Annals of Operations Research (2023) 325:1–7
https://doi.org/10.1007/s10479-023-05344-z

EDITORIAL

Editorial: Big data and data science in sport

Pierpaolo D'Urso[1] · Livia De Giovanni[2] · Tim Swartz[3]

Published online: 25 April 2023

Data and big data in sports are being collected and analyzed, with the integration of physical and digital sources, increasing the knowledge of professional sports for all parties involved.

The statistical analysis of data in sports can enhance decision making capabilities related to the performance of players or teams, health and safety of players, fan engagement, marketing strategies, revenues, economics of sport, sport practice, and well-being. The sources of data include private and public institutions, Internet of Things, and social networks. Different statistical learning and operations research methods have been proposed, depending on the type of sport, the data, and the goals of the analysis. Data are collected during training and matches to gain insights into the factors that impact the success of players and teams, including fairness of competition, player evaluation, scheduling, tactics, identification of key performance indicators, drafting, rules, and ranking. Data from fans influence marketing efforts and decisions about sporting events, variable price ticketing, and fan retention. Data improve the effort spent on forecasting outcomes of sporting events and the factors which are assumed to contribute to those outcomes. Data collected in sports can be used to manage the efficiency of gambling markets.

The contributions collected in the special issue have been divided by sport and, where possible, by statistical methodology within each sport.

✉ Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Livia De Giovanni
ldegiovanni@luiss.it

Tim Swartz
tswartz@sfu.ca

[1] Sapienza University of Rome, Rome, Italy

[2] Luiss University, Rome, Italy

[3] Simon Fraser University, Burnaby, Canada

1 Football

In a first group of papers *clustering* methods are used.

In the paper "A robust method for clustering football players with mixed attributes", P. D'Urso, L. De Giovanni, and V. Vitale develop a robust fuzzy clustering model for mixed data. For each variable, or attribute, a dissimilarity measure is proposed, and the clustering procedure combines the dissimilarity matrices with weights objectively computed during the optimization process. The weights reflect the relevance of each attribute type in the clustering results. The model is used to cluster football players with respect to mixed data on performances.

In the paper "Clustering of variables methods and measurement models for soccer players' performances", M. Carpita, P. Pasca, S. Arima, and E. Ciavolino investigate the ability of various composite indicators to define a measurement structure for global soccer performance. The theoretical soccer performance dimensions are based on a set of 29 players' attributes periodically produced by Electronics Arts (EA) Sports experts. The players' performance attributes or variables are considered and processed with three different techniques: Cluster of variables around Latent Variables (CLV), Principal Covariates Regression (PCovR) and Bayesian Model-Based Clustering (B-MBC), and the resulting clusters have been embedded into structural equation models with Partial Least Squares (PLS-SEMs) with a Higher-Order Component (that is, the overall soccer performance). Results show the validity of composite indicators.

In the paper "Community detection in attributed networks for global by transfer market", G. P. Clemente and A. Cornaro propose a community detection methodology in the framework of complex networks. Countries are clustered according to similarities in their roles in the transfer market of football players and to the presence of indirect connections due to common neighbours. Numerical results show a relationship between the composition of clusters and the economic value of the football leagues of different countries. In particular, various European countries are identified which represent top leagues in the soccer market and are also involved as central countries in economic trades.

In the paper "A rank-size approach to analyse soccer competitions and teams: the case of the Italian football league 'Serie A'", V. Ficcadenti, R. Cerqueti, and C. Hosseini Varde'i introduce a data-analysis rank-size approach to assess the features of soccer competitions and competitors. A cluster analysis of the estimated rank-size law parameters based on a k-means algorithm is developed to provide additional insights and capture similarities and deviations among championships and teams. The championship rankings and the teams' final scores in the most relevant Italian league, the "Serie A," between 1930 and 2020 are analysed to explore the presence of rank-size regimes in the various yearly championships. In a second group of papers *forecasting* methods are used.

