0.1$ . Eventually, the user could also bet on a specific time period. This is can be easily done setting the parameters `start_oos` and `end_oos`, which have to be formatted as “YYYY”. For instance, if the user is interested in the time interval from 2021 to 2022, then he/she has to format `start_oos = “2021”` and `end_oos = “2022”`.<sup>4</sup>

<sup>4</sup>The time interval from 2021 to 2022 would require a larger dataset including also data for 2022.

For comparison purposes, the `welo` package includes also another betting function, labeled `random_betting`. Such a function is useful when the user wants to evaluate if randomly betting on players  $i$  and  $j$  is a winning strategy with respect to the decision on the basis of the WElo and Elo probabilities. To make a fair comparison, `random_betting` shares almost all the inputs with the function `betting`: this means that the user can set the two functions similarly to select the same matches. As said before, in the case of `random_betting`, the players  $i$  and  $j$  are randomly selected. To reduce the impact of this randomness, the `random_betting` function repeats the random selection  $B$  times, which is the only (optional) parameter of the `random_betting` function not included in the `betting` function. By default,  $B = 10000$ . The resulting matrix reports the overall mean of the ROI (in percentage) across the  $B$  values for every threshold  $r$  used. The code will be:

```
R> res_rand_bet<-random_betting(res,r=seq(1,1.3,0.05),
q=0.3,model="WELO")
      r # Bets   ROI(%)
[1,] 1.00  1002 2.589963
[2,] 1.05   823 2.907354
[3,] 1.10   655 3.340112
[4,] 1.15   533 2.776383
[5,] 1.20   438 3.015227
[6,] 1.25   338 3.906786
[7,] 1.30   265 4.292838
R> res_rand_bet<-random_betting(res,r=seq(1,1.3,0.05),
q=0.3,model="ELO")
      r # Bets   ROI(%)
[1,] 1.00  1096 2.328365
[2,] 1.05   893 2.419626
[3,] 1.10   713 2.052510
[4,] 1.15   607 2.458521
[5,] 1.20   506 3.137259
[6,] 1.25   409 2.724166
[7,] 1.30   329 3.513878
```

From the last two R outputs, it can be noted that the random selection of players on which place a bet, even if repeated  $B$  times, does not yield larger ROI(%) with respect to the previous two ROI(%) obtained from the WElo and Elo rates.

## 5. CONCLUSIONS

The present contribution aimed at explaining the details of the `welo` package, an R package for the calculation of the standard (that is, unweighted) and weighted Elo (WElo) rates for tennis. The `welo` package has some interesting features: (i) the direct download of data for male and female professional tennis matches (via the `tennis_data` function); (ii) the cleaning of the tennis data (through the `clean` function); (iii) the calculation of standard and WElo rates by the core function of the package labeled `welofit`, with the possibility of weighting differently some tournaments and surfaces and having constant or time-varying scale factor; (iv) the plot of the resulting Elo and WElo rates with the `welo_plot` function; (v) the economic evaluation of the returns-on-investment (ROI) obtained from the predicted probabilities of the Elo and WElo rates, according to the betting rule of Angelini et al. (2022) and references therein; (vi) the comparison of the previous ROI with the ROI obtained from the random betting strategy.

The current paper can serve as a guide for practitioners and R users for the first time dealing with the calculation of the Elo and WElo rates. Further extensions of the `welo` package could enlarge the sports under consideration, like basket, volley and so forth.

## References

- Angelini, G., Candila, V. and De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. In *European Journal of Operational Research*, 297 (1): 120–132.
- Angelini, G. and De Angelis, L. (2017). PARX model for football match predictions. In *Journal of Forecasting*, 36 (7): 795–807.
- Arcagni, A., Candila, V. and Grassi, R. (2022). A new model for predicting the winner in tennis based on the eigenvector centrality. In *Annals of Operations Research*, 1–18.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. In *Monthly Weather Review*, 78 (1): 1–3.
- Carbone, J., Corke, T. and Moisiadis, F. (2016). The rugby league prediction model: Using an Elo-based approach to predict the outcome of National Rugby League (NRL) matches. In *International Educational Scientific Research Journal*, 2 (5): 26–30.

- Chasnovski, E. (2020a). *comperank: Ranking Methods for Competition Results*. URL <https://CRAN.R-project.org/package=comperank>. R package version 0.1.1.
- Chasnovski, E. (2020b). *comperes: Manage Competition Results*. URL <https://CRAN.R-project.org/package=comperes>. R package version 0.2.5.
- Del Corral, J. and Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? In *International Journal of Forecasting*, 26 (3): 551–563.
- Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46 (2): 265–280.
- Elo, A.E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub.
- Feldblum, J., Foerster, S. and Franz, M. (2021). *EloOptimized: Optimized Elo Rating Method for Obtaining Dominance Ranks*. URL <https://CRAN.R-project.org/package=EloOptimized>. R package version 0.3.1.
- Foerster, S., Franz, M., Murray, C.M., Gilby, I.C., Feldblum, J.T., Walker, K.K. and Pusey, A.E. (2016). Chimpanzee females queue but males compete for social status. In *Scientific Reports*, 6 (1): 1–11.
- Heinzen, E. (2022). *elo: Ranking Teams by Elo Rating and Comparable Methods*. URL <https://CRAN.R-project.org/package=elo>. R package version 3.0.1.
- Hvattum, L.M. and Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. In *International Journal of Forecasting*, 26 (3): 460–470.
- Klaassen, F.J. and Magnus, J.R. (2003). Forecasting the winner of a tennis match. In *European Journal of Operational Research*, 148 (2): 257–267.
- Koopman, S.J. and Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. In *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178 (1): 167–186.

- Kovalchik, S.A. (2016). Searching for the GOAT of tennis win prediction. In *Journal of Quantitative Analysis in Sports*, 12 (3): 127–138.
- Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. In *International Journal of Forecasting*, 36: 1329–1341.
- Kovalchik, S. and Reid, M. (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis. In *International Journal of Forecasting*, 35 (2): 756–766.
- Leitner, C., Zeileis, A. and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. In *International Journal of Forecasting*, 26 (3): 471–481.
- Lisi, F. and Zanella, G. (2017). Tennis betting: Can statistics beat bookmakers? In *Electronic Journal of Applied Statistical Analysis*, 10 (3): 790–808.
- Mattera, R. (2021). Forecasting binary outcomes in soccer. In *Annals of Operations Research*, 1–20.
- McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. In *International Journal of Forecasting*, 27 (2): 619–630.
- Neumann, C. (2019). *EloChoice: Preference Rating for Visual Stimuli Based on Elo Ratings*. URL <https://CRAN.R-project.org/package=EloChoice>. R package version 0.29.4.
- Neumann, C. and Kulik, L. (2020). *EloRating: Animal Dominance Hierarchies by Elo Rating*. URL <https://CRAN.R-project.org/package=EloRating>. R package version 0.46.11.
- Ryall, R. and Bedford, A. (2010). An optimized ratings-based model for forecasting Australian rules football. In *International Journal of Forecasting*, 26 (3): 511–517.

## A CAUSAL INVESTIGATION OF PACE OF PLAY IN SOCCER

**Nirodha Epasinghe Dona, Tim Swartz<sup>1</sup>**

*Department of Statistics and Actuarial Science, Simon Fraser University,  
Burnaby BC, Canada V5A1S6*

**Abstract** *This paper provides a comprehensive investigation of playing with pace in soccer. The investigation begins by introducing two quantitative definitions of pace whose calculations are facilitated through the availability of player tracking data. In the study, the primary scientific question concerns whether playing with pace is an advantageous strategy in terms of playing style. This is a question that has not been adequately resolved in either soccer or hockey. Here, we use methods of causal inference to investigate the relationship between pace in soccer and shots. It is determined that playing with more pace than the opponent throughout a match confers an advantage of approximately two additional shots per game. As a byproduct of our analysis, other soccer insights related to pace are obtained.*

**Keywords:** *Big data, Causal inference, Player tracking data, Spatio-temporal analyses, Style of play.*

### ACKNOWLEDGMENT:

Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work has been carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics. The authors thank Daniel Stenz, former technical director of Shandong Luneng Taishan FC who provided the data used in this paper. We also thank an associate editor and three anonymous reviewers whose comments have improved the paper.

---

<sup>1</sup>Corresponding Author: Tim Swartz, [tswartz@sfu.ca](mailto:tswartz@sfu.ca)

## 1. INTRODUCTION

In team sports, playing style is a much discussed topic and an important component of success. For example, in soccer, we hear about the gegenpress (Tweeddale 2022), total football (McLellan 2010) and parking the bus (Guan, Cao and Swartz 2022). However, playing style is notoriously difficult to quantify in soccer. It is difficult to quantify since playing style is a team concept, which relies on the actions of multiple players whose movements are fluid in both time and space.

