Mathsport International 8 2021 Proceedings

J. James Reade*

*Department of Economics, University of Reading + email address: j.j.reade@reading.ac.uk

1 Introduction

This is a collection of papers to be presented at the 8th Mathsport International Conference, hosted virtually by the University of Reading on June 24–25 2021. The 25 papers contained within this document constitute short papers based around the presentation by one (or more) of the authors at the conference.

As with the conference, the papers here cover a wide range of sports, of methods, and of topics within sport. Hence there are two papers on tennis (Kovalchik and Albert, and Kouřim), two on rugby (Bunker et al and Bracewell et al), multiple papers on football, and others on ice hockey, basketball, baseball, cricket and cycling. There are papers on scheduling and the design of tournaments, and there are multiple papers considering the impact of Covid-19 on sport. There are bibliometric studies, theoretical papers, and applied work included in this volume.

The table of contents below has page numbers, but the hyperlinks associated with each page number is rather unreliable. The papers are also presented in no particular order.

A Gaussian Mixed Membership Model with Latent Style Allocation for Describing Serve Return Patterns in Professional Tennis

Stephanie Kovalchik* and Jim Albert**

*Zelus Analytics, Austin, Texas Email: skovalchik@zelusanalytics.com
** Bowling Green State University, Bowling Green, Ohio

Abstract

The spread in the use of tracking systems in sport has made fine-grained spatiotemporal analysis a primary focus of an emerging sports analytics industry. Recently publicized tracking data for men's professional tennis has created the opportunity for the first detailed spatial analysis of the return impact characteristics of elite tennis players. Mixture models are an appealing model-based framework for spatial analysis in sport where latent variable discovery is often of primary interest. Although finite mixture models have the advantages of being highly interpretable and scalable, most implementations have the conditional spatial distribution within each latent subgroup as a standard parametric distribution. In this paper, we present a more flexible alternative that allows the conditional distribution in each latent category to be a mixed member of finite Gaussian mixtures. Our method uses an ordered stick-breaking procedure to define the distribution of mixture components in each style group and assigns a player-specific membership distribution over the style groups, which is a finite analog to a hierarchical Dirichlet process that induces partial pooling in the latent mixture components. Our model was motivated by our efforts to describe common styles of return impact location of professional tennis players and is the reason we name the approach a 'latent style allocation' model. In a fully Bayesian implementation, we apply the model to 142,803 return points played by 141 top players at Association of Tennis Professional events between 2018 and 2020 and show that the latent style allocation improves predictive performance over a finite Gaussian mixture model and identifies six unique impact styles on the first and second serve return.

1 Introduction

The last two decades have witnessed an explosion in the use of tracking data systems in professional sports [5]. Tracking systems capture fine-grained spatiotemporal information during competitive events and provide detailed summaries of player performance. Analysis of tracking data is regarded as a critical tool for gaining a competitive edge in sport and, as such, has become the central focus of an emerging sports analytics industry [10, 21].

Tracking systems began to be used at professional tennis events in the late 2000s for line call review [9]. In the past decade, the use of camera-based tracking systems have become a mainstay of major events, resulting in the collection of detailed positional data for hundreds of men's and women's tennis matches each year. Despite the wealth of positional data in the sport, the proprietary restrictions on these data have been a hindrance to research into tennis performance. As a result, only a limited number of studies have described the spatiotemporal characteristics of shots in tennis [14], and most have focused only on the serve [12, 19, 30].

The serve return is the first shot of the receiver, making it the most important receiving shot in tennis. Despite it's importance, few studies have considered the serve return characteristics of professional players. Hizan and colleagues looked at the distribution of landing location of elite junior players by categorizing the location into three different zones from which they observed that 50-70% of serve returns were hit to the middle of the court [11]. Reid, Morgan and Whiteside examined individual summary statistics of the serve return among Australian Open players, including depth of position and contact height [24]. Earlier studies had spatial data of the serve return yet restricted their analyses to univariate summaries of the physical properties of the tennis shot. To our knowledge, no prior work has directly modeled the multi-dimensional spatial characteristics of the serve return.

Across sports, a number of strategies for modeling spatial data have emerged in recent years. Popular strategies involve coarsening of spatial coordinates—either as counts on a segmented field [20, 31] or as image masks for applications in computer vision [8, 23]—each losing information from the outset. Owing to the complexity of spatial data in sport, a number of authors have proposed non-parametric methods to model tracking data. Gaussian processes have been a common non-parametric method in sports applications, with multiple examples in basketball [4] and soccer [2, 6]. The major strength of such non-parametric approaches is their flexibility. However, that flexibility often comes at the cost of interpretability and scalability [18].