In the paper "Forecasting binary outcomes in soccer", R. Mattera studies the prediction of binary outcomes of football matches by using the generalized autoregressive score (GAS). The binary events are: the presence of a red card, the scoring of at least one goal by each team, and the scoring of a certain number of total goals. The model is applied to the English Premier League and the Italian Serie A.

In the paper "Betting market efficiency and prediction in binary choice models", R.H. Koning and R. Zijm introduce a method to derive winning probabilities from betting odds. The method allows for residual favourite-longshot bias. The approach allows for incorpo-

ration of match specific variables that may determine the relationship between the actual probability of the outcome and the implied winning probabilities. The method has been used for the estimation of the actual probability of outcomes in the English Premier League and in the Spanish La Liga.

In the paper "Influence of Red and Yellow cards on team performance in elite soccer", L. Badiella, P. Puig, C. Lago-Peñas, and M. Casals analyse the effects of red and yellow cards on the scoring rate in elite soccer. All events were structured in 5-minute intervals and were analyzed by means of a Generalized Linear Mixed Model with a Poisson distribution, considering the presence of correlated data, where the dependent variable is the scoring rate. The sample was composed of 1,826 matches in the top five European leagues.
The final group of football papers are described.

In the paper "Football tracking data: a copula-based hidden Markov model for classification of tactics in football", M. Ötting and D. Karlis propose a data-driven approach for automated classification of tactics in football. A copula-based HMM model considers the effective playing space of both teams, obtained by high-frequency tracking data on the positions of players, to account for the competitive nature of the game. The model provides an estimate of a team's playing style and tactics.

In the paper "Service quality in football tourism: an evaluation model based on online reviews and data envelopment analysis with linguistic distribution assessments", A.P. Darko, D. Liang, Y. Zhang, and A. Kobina develop a decision support model to investigate the satisfaction of sports tourists. The proposed model employs text mining techniques to discover service quality attributes from text reviews and reveals the sentiment polarities of the text reviews. Furthermore, a bestpoint slack-based measure (BP-SBM) handling both positive and negative data values is proposed to compute the degree of tourist satisfaction and benchmarking goals.

In the paper "An extension of correspondence analysis based on the multiple Taguchi's index to evaluate the relationships between three categorical variables graphically: an application to the Italian football championship", A. D'Ambra and P. Amenta propose the use of three-way correspondence analysis with ordinal categorical variables to evaluate the relationships between the rankings of the Italian football "Serie A" championship of the last 10 seasons and a set of two factors defined by average percentage of ball possession and number of tags for each team. They introduce a multiple extension of Taguchi's index as an alternative to Pearson's statistic for ordered contingency tables.

In the paper "Does luck play a role in the determination of the rank positions in football leagues? A study of Europe's 'big five'", S. Sarkar and S. Kamath provide an empirical study to assess the role of luck in the determination of rank positions in football leagues. ANOVA analyses are provided on data from seven seasons (2014–15 to 2020–21) in the top tier leagues of England, Spain, Germany, Italy, and France. The X-factor effect on performance is defined as the difference between actual and expected goals.

2 Cricket

In the paper "Understanding the effect of contextual factors and decision making on team performance in Twenty20 cricket: an interpretable machine learning approach", P. Puram, S. Roy, D. Srivastav, and A. Gurumurthy propose tree-based machine learning (ML) models

such as gradient boosting, regression trees, bagging, random forests, and Bayesian additive regression trees (BART) to determine the effect of contextual factors and subsequent decisions taken on team performance in Twenty20 (T20) cricket, consisting of nine seasons of the Indian Premier League involving 563 matches.

In the paper "Optimization of team selection in fantasy cricket: a hybrid approach using recursive feature elimination and genetic algorithm", A. Jha, A. Kumar Kar, and Agam Gupta develop a two-step methodology for player assessment and team selection in fantasy cricket. Player assessment is carried out using recursive feature elimination with random forests, in which context relevant player metrics are considered and the selection of players is based on a modified genetic algorithm. The efficacy of the proposed method is assessed on Dream11, a popular fantasy sports application.

