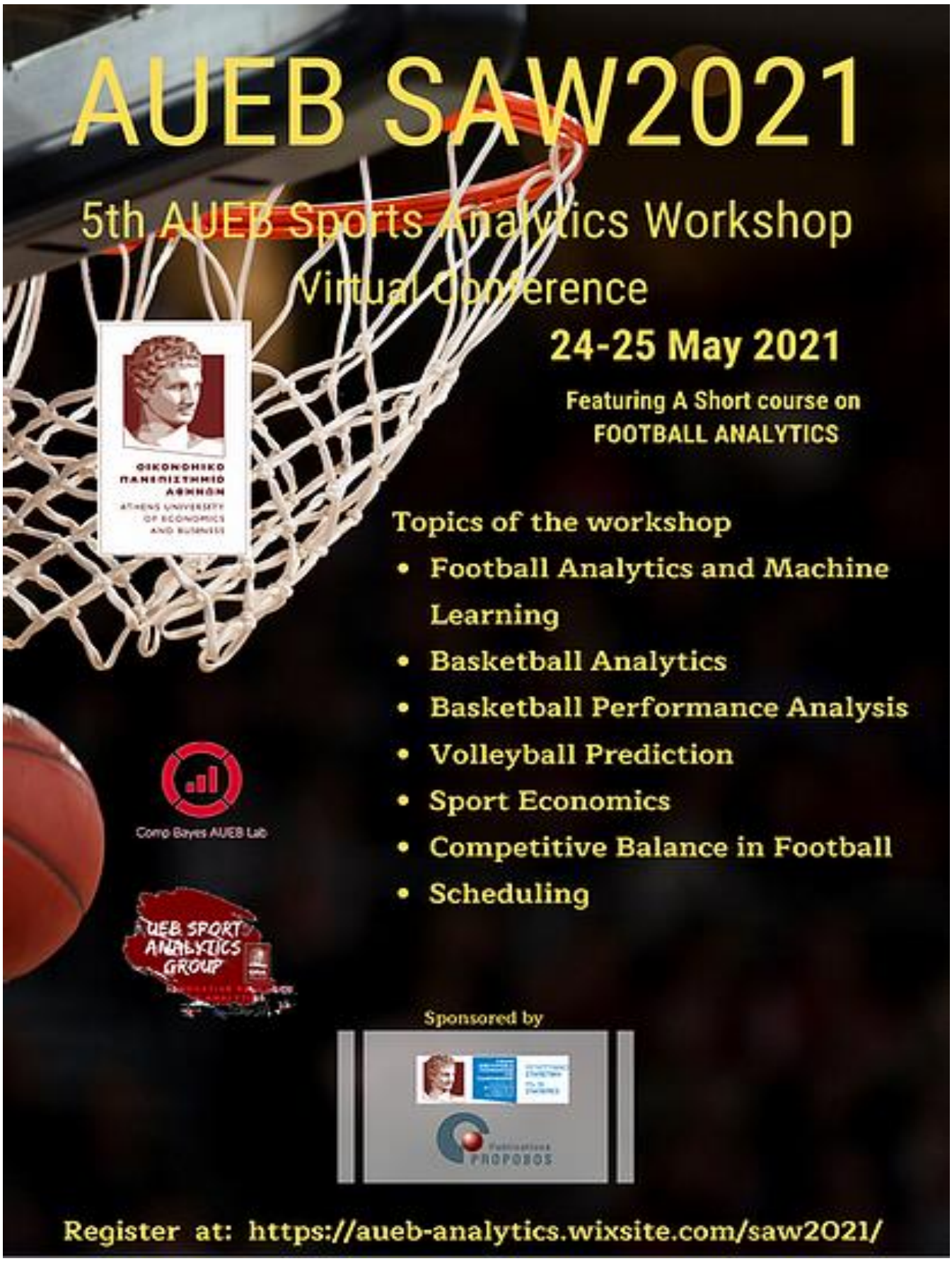

Workshop Abstracts

Monday 24 May 2021

The Probabilistic Final Standing Calculator: a fair stochastic tool to handle abruptly stopped football seasons

Christophe Ley (University of Ghent, Belgium)

The COVID-19 pandemic has left its marks in the sports world, forcing the full-stop of all sports-related activities in the first half of 2020. Football leagues were suddenly stopped and each country was hesitating between a relaunch of the competition and a premature ending. Some opted for the latter option, and took as the final standing of the season the ranking from the moment the competition got interrupted. This decision has been perceived as unfair, especially by those teams who had remaining matches against easier opponents. In this talk, I will present a tool to calculate in a fairer way the final standings of domestic leagues that have to stop prematurely: our Probabilistic Final Standing Calculator (PFSC). It is based on a stochastic model taking into account the results of the matches played and simulating the remaining matches, yielding the probabilities for the various possible final rankings. We have compared our PFSC with state-of-the-art prediction models, using previous seasons which we pretend to stop at different points in time. We illustrate our PFSC by showing how a probabilistic ranking of the French Ligue 1 in the stopped 2019-2020 season could have led to alternative, potentially fairer, decisions on the final standing.