However, the landscape for studying playing style has changed in recent years with the advent of player tracking data. With player tracking data, the location coordinates for every player on the field are recorded frequently (e.g. 10 times per second in soccer). With such detailed data, the opportunity to explore novel questions in sport has never been greater. The massive datasets associated with player tracking also introduce data management issues and the need to develop modern data science methods beyond traditional statistical analyses. Gudmundsson and Horton (2017) provide a review of spatio-temporal analyses that have been used in invasion sports where player tracking data are available.

This paper is concerned with “pace of play” in soccer, a relatively underexplored topic. In some sports, pace is readily defined. For example, in basketball, team pace may be defined as the average number of possessions per game. In the NBA, this is a well-studied statistic which is available from various websites including <https://www.nba.com/stats/teams/advanced/>

In American football, although there is a clear notion of pace of play, there is no commonly reported statistic that directly measures pace. In the National Football League (NFL), the number of plays per game is available for each team from standard box scores. Although this statistic is related to pace, it is obvious that poor offensive teams who rarely make first downs have fewer plays per game. Therefore, in football, the average number of plays per game for a team is confounded with offensive strength, and consequently, the number of plays is not a pure measure of pace. Pace in football can be increased for a team by using a “hurry-up offense” which affords more plays in a given period of time provided that the team continues to make first downs. Furthermore, teams that frequently pass the ball (as opposed to run the ball) typically use up less of the clock and have more plays from scrimmage.

In ice hockey, the definition of pace is even less clear. See, for example, Silva, Davis and Swartz (2018) where various definitions of pace are considered. Yu et al. (2018) revisit the hockey problem and suggest an alternative definition of pace.

The sport of soccer shares some of the same challenges as hockey with re-

spect to the definition of pace. For example, how is possession determined? How do successful passes contribute to pace and should pace calculations involving a pass be counted differently than when dribbling? Shen, Santo and Akande (2022) builds on the aforementioned hockey papers and uses event data to investigate pace in soccer.

This paper differs from Shen, Santo and Akande (2022) in a number of key directions. First, this paper uses tracking data rather than event data to study pace. Second, alternative definitions of pace are provided. In particular, we define *attacking pace* which is related to “direct play”, a much discussed tactic in soccer. Third, we provide various sporting implications associated with pace. Finally, our primary goal addresses the key question of whether playing with pace is strategically sound. Many soccer experts believe that moving the ball quickly is advantageous. When you move the ball quickly, the logic is that it affords the defensive team less time to transition to solid defensive formations. However, to our knowledge, this basic tenet of soccer has never been tested. Is it better to play with pace? We address this question by using methods of causal inference (Pearl 2009). Obviously, decisions that are made on the field are often instantaneous. Therefore, it is impossible to use traditional randomized trials to determine the cause-and-effect relationship between playing with pace and success. With match data, we have “studies” as opposed to “experiments”. Fortunately, the methods of causal inference allow us to address causality in studies provided that confounding variables can be identified and measured. With tracking data and our subject knowledge of soccer, confounding variables are accessible.

Related to our investigation of pace, they have been many investigations of determinants of success in soccer. A sample of recent papers include Lepschy, Wäsche and Woll (2021), Merlin et al. (2020), and in the women’s game, de Jong et al. (2020).

In soccer, the most investigated aspect of playing style concerns formations. For example, the book “Inverting the Pyramid” (Wilson 2013) considers the history of soccer tactics throughout the world with an emphasis on positional play and player roles. It is also now common during television broadcasts to provide graphical statistics that depict the average location of each player during a match. Such information is useful in determining match strategy as it can point out features such as gaps in player alignment. There have also been many technical papers written on player formation. For example, Shaw and Glickman (2019) use tracking data and clustering methods to determine a team’s offensive and defensive formations. This is useful as the fluidity of the sport and changing tactics



sometimes makes it difficult to distinguish between formations (e.g. 4-4-2 versus 3-5-2). Goes et al. (2021) also identify formations using tracking data and relate attacking success to formations.

The association between style and results in soccer has been well investigated. For example, in their Table 1, Kempe et al. (2014) list various ball possession and passing metrics which have been explored in the literature. Kempe et al. (2014) also propose aggregate metrics and relate these to success. However, a distinguishing feature of our work is that we consider a causal approach rather than one of association. This is made possible by the availability of player tracking data.

In Section 2, we introduce and motivate two definitions of pace. We contrast these definitions with alternative definitions that have been presented in the literature. In Section 3, we describe the player tracking dataset and discuss the challenges involved in pace calculations. One of the challenges is the determination of possession. In Section 4, we provide exploratory data analyses which provide various sporting insights on pace. The sporting insights are highlighted with the letters A-E. This section is also useful in identifying confounding variables that are related to pace. In Section 5, we present a causal analysis concerning the benefit of playing with pace. This involves the fitting of a MANOVA model which is the foundation for the determination of propensity scores and matching. The main result of this section is that playing with pace is a beneficial team strategy in soccer in terms of generating more shots. We conclude with a short discussion in Section 6.

## **2. DEFINITIONS OF PACE IN SOCCER**

Dan Blank's paperback on soccer (Blank 2002) provides 54 chapters on different tactics and advice on playing the game well. The first chapter which is titled the "Holy Grail" provides an inspiration for our investigation. In this chapter, Blank claims that playing fast is better than playing slow. In other words, Blank argues that teams should play with pace. Although the heuristic may be appealing, it does not seem that the belief has ever been corroborated against data. If the belief is true, then a measurable and sensible definition of pace may lead to important soccer insights.

First, we review some of the previous definitions of pace. In the original investigation of pace in hockey, Silva and Swartz (2018) were limited to the analysis of event data. With event data, a finite number of event types are recorded along with a timestamp. A shortcoming of the analysis is that the skating paths

between events (which are relevant to pace) are unknown. Consequently, Silva and Swartz (2018) only measured horizontal distances (i.e. down the length of the rink) during which possession was maintained. Furthermore, Silva and Swartz (2018) only evaluated pace for a game and did not differentiate pace of play between the two teams. Yu et al. (2019) used more extensive event data with events recorded approximately every second on average. With this data, they were able to define pace in various directions and considered zonal, league, team-level and player-level analyses. The pace metric defined by Yu et al. (2019) appears to be an average of velocities over event intervals and therefore differs conceptually from the Silva and Swartz (2018) definition which is based on total distance travelled. In soccer, Shen, Santo and Akande (2022) also used velocity as a pace measurement but restricted analyses to sequences where possession is retained over three or more events.

A commonality amongst all of the above pace analyses is that they were based on event data. With event data, distance calculations between events assume that the ball/puck travels in a straight line. Shen, Santo and Akande (2022) described the assumption as a major limitation. In this paper, the more detailed tracking data allows us to consider the actual paths where the ball travelled.

We begin with an analogy related to our definition of pace. We suggest that a painter is painting quickly (i.e. with pace) if they are able to apply a lot of paint on a canvas in a short period of time. In soccer, we view the brush strokes as the paths where players carry the ball and the paths where a ball is successfully passed. If a team is able to move the ball quickly, then they are playing with pace. The concept of possession is important; if a team is simply punting the ball downfield, in our view, they are not playing with pace. To operationalize these ideas, we consider the non-contiguous time intervals  $(t_1, t_2), \dots, (t_n, t_{n+1})$  in a match where a team has possession. During the possession interval  $i$ , the team moves distance  $d_i$ ,  $i = 1, \dots, n$ . Then, following the painting analogy, we refer to the team's *general pace* in the match as

$$GP = \frac{\sum_{i=1}^n d_i}{\sum_{i=1}^n (t_{i+1} - t_i)}. \quad (1)$$

Similarly, there is a corresponding pace formula (1) for the opponent. Note that the two teams will differ in the amount of time possession during the match. Therefore, the pace measure (1) is reflective of their style of play while in possession, and is insensitive to their total time of possession. Although the general pace metric (1) is defined in terms of a match, it can also be calculated for shorter periods of time (e.g. a half) or even for a single possession.

We contrast the general pace metric (1) with the quantity

$$P_{SSA} = \frac{1}{n} \sum_{i=1}^n \frac{d_i}{t_{i+1} - t_i} \quad (2)$$

which is related to the velocity concept of pace utilized by Shen, Santo and Akande (2022); note, however that Shen, Santo and Akande (2022) used medians rather than means. When comparing (1) with (2), we observe that (2) is sensitive to and is inflated by very fast passes (i.e. typically large  $d_i$  that occur over moderate time intervals  $t_{i+1} - t_i$ ). We believe that (1) better reflects pace as the totality of distance covered with respect to the cumulative time of possession.