Recently, mixture models have emerged as a more scalable and interpretable model-based framework for spatial data problems in sport. One of the most appealing features of mixture models for sports applications is that they directly embed a latent group factor, which can be interpreted as an unobserved subgroup category. Gaussian mixture models have been used in tennis to build a taxonomy of shots [14] and a generative model for shot events [15]. Dutta, Yurko and Ventura (2020) applied finite Gaussian mixture models to discover coverage types for passing plays from NFL tracking data [7]. Hu, Yang and Xue (2020) used a log Gaussian Cox process with a mixture of finite mixtures to describe shooting styles among NBA players [13].

There is a growing recognition of the advantages of mixture models for sports spatial data. Yet current mixture model approaches have a major limitation: the conditional distribution for the spatial outcome, conditioned on the latent category, follows a standard multivariate distribution. As a result, latent clusters are modelled with parametric distributions which may be overly restrictive for some spatial applications. In this paper, we address this limitation by introducing a Gaussian mixture model with latent style allocation, where the style-specific distribution is itself a finite mixture of multivariate Gaussians. We apply this model to newly released public data on the return impact position of professional tennis players and show that it improves the unsupervised classification of the styles of player positioning when receiving serve.

2 Study Data

The Association of Tennis Professionals (ATP) is the main organizer of tournaments in men's professional tennis. Beginning in 2018, the ATP began to provide summaries of tracking data on its website, including the location of ball at impact on the serve return (`www.atptour.com`). The data on returns include shots ending in an error or shots in play. When the receiver does not make contact with the serve, when a serve is an ace,

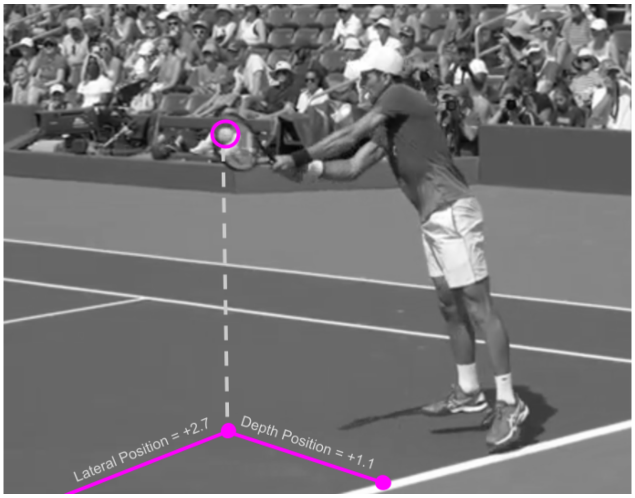


Figure 1: Illustration of the return impact location and its 2D spatial measurement. The ‘lateral position‘ is the length in meters of a straight line from the ball location at impact to the centre line of the court. The ‘depth position‘ is the length in meters of a straight line from the ball location to the baseline, with positions beyond the baseline taking a negative value.

for example, there is no return impact information. The data for the present study comprises return impact location, serve number, serving player, returning player, court side, match surface, event name and date for all matches with published tracking data between 2018 and 2020. Matches with 30 or more return points were considered complete and were retained for analysis. Further, to focus on top players, only receivers with 3 or more matches were included. After applying these inclusion criteria, the final sample included 141 receiving players, 1,334 matches, and 142,803 return points.

The ‘return impact’ is the event when the receiver’s racquet makes contact with the ball on the serve return. The 2D position consists of the lateral and longitudinal location of the ball, which will be referred to as the ‘lateral position’ and ‘depth position’, respectively. All coordinates are in meters. For the lateral position, the coordinate is the distance from the centre line down the middle of the court, with negative values being left of center and positive values right of center. The depth position is the distance from the baseline with positive values indicating a location inside the court and negative values a location beyond the baseline. An illustration of some of the return impact data is shown in Figure 1.

A typical sample of the return impact locations is shown in Figure 2. The first row of plots show the return impact locations for Dominic Thiem for one clay court match against Roger Federer played in 2019. On both first serve (to the left) and second serve (to the right), Thiem’s impact locations show strong evidence of clustering where the region within 2-3 meters behind the baseline was rarely the depth of impact in this match. Multiple modes in both the lateral and depth dimensions are also seen in the lower panel that shows sample positions of Stan Wawrinka from one hard court match played against Andy Murray in 2019. Wawrinka was more often inside the court on the second serve return than the first serve return, though he did make impact at depth of 3 meters or more behind the baseline on some second serves.