Characterizing Team Passing Behavior in Elite Football: A Topological Approach

Abdulah Zafar (University of Toronto, Canada) Christophe Ley (University of Ghent, Belgium)

Ball movement is critical in football performance and match outcome as it is the primary method by which players 'interact' with each other. Pass networks are typically used to model these interactions, although they are only able to capture dyadic relations. In order to illustrate higher-order relations between players and to identify differences in overall team passing behaviour, the passing network was considered as a simplicial complex and methods from topological data analysis were applied. A filtration was produced from the network weights and the persistence intervals for the 0-th and 1-st homology groups were computed. The intervals were embedded into a real coordinate space and the 2-Wasserstein metric was used to compute distances between embeddings. The proposed method was implemented on performances from the 2018 FIFA World Cup and found clusters of performances which aligned with opposition team play style, pass network features, and match performance statistics.

Analyzing football tactics using high-frequency tracking data

Marius Ötting (University of Bielefeld, Germany)

Driven by recent advances in technology, tracking devices allow to collect high-frequency data on the position of players in association football matches. Corresponding data sets cover the exact locations of the players with a frequency of 10-25Hz, and hence one can monitor and measure in detail player attributes and behaviours. Such data can help coaches and scouts in several aspects, including game strategy and tactics, player evaluation, goal analysis, judging referee decision, and talent identification, to name but a few. In this talk, we consider a unique tracking data set which covers information on a single match that has taken place in one of Europe's top five leagues. It includes the (x; y) positions of all players and the ball, which are sampled with a resolution of 25 Hz. In our analysis, we focus on the convex hull created by the players of a team, thus examining several tactical aspects. It is also referred to as Effective Playing Space (EPS), calculated as the surface area (in square meters) of the convex hull of all players (excluding goalkeepers) as a measure of the playing area used by the players. For our analysis, we consider hidden Markov models (HMMs) for modelling the EPS time series data, as they naturally accommodate the idea of a match progressing through different phases, with potentially changing tactics. The unobserved states in our HMM serve for the underlying tactics of a team (e.g. defensive vs. offensive style of play). To draw a comprehensive picture of teams' tactics and their interactions, we explicitly model any within-state dependence of the two teams' EPS using copulas. Our results suggest a high persistence in the three HMM states considered, with the different states being associated with different tactics (Team A in open play and Team B defending; Team A in open play and Team B pressing; Team B attacking and Team A pinned back in its own half).

Positional value in soccer

Kostas Pelechrinis (University of Pittsburgh, USA)

Metrics applied successfully in other sports for evaluating player performance such as +/- are not appropriate in the context of soccer. Recently a number of different metrics/approaches have been proposed taking advantage of the available player tracking data. However, these data are proprietary for the most part and hard to replicate in public. In this talk, I will present an open framework that can be used to estimate the positional value in soccer, and consequently the expected contributions of a player to his team's winning chances. The framework is based on a Skellam regression, and uses data from approximately 20K games from 11 European competitions, as well as, player ratings from the FIFA videogame. The learnt model allows us to estimate the importance of every line in winning a soccer game. I then present how one can translate this to expected league points added above replacement player (eLPAR). Finally, I present some applications/case studies of using eLPAR for allocating salaries to players.

Understanding the Impact of Covid-19 on Football

James Reade (University of Reading, UK)

Covid-19 has had an obvious impact on professional sport, and football in particular - with competitions and events being postponed, cancelled, or taking place without fans in tightly controlled scenarios to avoid the spread of the virus. The extensive measurement of sport, alongside the abundance of Covid-19 related data on cases and mortality affords potentially greater insights into how a virus like Covid-19 can impact outcomes. In this work we look at how the spread of the virus throughout 2020 and 2021 has impacted football outcomes using a highly detailed dataset on events within football matches all around Europe, and using information on the spread of Covid-19 in the local areas that football clubs are located in, and in the football clubs themselves.