We now introduce a variation of the general pace metric  $GP$  defined in (1). We note that there are differences in scoring intent based on the type of passes and dribbling. For example, the “tiki-taka” approach adopted by the Spanish national team in 2006 relied on many consecutive short passes that emphasized possession. Based on the metric  $GP$ , the tiki-taka approach would be characterized as a pacey style since passes typically have larger pace contributions than dribbling. But is tiki-taka pacey?

The aforementioned tiki-taka style allows one to reflect on stylistic differences between hockey and soccer. In hockey, the playing surface is smaller and players skate at great speeds. Therefore, it is more difficult to retain possession in hockey. Consequently, possession sequences tend to be of shorter duration than in soccer. To investigate pace in soccer with an emphasis on direct play, we modify  $GP$  and introduce *attacking pace*  $AP$  where the distances  $d_i$  now correspond to displacements down the field in the direction of the opposing goal. For example, passes back to the keeper (which have no attacking intent) do not positively contribute to attacking pace  $AP$ . Large positive contributions to the statistic  $AP$  will involve transitions such as the counter-attack.

To define  $AP$ , we refer to Figure 1 where an  $AP$  contribution is illustrated. In the plot, the “most attacking” pass that could possibly be made from point  $A$  would be to the middle of the opponent’s goal line  $C$ . This potential pass has associated distance  $d_{AC}$ . Instead, the pass was made from  $A$  to  $B$ , and the attacking distance from this new point  $B$  to  $C$  is denoted  $d_{BC}$ . Therefore, the contribution (in terms of attacking) from  $A$  to  $B$  is given by the residual distance

$$d = \begin{cases} d_{AC} - d_{BC} & d_{AC} \geq d_{BC} \\ 0 & d_{AC} < d_{BC} \end{cases} \quad (3)$$

In (3),  $d_{AC} - d_{BC}$  represents the reduction of the greatest attacking distance that was made due to the pass from  $A$  to  $B$ . Therefore, the new statistic  $AP$  has the

same form as  $GP$  in equation (1) where the  $d$  in equation (3) assumes a subscript  $i$  corresponding to the  $i$ th possession. We also note that the same type of calculation is carried out whether a possession involves passes or dribbles. It is important to note that the tracking data allows us to deal with path curvature when dribbling by breaking up dribbling sequences into small time intervals. If event data had been used (in contrast to tracking data), only the starting point and ending point of a dribbling sequence would be known.

A feature of the construction of the metric  $AP$  is illustrated through a possession sequence where the ball travels in a forward direction from  $A$  to  $B$  to  $C$  and where we denote the center of the goal by  $G$ . Using obvious notation for distances, the total attacking distance (3) is given by  $d = (d_{AG} - d_{BG}) + (d_{BG} - d_{CG}) = d_{AG} - d_{CG}$  which demonstrates that the metric is additive over the possession path.

Whereas the metrics  $GP$  and  $AP$  describe style of play while a team is in possession, insights may also be provided by considering the extent to which teams play a given style. For example, if a team is rarely in possession, then they are rarely executing their style. Therefore, we could also introduce the metric  $GP^*$  which is similar to  $GP$  except that we omit the denominator in (1). Therefore,  $GP^*$  may be thought of as total distance travelled by the team. Therefore, while  $GP$  is a statistic that describes pace during possession,  $GP^*$  takes possession into account such that teams with little possession are not playing with pace. Similarly, we could introduce the metric  $AP^*$  which is the total attacking distance during the match. However, for the remainder of our investigation, we only focus on the general pace statistic  $GP$  and the attacking pace statistic  $AP$ . Note that all of the proposed definitions of pace are properties of the possessing team.

### 3. DATA

For this investigation, we have a big data problem where both event data and player tracking data are available for 237 regular season matches (three matches missing) from the 2019 season of the Chinese Super League (CSL). The schedule is balanced where each of the 16 teams plays every opponent twice, once at home and once on the road.

Event data and tracking data were collected independently where event data consists of occurrences such as tackles and passes, and these are recorded along with auxiliary information whenever an “event” takes place. The events are manually recorded by technicians who view film. Both event data and tracking data have timestamps so that the two files can be compared for internal consistency. There are various ways in which tracking data are collected. One approach in-

volves the use of RFID technology where each player and the ball have tags that allow for the accurate tracking of objects. In the CSL dataset, tracking data are obtained from video and the use of optical recognition software. The tracking data consists of roughly one million rows per match measured on 7 variables where the data are recorded every 1/10th of a second. Each row corresponds to a particular player at a given instant in time. Although the inferences gained via our analyses are specific to the CSL, we suggest that the methods are applicable to any soccer league which collects tracking data.

### 3.1. Possession

A possession is defined as a period where a team has control of the ball. The event data is used to identify the possession sequences of a team. The event data contains all the events that occurred during a match and therefore tells us when possession sequences began and ended. Events where neither team is determined to have possession include injuries, cards, out-of-bounds, preliminary time to the beginning of set pieces (eg corners, throw-ins, free kicks, penalties, etc). Also, when determining possession sequences, we exclude time beyond 90 minutes since different matches have different amounts of added time. Among all the matches, there is an average of 373 possessions per match.

In Figure 2, we provide histograms (*GP* and *AP*) of the length of the possession sequences in metres. The histogram is right skewed. We observe a mean length of 46.6 (21.0) metres, minimum length 0.1 (0.0) metres, and maximum length 456.8 (82.2) metres corresponding to *GP* and *AP*, respectively.

### 3.2. The Pace Datasets

We pre-processed the CSL tracking and event data. Originally, the data were provided in xml files and we extracted the content using the `read_xml` function from the `XML` package using R software. The resulting tracking and event data were written into csv file format.

Ultimately, we constructed a pace dataframe for each match. This is a comprehensive dataset that allows us to investigate various questions of interest. The pace dataset is a matrix where the rows correspond to pace contributions made by an individual player during a possession. The columns consist of the following variables: start time of pace contribution, end time of pace contribution, the displacement  $d_i$  (both Euclidean distance and attacking distance (3)), match score at the beginning of the pace contribution, match score at the end of the pace contribution, the player who contributed to the pace contribution, the team of the player

who made the pace contribution, whether the contributing player plays for the home or road team, and the number of playing minutes during the match for the player. We note that Yu et al. (2018) shared the pace contribution equally between the player who made the pass and the player who received the pass. In our construction, we assign credit only to the player who made the pass.

To create the pace dataframe, we looped frame by frame through the tracking data, where we matched events and time using the event data. This permitted the calculation of the relevant distances during each possession. The process required approximately 15 minutes of computation for all 237 matches. Another challenge related to the calculation of AP involved slight differences in pitch size where the coordinates of point C in Figure 1 varied across pitches.

#### 4. EXPLORATORY DATA ANALYSES

A main objective of exploratory data analysis (EDA) is to reveal insights that can be more thoroughly investigated via modelling and inferential techniques. In this section, we use EDA to gain insights related to the pace statistics *GP* and *AP* together with other variables of interest. Below, EDA reveals five insights, labelled A-E.

In Figure 3, we produce scatterplots of *GP* and *AP* related to the home and road teams for all of the 237 available matches during the 2019 season of the CSL. We obtain a mean value of 0.66 (0.66) metres/sec, minimum value 0.42 (0.48) metres/sec and maximum value 0.78 (0.82) for the home and road team, respectively, using *GP*. We obtain a mean value of 0.25 (0.26) metres/sec, minimum value 0.14 (0.15) metres/sec and maximum value 0.38 (0.53) for the home and road team, respectively, using *AP*. Therefore, we observe that the pace statistics differentiating home and road teams are minor.

Initially, we were unsure whether pace was a property of the match (e.g. both teams play at high pace due to the particular style of the game) or whether each team has control of their respective pace. We observe that the sample correlation coefficients for *GP* and *AP* are 0.16 and 0.06, respectively. It is possible to carry out a test of correlation  $H_0 : \rho = 0$  in the two cases. The p-values are given by 0.014 and 0.358, for *GP* and *AP*, respectively. Although the first correlation is statistically significant, it is not strong in magnitude. The lack of strong correlations lead to the following insight.

**Insight A:** In a given match, each team has control of whether they play a pacey style. The pace of one team is not dictated by the pace of its opposition.

Next, we are interested in whether pace is a characteristic that can be attributed to teams. In Figure 4, we produce boxplots of the pace (*GP* and *AP*) for each of the 16 teams in the CSL where a single datapoint refers to the pace calculation in a match. We observe that there are only minor differences in the pace distributions across teams. Using a one-way ANOVA design for testing differences across teams, we obtain p-values of 0.0278 and 0.0106, for *GP* and *AP*, respectively. This leads to the following insight.

**Insight B:** Although some teams may play at slightly different average pace than other teams, such differences are small (particularly with *GP*). Pace is primarily a property of how a team plays in a particular match rather than a general property of the team.