In the paper "Best strategy to win a match: an analytical approach using hybrid machine learning-clustering-association rule framework", P.R. Srivastava, P. Eachempati, A. Kumar, A. Kumar Jha, and L. Dhamotharan introduce a hybrid machine learning-clustering-associative rules model helpful for both team management and for players to improve their winning strategy and for the discovery of emerging players. Variable importance with respect to match-winning is computed using machine-learning techniques and is further statistically validated through the regression model. The emerging talented players are identified by clustering.

3 Basketball

In a first group of papers clustering methods are used.

In the paper "Complex networks for community detection of basketball players", A. Chessa, P. D'Urso, L. De Giovanni, V. Vitale, and A. Gebbia propose the use of a weighted complex network to detect communities of basketball players on the basis of their performances. A sparsification procedure to remove weak edges is also applied, confirmed by the normalized mutual information, so that not only the best distribution of nodes into communities is found, but also the ideal number of communities as well. An application to community detection of basketball players for the NBA regular season 2020–2021 is presented.

In the paper "Home advantage and mispricing in indoor sports' ghost games: the case of European basketball", L. De Angelis and J.J. Reade apply linear probability models to investigate the effect of ghost games in basketball with a special focus on the possible reduction of the home advantage due to the absence of spectators inside the arena. The models have been applied to a large data set covering all seasons since 2004 for the 10 most popular and followed basketball leagues in Europe, with consequences on the betting markets.

In a second group of papers, team and player prediction is considered.

In the paper "A Bayesian network to analyse basketball players' performances: a multivariate copula-based approach", P. D'Urso, L. De Giovanni, and V. Vitale propose the use of Bayesian networks to model the joint distribution of a set of indicators of player performance in basketball to discover the set of probabilistic relationships as well as the main determinants of winning percentage.

In the paper "Measuring players' importance in basketball using the generalized Shapley value", R. Metulini and G. Gnecco introduce a generalized version of the Shapley value for measuring basketball players' importance. The approach is based on the average dif-

ference in win probabilities when a player is included and not included in the lineup. Such probabilities are estimated by applying a logistic regression model in which the response is represented by the game outcome.

In the paper "Will more skills become a burden? The effect of positional ambiguity on player and team performance", J. Wang and F. Liu propose the use of a multilevel (player and team) mixed-effect regression with random-coefficient model to study the effects of positional concentration.

In a third group of papers spatial analysis is considered.

In the paper "Spatial performance analysis in basketball with CART, random forest and extremely randomized trees", P. Zuccolotto, M. Sandri, and M. Manisera use random forests and extremely randomized trees to represent maps of the court visualizing areas with different levels of scoring probability of the analysed player or team. The approaches are demonstrated by the analysis of data from the NBA regular season 2020–2021.

In the paper "Filtering active moments in basketball games using data from players tracking systems", T. Facchinetti, R. Metulini, and P. Zuccolotto develop an algorithm to automatically identify active periods by using players' tracking data in basketball. Data analysis of positioning information during the actions of a game allows a deep characterization of the performance of single players and teams.

4 Cycling

In the paper "Result-based talent identification in road cycling: discovering the next Eddy Merckx", D. Van Bulck, A. Vande Weghe, and D. Goossens develop a computer-aided system based on statistical learning methods (linear regression and random forest techniques) to assist in the identification of talented new riders (based on U23 results) who are likely to become top professional riders.

In the paper "Predicting the next Pogacar: a data analytical approach to detect young professional cycling talents", B. Janssens, M. Bogaert, and M. Maton design a data analytical system that is capable of predicting future performance of cycling athletes based on youth race performances. To facilitate the deployment of prediction algorithms in situations without complete cases, they propose an adaptation to the k-nearest neighbours imputation algorithm which uses expert knowledge. Their proposed method correlates strongly with eventual rider performance and can aid scouts in targeting young talents.