Basketball performance variability and line-up interactions

Marica Manisera, Paola Zuccolotto and Marco Sandri (University of Brescia, Italy)

In basketball and other team sports, measures of individual player performance should focus not only on performance level measurement, but also on performance variability. In this contribution, we model performance variability by using Markov switching models. We assume the presence of two performance regimes, good and bad performance, related to the positive or negative synergies among specific combinations of players. The relationships between each player's performance variability and line-up composition is then investigated. Based on the interactions between teammates, a score is defined for the line-ups, which correlates with the line-up's shooting performance. This shows how interactions between teammates affect team performance. This information can give important insights to coaches when deciding the best strategy in line-up composition.

Estimating team possessions in Euroleague Basketball competition

Christos Marmarinos (University of Athens, Greece)

The reference to possessions is the building block of basketball analytics. The paradox is that there is no official register of the possession count in a basketball game. This can be calculated by the play-by-play data or by a formula introduced by Kubatko at al. (2007) with the use of NBA data. The purpose of this presentation is to introduce a reliable formula for possessions estimation from box score statistics, based on data from Euroleague competition. To achieve this goal, linear regression was used. The preferred model was determined according to its fit to the data, prediction accuracy and ease of use. After the calculation of the possessions, we applied a stepwise regression analysis and we introduced a model with three indicators that explains a statistically significant part of the variation of the team's win percentage.

Tuesday 25 May 2021

Statistical methodologies and approach for the prediction of athletes' fantasy performance ratings

Vasilis Palaskas & Bill Mexias (Fantasy Sports Interactive, UK & Greece)

Fantasy Sports is a type of game involving users who try to select real-life athletes to construct their imaginary teams. A competition between players based teams is based on their actual performance in the pitch. This actual performance is measured by a rating which is composed of as sum of several skill actions executed by each athlete during a match. FSI provides markets such as the U/O, Head2Head, etc. between athletes' individual performances across several sports like Soccer, Basketball, NFL, etc. First, we investigate through a statistical exploratory analysis some interesting assumptions and patterns exist in soccer literature within the fantasy sports framework. Our foremost challenge in this work is to study and develop statistical methodologies to predict football athletes' fantasy rating using Bayesian hierarchical models based on both athletes' and teams' performance analytics and additional pre-match as well as betting odds available information (similar work can be found in work of Egidi & Gabry, 2017). A comparison between fitted models using cross-validation checks is illustrated using data from English Premier League of both 2019/20 and 2020/21 seasons.

CRM Analytics in the Greek online sports betting market. "Applied data science empowers the ultimate customer experience"

George Marinakos (OPAP, Greece)

This work shares real market experience between the online sports betting data science hub and the sports analytics scientific community, regarding how behavioral analytics methodologies meet their destination which is the sports bettor's device. Players leave their digital journey trace, surfing through their favorite football teams, leagues of preference, matches and odds to bet on, as they jump from one website to another to collect tips, statistics and recommendations. Deep understanding of the customer's digital journey is the crucial element of success for any online bookmaker competing in a such a dynamic domain. The role of the data scientist in this environment becomes critical as the art of leveraging data engineering techniques, machine learning methodologies combined with the ability to automate end to end business to customer promotional recommendation engines is the key CRM enabler to achieve relevant, personalized and appealing promotional proposition on a near real time basis. The present work aims to outline how sports betting behavioral patterns are recognized through clustering practices, how behaviors are being analyzed to predict the next customer choice through classification algorithms and what is the architectural process transforming those techniques into a single automated streamline near real time recommendation engine deployed directly in the CRM live production systems.

Clustering analysis for the Balloon D'Or partial ranking datasets of the period 2010-15

Dimitris Gkoumas (AUEB, Greece)

Balloon D'Or is one of the most prestigious awards for football players every year. It is presented by the French news magazine "France Football". The award is nominated to the male player that deemed to have performed the best, over the previous year. In the current version, it is based on a voting procedure by football journalists (one from each country member of FIFA), coaches and captains of national teams. The eligible voters are asked to rank the top 3 players, out of a list including 23, based on their preference. Since the voters are asked to rank only the top 3 players, the data are partial ranking data. This fact makes the statistical analysis very interesting. In the present paper we use data from the period 2010 – 15 to examine clustering effects of the voters towards the ranked players. We present and use appropriate methods for clustering partial ranking data to examine the clustering effects. Moreover, we distinguish voting behavioral patterns through further analysis of the clustering results. It has been debated long time, that geographical (and perhaps political) issues enter in the competition. We display the way that such an extrinsic factor could affect the final preference of a voter.