Next, we are interested in whether pace depends on playing position. In Figure 5, we produce boxplots of the pace (*GP* and *AP*) for defenders, midfielders and forwards in the CSL where a single datapoint refers to the pace calculation in a match. We observe differences across the three positions. Due to the constraints of the field and positioning, it is logical that defenders have more open space in front of them than midfielders, and that midfielders have more open space in front of them than forwards. Therefore, it coincides with our intuition that pace should decrease according to defenders, midfielders and forwards, respectively. Using a one-way ANOVA design for testing differences across positions, we obtain highly significant test results with p-values of  $4.13e^{-10}$  and  $5.18e^{-6}$ , for *GP* and *AP*, respectively. This leads to the following broad insight.

**Insight C:** Defenders play at higher pace levels than midfielders who in turn play at higher pace levels than forwards.

Next, we are interested in whether pace is related to the time of the match. In Figure 6, we produce boxplots of the total pace by both teams (*GP* and *AP*) according to the time of the match broken into 15-minute intervals from 0 to 90 minutes. Although pace changes throughout the match, we observe different patterns according to *GP* and *AP*. With general pace *GP*, when a match begins, we expect that teams are alert and maintain defensive discipline. As the match continues, players tire, and they discontinue running with the same pace as before. At halftime, there is a rest period where teams recover slightly, and then they continue to tire during the second half.

With attacking pace *AP*, the interplay between exhaustion and defensive discipline is expressed differently. As the match continues, players tire and this allows

for more open space and the opportunity to seek gaps downfield. This causes a gradual increase in attacking pace with more pronounced increases in the latter stages. Using a one-way ANOVA design for testing differences across time intervals, we obtain highly significant test results with p-values of  $3.66e^{-8}$  and  $4.56e^{-5}$ , for *GP* and *AP*, respectively. This leads to the following insight.

**Insight D:** Teams plays at higher attacking pace as the match progresses.

Next, we are interested in whether pace is related to goal differential. In Figure 7, we produce boxplots of *GP* and *AP* corresponding to five goal differential categories as explained in the caption. The calculation of pace is taken over five-minute intervals for all teams and matches during the season. When a goal is scored during a five-minute interval, then the pace observation for that interval is excluded since the goal differential during the interval is not constant. With respect to the home team, we observe an interesting pattern with a slight increase in the median value of *AP* from  $GD = -2$  to  $GD = -1$ , from  $GD = -1$  to  $GD = 0$ , from  $GD = 0$  to  $GD = 1$ , followed by a drop in pace at  $GD = 2$ . Note that due to the home team advantage,  $GD = 2$  is a more common situation than  $GD = -2$ . Our nuanced intuition corresponding to these observations begins with the case  $GD = -2$  where the home team is losing badly. In this case, we expect that the road team is playing defensively as argued by Guan, Cao and Swartz (2002). The home team is therefore dominant in their offensive zone (i.e. near the road team's goal). On average, there is little room downfield for the home team and, consequently, they will be unable to make significant positive contributions to *AP*, and the *AP* measurement (as observed), will be low. As  $GD$  changes from -2 through 1, we would expect the road team to play less defensively, and as previously argued, *AP* will increase (as observed). However, a different behavioural mechanism occurs when  $GD = 2$ . In this case, the home team has a dominant lead. Their lead is so great, that they have little fear of losing. Hence, when  $GD = 2$ , the home team is not playing ultra-defensive (i.e. largely contained in their own zone, with predominantly long passes having a high *AP* contribution). Rather, when  $GD = 2$ , the home team is playing free, and this causes a reduction in *AP* from  $GD = 1$  to  $GD = 2$ . Using a one-way ANOVA design for testing differences in pace across goal differentials, we obtain significant test results with p-values of 0.041 and 0.028 for *GP* and *AP*, respectively. This leads to the following insight.

**Insight E:** Teams play at different pace levels depending on the goal differential.



## 5. CAUSAL ANALYSIS

In this section, we return to our primary question whether it is advantageous to play with pace. With a sensible definition of pace and the availability of tracking data, the issue can be addressed.

Recall that questions of cause and effect are traditionally addressed using randomization in experimental contexts. For our problem, this would require the random assignment of pace to the two teams. Of course, matches are not experiments, but rather observational studies where randomization does not occur. Therefore, we address cause and effect through methods of causal inference (Pearl 2009). Although causal inference has received great attention, the methods are often difficult to implement due to the necessity of specifying and measuring relevant confounding variables. Fortunately, sport is much simpler in its objectives than many other scientific domains, and via the spatio-temporal tracking data and the EDA investigations of Section 2, confounding variables are accessible. Therefore, together with some novel ideas, and referring to the approach introduced in Wu et al. (2021), we are able to address cause and effect associated with pace.

### 5.1. Propensity Scores

Using causal terminology, we think of pace as the treatment which we denote  $X_h$  and  $X_r$ , corresponding to the home and road teams, respectively. We denote  $W$  as the vector of confounding variables which we believe are predictive of the pace  $X = (X_h, X_r)'$ . With this structure, we wish to specify propensity scores  $\text{Prob}(X_h - X_r > 0 \mid W)$  that describe the probability that the home team plays at greater pace than the road team given the relevant circumstances of the match. With insights gained from the EDA of Section 2, we specify a statistical model that leads to propensity scores. For reference, we define all of our relevant variables below.

$$\begin{aligned}
 t &\equiv \text{time of the match in minutes, } t \in (0, 90) \\
 X(t) &\equiv \text{pace vector for home and road teams at time } t; \text{ either } GP \text{ or } AP \\
 GD(t) &\equiv \text{goal differential in favour of the home team at time } t \\
 O &\equiv \text{pre-match betting odds corresponding to the home team} \\
 Y(t_1, t_2) &\equiv \text{excess shots by home team relative to road team during } (t_1, t_2)
 \end{aligned} \tag{4}$$

Looking ahead, our interest is in determining a cause-effect relationship regarding the impact of pace  $X$  on success  $Y$ . A natural success variable would be goals. However, in soccer, goals are rare events with less than three goals per game on average in top professional leagues. We therefore use the surrogate variable shots as defined in (4) to assess success. Of course, not all shots lead to goals

but shots are an indication of success. However, let's return to the first step of the causal investigation which involves the construction of a propensity score model.

We first bin the data to define levels for each of the three confounding variables  $W(t) = (t, GD(t), O)'$ . We segment the time  $t$  into 18 five-minute intervals:  $(0, 5)$ ,  $(5, 10)$ ,  $\dots$ ,  $(85, 90)$ . We do not include added time beyond 90 minutes since the amount of added time differs across matches.

For the second variable, we restrict  $GD(t)$  to five states with goal differentials  $-2$ ,  $-1$ ,  $0$ ,  $1$  and  $2$  corresponding to the home team at time  $t$ . Note that  $GD(t) = -2$  corresponds to the home team losing by two or more goals and that  $GD(t) = 2$  corresponds to the home team winning by two or more goals. For a given match, we consider each of the 18 time intervals, and if the goal differential is constant throughout the interval (either  $-2$ ,  $-1$ ,  $0$ ,  $1$  or  $2$ ), then an observation is recorded.

For the third variable  $O$ , we access pre-match betting odds available from the website <https://www.oddsportal.com/soccer/china/super-league-2019/results/>. The betting odds (reported in decimal format) provide us with the relative strength of the two teams. Ignoring the vigorish imposed by the bookmaker, the interpretation of betting odds  $o$  for a team is that the team has a pre-match probability  $1/o$  of winning the match. Therefore, values of  $o$  slightly greater than  $1.0$  indicate a strong favourite whereas large values of  $o$  indicate an *underdog*. For a given match, we define four bins for the decimal odds of the home team:  $[1.3, 1.7)$ ,  $[1.7, 2.3)$ ,  $[2.3, 3.0)$  and  $[3.0, 8.0)$ . The odds are restricted so that only competitive matches are included, and the endpoints are selected to provide comparable numbers of observations across bins. Note that the betting odds  $O$  do not depend on the time  $t$ .

The variable  $O$  was obtained using the standard three-way betting odds for soccer corresponding to home wins, draws and losses. Ideally, relative strength would be better measured with *moneyline* odds corresponding to wins where wagers corresponding to draws are refunded. The reason why three-way betting odds are not ideal is that two matches can have identical win odds yet different draw and loss odds. However, the difference in odds in these two situations is typically minor.

For the response variable in the propensity score model, we calculate  $X_h(t)$  and  $X_r(t)$  during the time interval which is intended to convey the style of pace over the time period. We use attacking pace  $AP$  for the pace calculation, as it is more definitive and perhaps more interesting than general pace  $GP$ . We also emphasize that the response variable  $X(t) = (X_h(t), X_r(t))'$  is bivariate which makes the causal investigation nonstandard.