In the paper "A hydraulic model outperforms work-balance models for predicting recovery kinetics from intermittent exercise", F.C. Weigend, D.C. Clarke, O. Obst, and J. Siegler introduce one hydraulic model based on discretized differential equations and three work-balance models to study energy recovery dynamics. These models are compared on data extracted from five studies in cycling. The hydraulic model outperformed established work-balance models on all defined metrics, even those that penalize models featuring higher numbers of parameters. The results incentivize further investigation of the hydraulic model as a new alternative to established performance models of energy recovery.

5 Tennis

In the paper "A new model for predicting the winner in tennis based on the eigenvector centrality", A. Arcagni, V. Candila, and R. Grassi describe a model for computing the abilities/ratings of tennis players in a network context. By doing so, each match updates the full network, rather than only the ratings of the players involved. They propose a new measure called the B-score to rate the tennis players, who are considered nodes of a network. The network incorporates the memory over time of previous matches, and the B-scores are obtained through the so-called Bonacich centrality. These scores are used in a logit regression model to determine the winning probabilities. The proposed approach has been extensively evaluated under two perspectives: forecasting ability and betting results.

In the paper "The analysis of serve decisions in tennis using Bayesian hierarchical models", P. Tea and T.B. Swartz investigate intended serve direction with Bayesian hierarchical models applied on an extensive data source of professional tennis players at Roland Garros. They find discernible differences between men's and women's tennis, and between individual players. General serve tendencies such as the preference of serving towards the body on second serve and on high pressure points are revealed.

6 Darts

In the paper "Analysing a built-in advantage in asymmetric darts contests using causal machine learning", D. Goller analyses a sequential contest with two players in darts where one of the contestants enjoys a technical advantage by being the first-mover. Using methods from the causal machine learning literature, the author analyses the extent of the built-in advantage.

7 Hockey

In the paper "Estimation of player aging curves using regression and imputation", M. Schuckers, M. Lopez, and B. Macdonald identify the statistical tools best equipped to estimate the shape of age curves. They generalize the Delta Method to allow for non-zero maxima of a mean age curve and refer to this generalization as the Delta Plus method. Since the form of mean age curves is assumed to be concave downward but the specific form of these curves can vary, they consider a range of different estimation approaches for mean age curves including models with fixed effects and models with random effects. They also consider different data to be included in the estimation process. For players whose performance is not observed since they did not play in the top league, e.g., the NHL in ice hockey, they consider imputation of these values to improve model performance.

8 Swimming

In the paper "Dyadic analysis for multi-block data in sport surveys analytics", M. Iannario, R. Romano, and D. Vistocco introduce a data processing method for dyadic data which provides an analysis of psychological factors affecting the actor/partner interdependence by means of a quantile regression. The obtained results based on psychological behaviour could be an asset to design strategies and actions both for coaches and swimmers.

9 American Football

In the paper "Sports analytics in the NFL: classifying the winner of the superbowl", Y.F. Roumani shows an application of sports analytics in the National Football League. In particular, the author compares the classification performance of several methods (C4.5, Neural Network, and Random Forest) in classifying the winner of the Superbowl using data collected during the regular season.

In the paper "Simulation-based decision making in the NFL using NFLSimulatoR", B. Williams, W. Palmquist, and R. Elmore develop a R software package for simulating plays and drives using play-by-play data from the National Football League. The simulations are generated by sampling play-by-play data from previous football seasons. The sampling procedure adds statistical rigor to any decisions or inferences arising from examining the simulations. The authors highlight that the package is particularly useful as a data-driven tool for evaluating potential in-game strategies or rule changes within the league. In particular, they demonstrate its utility by evaluating the oft-debated strategy of "going for it" on fourth down and investigating whether or not teams should pass more than the current standard.

10 Other (different sports)

In the paper "Who's watching? Classifying sports viewers on social live streaming services", H. Liu, K.H. Tan, and X. Wu are concerned with sport SLSSs (Social Live Streaming Services) firms understanding and engaging with viewers. Sports viewers are classified based on their engagement behaviour where the perceived value and contribution of each group of viewers is identified. In the study, the authors consider a vast range of worldclass sports events including table tennis, billiards, basketball, badminton, fighting, and racing.