Statistical modeling for the evaluation of basketball performance

Pavlos Kollias (Aristotle University of Thessaloniki, Greece)

Basketball is one of the fastest and most complicated games in sports. The impact of the coach's decision while managing the players' replacement increases the game's complexity to its end. In this study, we present two distinct methodological procedures, in order to quantify the performance of a basketball team, in terms of game-related statistics and rotation line-ups. In the first part, linear discriminant analysis is implemented in a sample of matches of Basketball World Cup 2019, which can reveal how victory or defeat is affected by the game-related statistics. In the second part, a Markov chain model is implemented in a sequence of games of Greek Basket League to estimate the probabilities of the associated formation's performance in the long run. The Markov model could allow us to distinguish between over performing and underperforming formations and reveal the probabilities over the evolution of the game, for each formation to be in a specific rating category. Such results provide information, which could operate as a supplementary tool for the coach's decisions during a basketball game.

Estimating NBA players' salary share according to their performance on court: A machine learning approach

Michalis Tsagris (University of Crete, Greece)

It is customary for researchers and practitioners to fit linear models in order to predict NBA player's salary based on the players' performance on court. On the contrary, we focus on the players' salary share (with regards to the team payroll) by first selecting the most important determinants or statistics (years of experience in the league, games played, etc.) and then utilize them to predict the player salaries by employing a nonlinear Random Forest machine learning algorithm. We externally evaluate our salary predictions; thus we avoid the phenomenon of over-fitting observed in most papers. Overall, using data from three distinct periods, 2017-2019 we identify the important factors that achieve very satisfactory salary predictions and we draw useful conclusions.

Bayesian Models for Prediction of the Set-Difference in Volleyball

Ioannis Ntzoufras (AUEB, Greece)

The aim of this paper is to study and develop Bayesian models for the analysis of volleyball match outcomes as recorded by the set-difference. Due to the peculiarity of the outcome variable (set-difference) which takes discrete values from 1 to 3, we cannot consider standard models based on the usual Poisson or binomial assumptions used for other sports such as football/soccer. Hence, the first and foremost challenge was to build models appropriate for the set-differences of each volleyball match. Here we consider two major approaches: a) an ordered multinomial logistic regression model and b) a model based on a truncated version of the Skellam distribution. For the first model, we consider the set-difference as an ordinal response variable within the framework of multinomial logistic regression models. Concerning the second model, we adjust the Skellam distribution in order to account for the volleyball rules. We fit and compare both models with the same covariate structure as in Karlis & Ntzoufras (2003). Both models are fitted, illustrated and compared within Bayesian framework using data from both the regular season and the play-offs of the season 2016/17 of the Greek national men's volleyball league A1.

Inquiry into the discriminant game variables that predict success and performance level in elite men's volleyball

Sotiris Drikos (AUEB Sports Analytics Group)

In volleyball, teams play the game aiming to reduce the failure rate of the execution of skills, as well as to increase the rate of excellent executions. It can be argued that the simultaneous evaluation of these two aspects of performance in a specific skill would result in a more representative outcome measure of evaluation of a team's performance. The inquiry into important game variables can guide coaches to focus their training time on developing skills that are most highly correlated to success, thus making it possible to predict future game conception more accurately and helping their team to improve more quickly than their competitors. In this frame, we used data from Men's European Volleyball Championship 2019, in which 24 national teams competed, and the analysis had two approaches: The micro-level approach with the result of a balanced set as the dependent variable, and the macro-level approach with the team's performance group as the dependent variable. The research of the important factors begins from a bucket of recording variables, faces the multicollinearity problem, ends up in 7 variables under consideration and through a discriminant analysis procedure it comes up with just two performance indicators and their relative benchmarks that determine the success either in a balanced set or in an elite tournament.

Network analysis to predict tennis career paths: The case of Tsitsipas

Aggelos Alexopoulos (University of Cambridge, UK and AUEB Sports Analytics Group)

We build a statistical method to analyze the career paths of a large number of professional tennis players. In particular, we aim to predict the evolution of their career based on their profile as well as on the large amount of data collected from games that they play each year. We construct a scalable data analysis method by representing the tennis players as nodes in a network where edges exist according to a similarity matrix computed from the profile and game statistics of the players. We utilize then network analysis methods in order to identify clusters of similar players. Finally, we predict the evolution of the career of a given player based on statistics calculated from the players within her/his cluster. The dataset that motivated our work is consisted of game statistics from the last twenty years as well as the profile of 3,192 male tennis players. We test the developed methodology by conducting an out-of-sample exercise where we predict the career path of older players assuming that we do not know their whole career. We apply the proposed method on the prediction of the career path of the Greek player Stefanos Tsitsipas.

SAW 2021

Organized by

I. Ntzoufras

D. Karlis

S. Drikos

Administrative Assistant

A. Kekempanos