To illustrate the variables in the propensity score model, consider a match where the score is 2-0 just prior to the 70-th minute. Following conventional notation where the first team in the scoreline is the home team, this implies that the home team is leading by two goals. Assume further that the home team is the favoured team with pre-match decimal betting odds  $o = 1.5$ . In this match, suppose that neither team scores during the time interval (65, 70) minutes, and that the *AP* statistics for the home and road teams during this period are 2.34 and 2.07, respectively. Then, for this time interval, we have the observed response  $X = (2.34, 2.07)$  and covariates  $W = (14, 2, 1)$  where  $t \in (65, 70)$  corresponds to the 14th time category,  $GD = 2$  is the goal differential (categorical), and odds  $o = 1.5 \in (1.3, 1.7)$  corresponds to the first category.

Based on the above considerations, we have 3679 observations recorded across  $18 \times 5 \times 4 = 360$  cells. For linear models based on categorical data, it is prudent to have adequate numbers of observations in each cell. For this reason, we consider a reduction in the number of cells by instead defining six time categories, (0,15),(15,30),..., (75,90) minutes. In this case, we have 944 observations recorded across  $6 \times 5 \times 4 = 120$  cells. The cell counts are provided in Table 1. In most cells, we have the recommended minimum number of five counts per cell; exceptions tend to occur with large goal differentials (i.e.  $GD = -2$  and  $GD = 2$ ), especially early in matches.

<i>O</i>	<i>GD</i> = -2				<i>GD</i> = -1				<i>GD</i> = 0			
	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)
$t \in (00, 15)$	0	0	0	2	0	0	0	3	48	51	33	37
$t \in (15, 30)$	0	0	0	3	6	3	2	8	32	32	19	24
$t \in (30, 45)$	1	0	0	3	6	3	4	5	24	24	17	20
$t \in (45, 60)$	0	0	2	8	5	5	6	7	18	18	11	16
$t \in (60, 75)$	1	1	3	13	3	4	3	12	15	15	7	7
$t \in (75, 90)$	1	3	4	7	3	4	4	11	9	16	6	8

<i>O</i>	<i>GD</i> = 1				<i>GD</i> = 2			
	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)	[1.3,1.7)	[1.7,2.3)	[2.3,3)	[3,8)
$t \in (00, 15)$	6	2	1	2	0	0	0	0
$t \in (15, 30)$	14	6	5	7	2	1	0	0
$t \in (30, 45)$	17	9	7	10	6	2	1	0
$t \in (45, 60)$	18	14	5	8	12	1	2	2
$t \in (60, 75)$	12	16	8	5	14	5	4	3
$t \in (75, 90)$	13	9	6	4	16	3	3	1

**Table 1: Cell counts for the  $6 \times 5 \times 4$  covariate categories where the categories correspond to the time  $t$ , the goal differential  $GD$  and the betting odds  $O$ .**

Our propensity score model is a multivariate analysis of variance (MANOVA)

model where the response variable  $X$  is two-dimensional and the covariate  $W = (t, GD, O)$  has  $6 \times 5 \times 4 = 120$  cells (as described above). The MANOVA model is preferred to two separate ANOVA models for  $X_h$  and  $X_r$  since the MANOVA model permits a covariance structure between  $X_h$  and  $X_r$ . Details on MANOVA models are given by Smith, Gnanadesikan and Hughes (1962).

We used MANOVA software using the `manova` function in the `Stats R` package. One of the assumptions of MANOVA concerns the normality of observations. A quantile plot of the residuals does not suggest any serious departures from normality. In Table 2, we present the results of fitting the MANOVA model where we have allowed for the possibility of first-order interaction terms. The main take-away is that the time of the match  $t$ , the goal differential in favour of the home team  $GD$  and the relative strength of the home team  $O$  are strongly associated with attacking pace  $X$ . There is also mild evidence of some first-order interactions involving  $t$ ,  $GD$  and  $O$ .

Variable	Df	Pillai	approx F	num Df	den Df	Pr(> F)
$t$	5	0.062411	5.7077	10	1772	1.801e-08 ***
$GD$	4	0.075761	8.7208	8	1772	9.575e-12 ***
$O$	3	0.126651	19.9665	6	1772	<2.2e-16 ***
$t*GD$	18	0.050637	1.2786	36	1772	0.12531
$t*O$	12	0.054758	2.0784	24	1772	0.00164 **
$GD*O$	15	0.035338	1.0624	30	1772	0.37509
Error	886					

**Table 2: Results from the MANOVA which relates pace  $X$  to the covariates  $W = (t, GD, O)$ .**

Finally, we need to induce the required probability  $\text{Prob}(X_h - X_r > 0 \mid W)$  from the fitted MANOVA model. The calculation is based on a simple result from mathematical statistics using properties of the normal distribution. For example, for a given match situation  $W$ , suppose that the MANOVA model yields  $X \sim \text{Normal}_2(\mu, \Sigma)$  where  $\mu = (\mu_1, \mu_2)'$  and  $\Sigma = (\sigma_{ij})$ . Then  $\text{Prob}(X_h - X_r > 0 \mid W) = \Phi((\mu_1 - \mu_2) / \sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}})$  where  $\Phi$  is the cumulative distribution function of the standard normal distribution. Note that the bivariate normal parameters are estimated through the fitting of the MANOVA model. For example, the estimated values of  $\sigma_{11}$ ,  $\sigma_{22}$  and  $\sigma_{12}$  are 0.0051, 0.0061 and 0.0003, respectively. Therefore, the estimated correlation between home and road attacking pace is  $\sigma_{12} / (\sqrt{\sigma_{11}\sigma_{22}}) = 0.054$  which indicates that the MANOVA formulation

(which takes into account the relationship between home and road pace) provides only a slight improvement over ANOVA.

## 5.2. Matching and Results

In the most basic randomized experiment, an experimenter randomly assigns  $M$  subjects from a population to receive a treatment and  $M$  subjects from the population to receive the control. The hope is that through random assignment, the treatment group will on average be similar to the control group, and that differences in the response between the two groups can be attributed to the treatment.

The use of propensity scores and matching (Austin 2011, Imbens 2004) attempts to mimic the basic randomized experiment in the context of observational studies. A propensity score for a subject in a clinical trial is the probability that the subject receives the treatment. In the pace problem,  $\text{Prob}(X_h - X_r > 0 | W)$  is the estimated probability that the home team will play at a higher pace than the road team. Therefore,  $\text{Prob}(X_h - X_r > 0 | W)$  serves as the relevant propensity score in the pace application.

In our problem, we have a dataset involving 944 pace observations (see Section 5.1) resulting in  $M_1 = 450$  cases where the home team plays at greater pace (the treatment) and  $M_2 = 494$  cases where the home team plays at lesser pace (the control). Since  $M_1 < M_2$ , the matching idea is that we attempt to match each of the  $M_1$  treatment cases with a corresponding control case so that each pair has a similar estimated propensity score based on the underlying match circumstances  $W$ . Then the resulting two groups ( $M_1$  treatments and  $M_2$  controls) will be similar in the match characteristics, and that differences between the two groups can be attributed to the treatment (i.e. pace).

There are many ways that the matching of propensity scores can be carried out (Stuart 2010), and caution ought to be exercised in the process. In our application, we begin with the  $M_1$  cases where the home team plays at a greater pace, and we use a nearest neighbour method for selecting the matched cases where the home team plays at a lesser pace. Specifically, we use the *Matching* package (Sekhon 2011) in the statistical programming language R to randomly select (with replacement) control cases that fall within a specified tolerance of the propensity scores for the treatment cases. Sampling with replacement tends to increase the quality of matching when compared to sampling without replacement. Unlike deterministic matching procedures, the random aspect of the nearest neighbour procedure allows us to repeat analyses to check the sensitivity of the inferences.

Following the implementation of the matching procedure, Figure 8 displays

the balance between the two groups with respect to the propensity scores. The similarity in the histograms is important as it provides confidence that the two groups are similar according to the characteristics that affect whether the home team plays at greater pace.

The inferential component of the investigation begins with a simple paired two-sample test between the two groups based on the response  $Y$  (excess shots by the home team) as described in (4). Again, we prefer to use shots rather than goals since goals are rare events. The quantity of interest is the average treatment effect  $ATE = \bar{Y}(1) - \bar{Y}(0)$  where  $\bar{Y}(1)$  is the excess number of resultant shots by the home team when they are playing at greater pace, and  $\bar{Y}(0)$  is the excess number of resultant shots by the home team when they are playing at lesser pace. We obtain  $ATE = 0.73 - 0.41 = 0.32$  with standard error 0.103. The result is significant and suggests that pace is beneficial in the sense of playing at a higher attacking pace.

To put the above result into context, suppose that the home team outpaces the road team during all six 15-minute intervals during the match. Then, we would expect the home team to have roughly  $6(0.32) = 2$  more shots during the match than the road team. Note also that we have been careful to distinguish the home and road teams. If we flipped the analysis to consider the average treatment effect due to the road team playing at pace, we would obtain  $ATE = -0.41 - (-0.73) = 0.32$ . Therefore, the benefit of outpacing the opposition applies to either team.

In Figure 9, we present a more nuanced view of the situation. For each group (treatment and control), we smooth the variable  $Y$  with respect to the propensity score. We observe that as the propensity score increases (i.e. conditions become more favourable for the home team to play at greater pace), the excess shots for the home team increases for both groups. We also observe that the excess shots by the home team remains relatively constant across the two groups as the propensity score increases. In practice, this means that the advantage of playing at pace persists no matter the circumstances that dictate whether a team should play at pace.

Therefore, the takeaway message is that playing with pace is a good strategy. It leads to more shots for than against. This provides support to Blank's thesis - the Holy Grail of tactics (in Chapter 1 of Blank (2012)) that fast is better than slow.

## 6. DISCUSSION

Despite its importance, style of play is an understudied aspect in team sport. In this paper, we investigate pace of play as it relates to soccer. Although the analyses were restricted to the study of tracking data in the Chinese Super League, it is conjectured that the broad results hold true for other high-level professional soccer leagues.

In particular, we found that teams that play at higher attacking pace are more advantaged in producing shots than teams that play at lower pace. For a team that outpaces its opponent throughout a match, this translates to roughly two extra shots. The conclusion was facilitated through the adaptation of causal methods. In particular, we sought confounding variables that were important in determining propensity scores. Furthermore, the propensity scores were obtained by reducing a bivariate normal distribution to a relevant Bernoulli distribution. The EDA produced additional sporting insights (A-E) related to pace.

There are possible future investigations related to pace of play. For example, we believe that similar analyses may be carried out in other invasion sports where tracking data are available. Also, it may be interesting to analyse pace separately in terms of passing and dribbling. The ball generally moves more quickly when passing, and there may be stylistic differences between teams in terms of how much they pass relative to how much they dribble.

A limitation in our work is that the response variable  $Y$  (shots) in the causal analysis correspond to rare events and is known to be noisy. A better response variable may be expected goals (Spearman 2018, Anzer and Bauer 2021), and this could be considered in future investigations. Another limitation of our work is the restriction to matches from the CSL. It would be good to see if the results also hold in top-level European leagues where the best players from all over the world compete. Although we argue that confounding variables can be identified with tracking data, for sure, there are latent variables that we have not utilized (e.g. level of player fatigue). This is also a limitation.

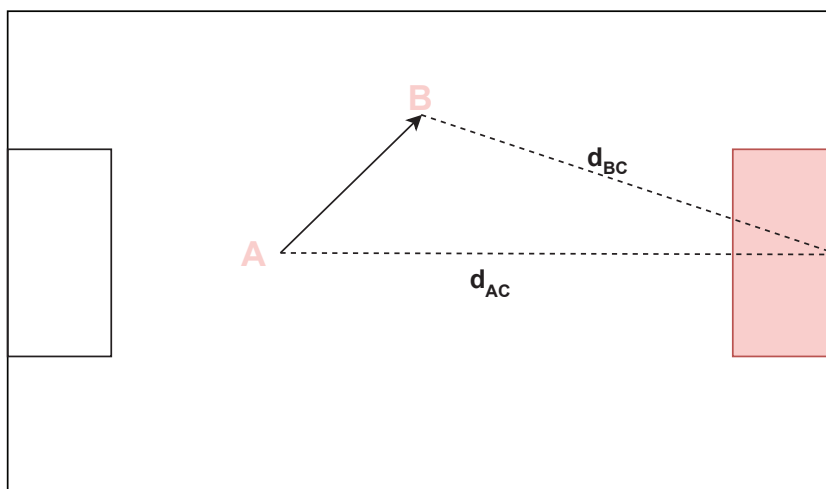
## REFERENCES

- Anzer, G. and Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). In *Frontiers in Sport and Active Living*, 3, Article 624475.
- Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. In *Multivariate Behavioral Research*, 46, 399-424.

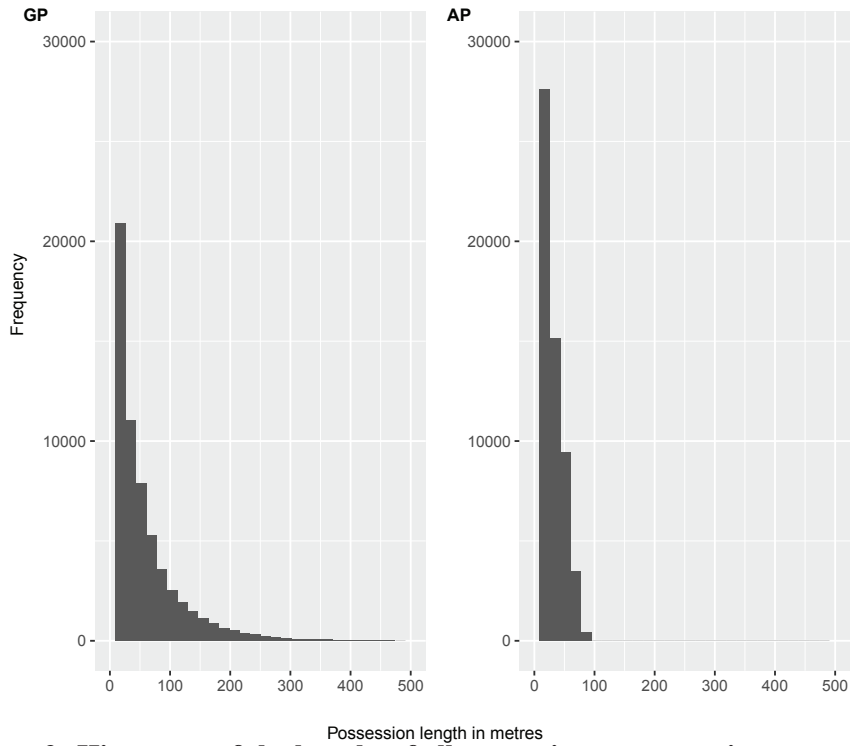
- Blank, D. (2012). *Soccer IQ*, [www.soccerpoet.com](http://www.soccerpoet.com)
- de Jong, L.M.S., Gastin, P.B., Angelova, M., Bruce, L. and Dwyer, B. (2020). Technical determinants of success in professional women's soccer: A wider range of variables reveals new insights. In *PLOS One*, 15(10), e0240992.
- Goes, F.R., Brink, M.S., Elferink-Gemser, M.T., Kempe, M. and Lemmink, K.A.P.M. (2021). The tactics of successful attacks in professional association football: Large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. In *Journal of Sports Sciences*, 39(5), 523-532.
- Guan, T., Cao, J. and Swartz, T.B. (2022). Should you park the bus? Manuscript under review. Available at <https://www.sfu.ca/~tswartz/>
- Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. In *ACM Computing Surveys*, 50(2), Article 22.
- Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. In *The Review of Economics and Statistics*, 86, 4-29.
- Kempe, M., Vogelbein, M., Memmert, D. and Nopp, S. (2014). Possession vs. direct play: Evaluating tactical behavior in elite soccer. In *International Journal of Sports Science*, 4(6A), 35-41.
- Lepschy, H., Wäsche, H. and Woll, A. (2021). Success factors in football: an analysis of the German Bundesliga. In *International Journal of Performance Analysis in Sport*, 20(2), 150-164.
- McLellan, I. (2010). Total football: Whatever happened to the beautiful game? *Bleacher Report: World Football*, Accessed June 2, 2022 at <https://bleacherreport.com/articles/321814-beautiful-game-what-ever-happened-to-total-football>
- Merlin, M., Cunha, S.A., Moura, F.A., Torres, R., Gonçalves, B. and Sampaio, J. (2020). Exploring the determinants of success in different clusters of ball possession sequences in soccer. In *Research in Sports Medicine*, 28(3), 339-350.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference, Second Edition*. Cambridge University Press: Cambridge.
- Sekhon, J.S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. In *Journal of Statistical Software*, 42, 1-52.
- Shaw, L. and Glickman, M. (2019). Dynamic analysis of team strategy in professional football. In *Barça Sports Analytics Summit*.
- Shen, E., Santo, S. and Akande, O. (2022). Analyzing pace-of-play in soccer using spatio-temporal event data. In *Journal of Sports Analytics*, 8(2), 127-139.



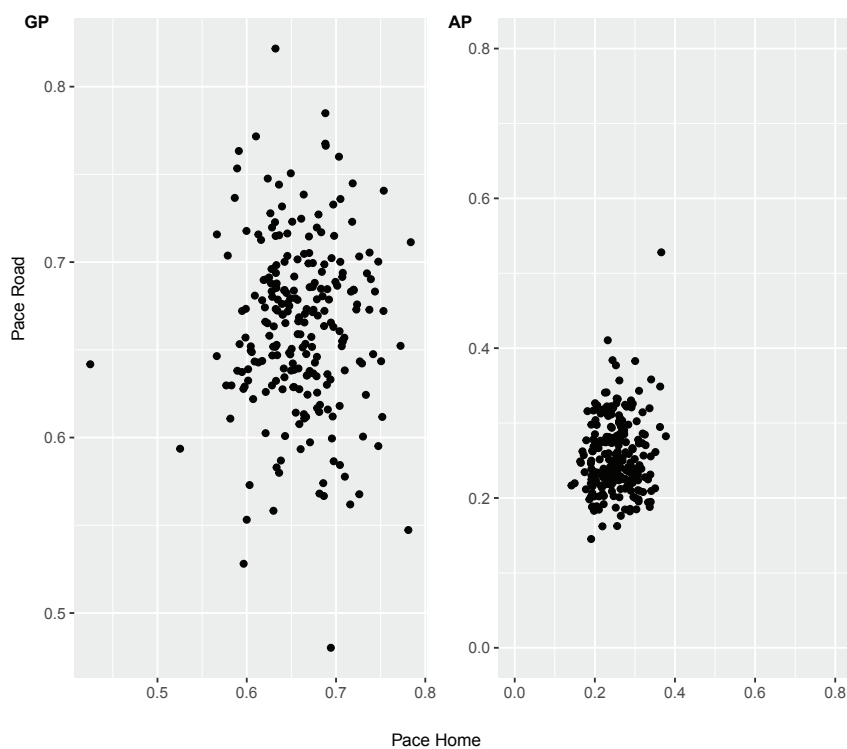
- Silva, R., Davis, J. and Swartz, T.B. (2018). The evaluation of pace of play in hockey. *In Journal of Sports Analytics*, 4, 145-151.
- Smith, H., Gnanadesikan, R. and Hughes, J.B. (1962). Multivariate analysis of variance (MANOVA). *In Biometrics*, 18(1), 22-41.
- Spearman, W. (2018). Beyond expected goals. *Proceedings of the 2018 MIT Sloan Sports Analytics Conference*, Accessed October 6, 2022 at [https://www.researchgate.net/publication/327139841\\_Beyond\\_Expected\\_Goals](https://www.researchgate.net/publication/327139841_Beyond_Expected_Goals)
- Tweeddale, A. (2022). Counter-pressing and the gegenpress: Football tactics explained. *The Coaches Voice: Coaching Knowledge*, Accessed June 2, 2022 at <https://www.coachesvoice.com/cv/counter-pressing-gegenpressing-football-tactics-explained-klopp-guardiola-bielsa-hasenhuttl/>
- Stuart, E.A. (2010). Matching methods for causal inference. A review and a look forward. *In Statistical Science*, 25(1), 1-21.
- Wilson, J. (2013). *Inverting the Pyramid*, Nation Books, New York.
- Wu, Y., Danielson, A., Hu, J. and Swartz, T.B. (2021). A contextual analysis of crossing the ball in soccer. *In Journal of Quantitative Analysis in Sports*, 17(1), 57-66.
- Yu, D., Boucher, C., Bornn, L. and Javan, M. (2019). Evaluating team-level pace of play in hockey using spatio-temporal possession data. *Proceedings of the 2019 MIT Sloan Sports Analytics Conference*, accessed October 18, 2021 at <https://arxiv.org/pdf/1902.02020.pdf>



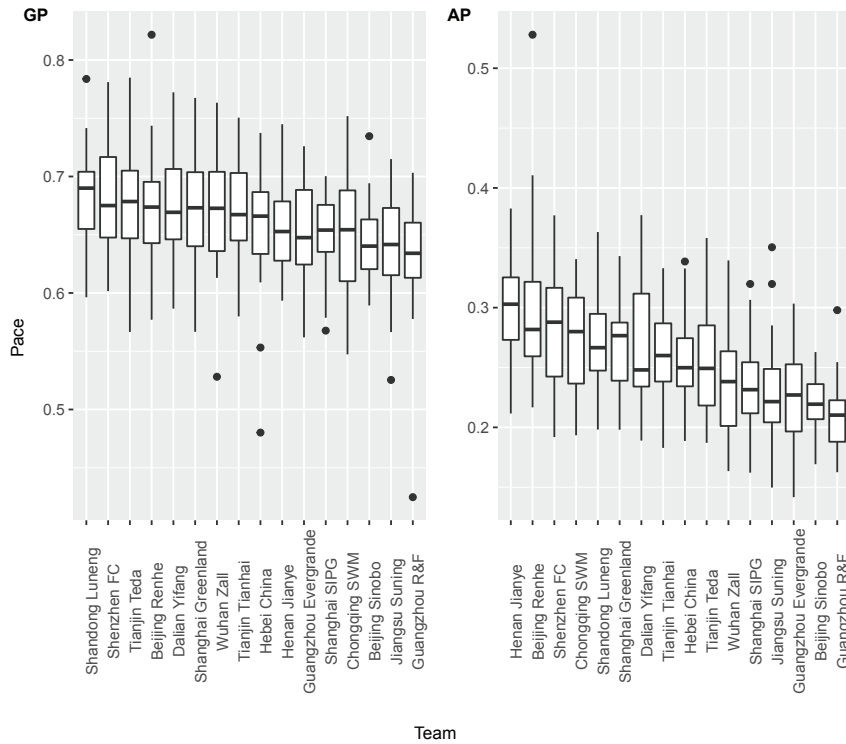
**Figure 1:** The plot illustrates a pass with attacking intent. A is the starting point of the pass, B is the end point of the pass and C denotes the middle of the goal line of the opponent. The values  $d_{BC}$  and  $d_{AC}$  represent the distances from B to C, and A to C, respectively. Attacking pace  $AP$  for this component of play is obtained using the distance  $d_{AC} - d_{BC}$ .



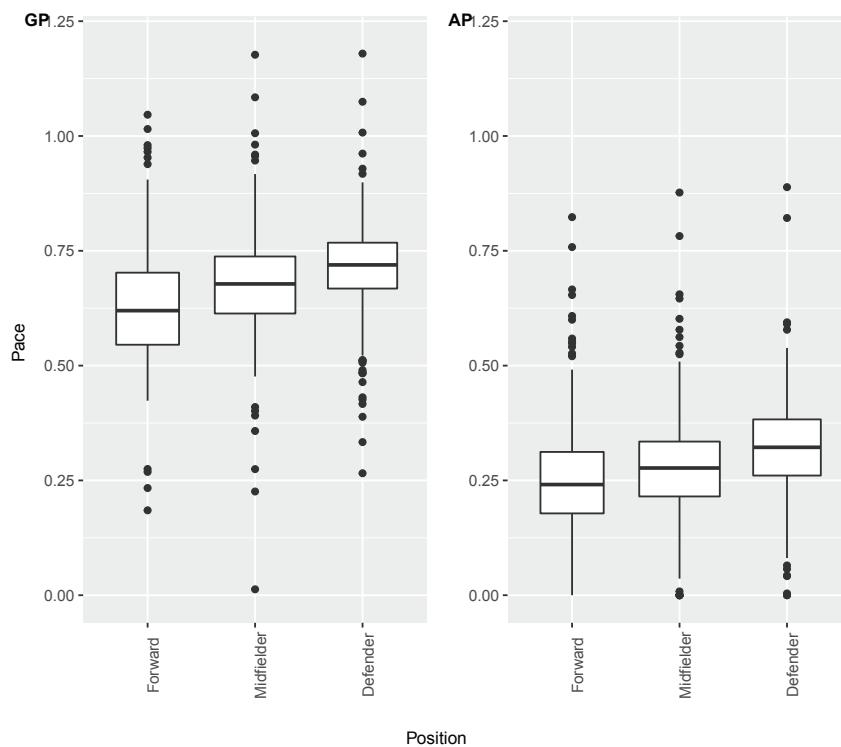
**Figure 2: Histogram of the lengths of all possession sequences in metres corresponding to *GP* and *AP*, respectively.**



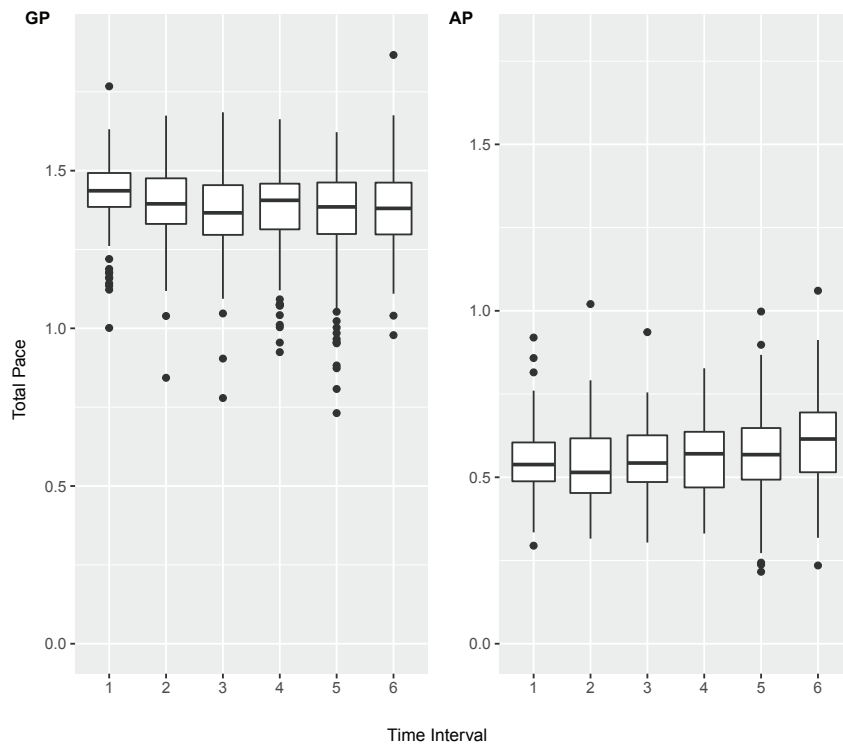
**Figure 3: Scatterplots for *GP* and *AP* related to the home and road teams for each match.**



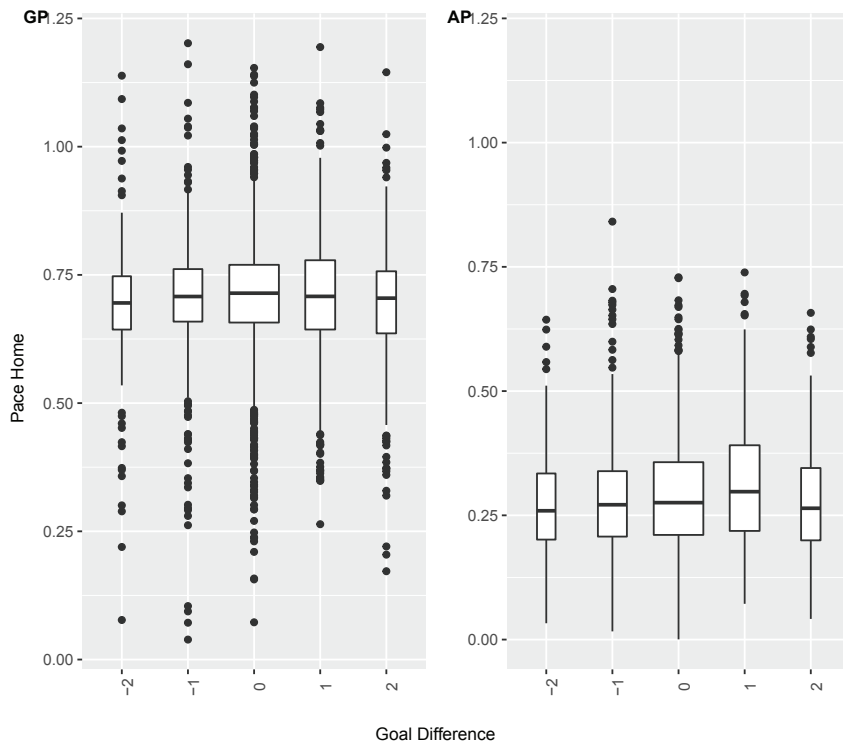
**Figure 4: Boxplots of GP and AP for each of the 16 teams in the CSL based on their 30 matches.**



**Figure 5: Boxplots of *GP* and *AP* for forwards, midfielders and defenders based on their individual match statistics.**

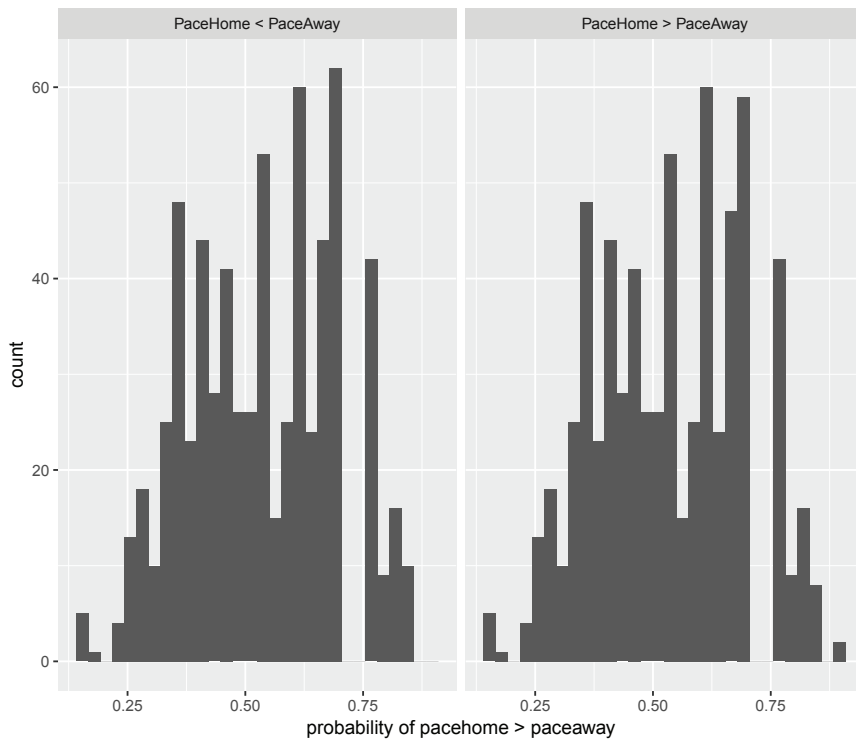


**Figure 6: Boxplots of *GP* and *AP* according to the time of the match where time is divided into six 15-minute intervals from 0 to 90 minutes. The pace calculations are the total pace corresponding to both teams.**

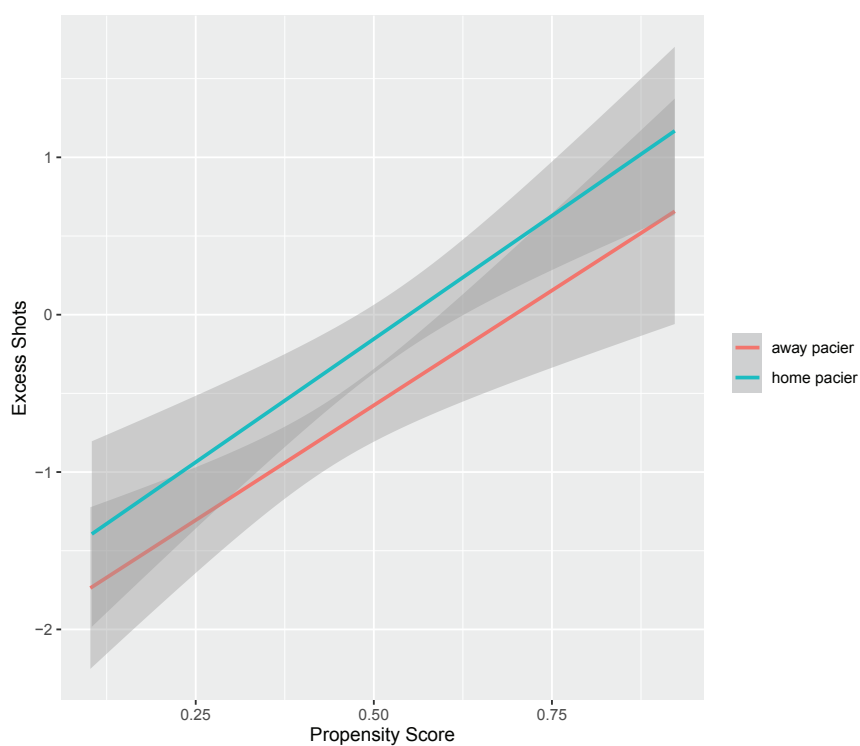


**Figure 7: Boxplots of  $GP$  and  $AP$  corresponding to the goal differential (GD) taken at 5-minute intervals where -2 corresponds to the home team losing by 2 or more goals, -1 corresponds to the home team losing by 1 goal, 0 indicates a tied match, 1 corresponds to the home team winning by 1 goal and 2 corresponds to the home team winning by 2 or more goals.**





**Figure 8: After matching, histograms of the two groups (treatment and control) are depicted where the horizontal variable is the propensity score.**



**Figure 9:** After matching, smoothed plots of the excess shot variable  $Y$  for the home team with respect to the propensity score under the treatment (blue) and the control (red).