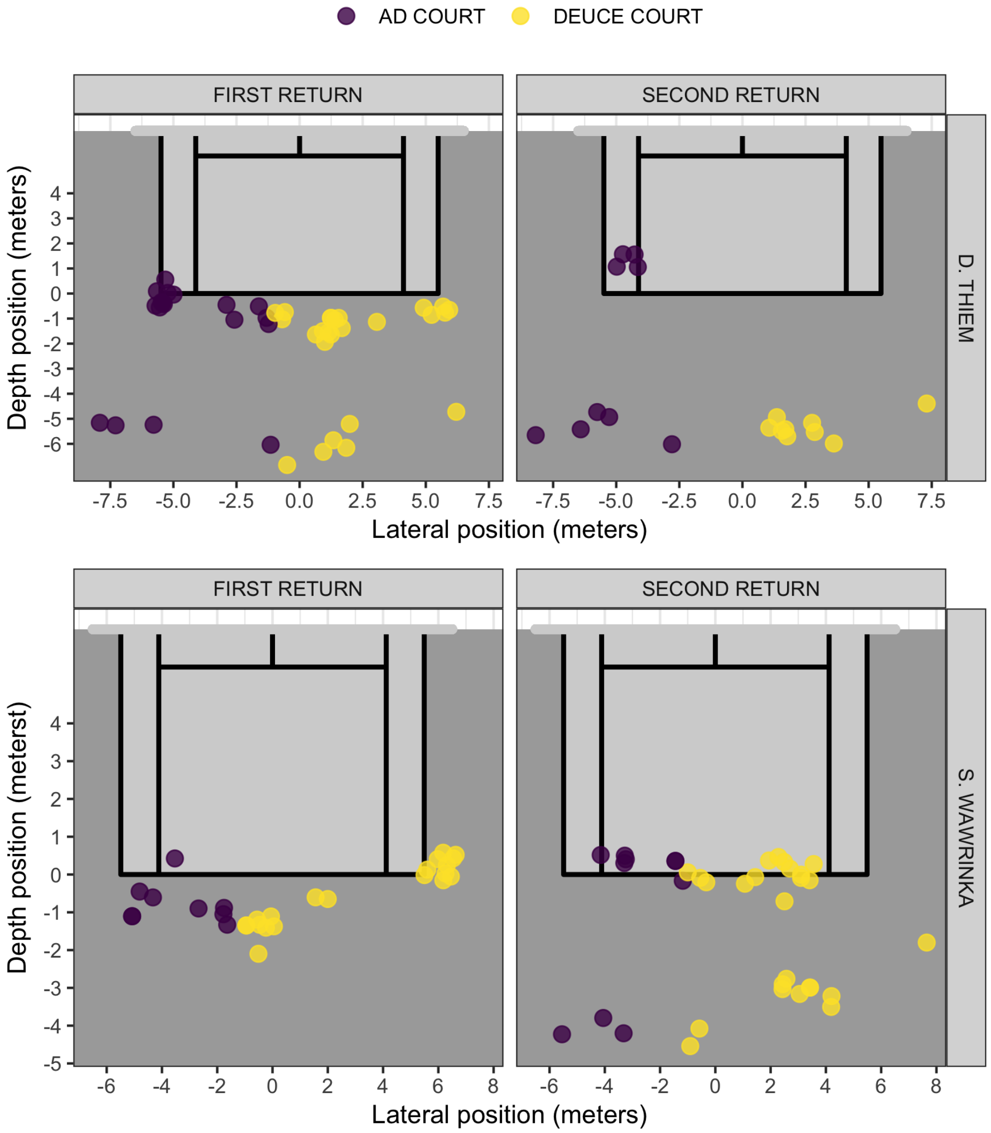


Figure 2: Return impact locations from a single match for two illustrative players: Dominic Thiem and Stan Wawrinka.

3 Model

Let $\mathbf{Y}_{ij}$ be the jth return point made by the ith receiver. We suppose that there are some M unknown patterns of return impact and each return point falls into one of these patterns. Given a pattern type, the return impact location is given a multivariate normal with pattern-specific mean and covariance,

$$\mathbf{Y}_{ij}|m_{ij} \sim MVN(\boldsymbol{\mu}_{m_{ij}}, \boldsymbol{\Sigma}_{m_{ij}}) \tag{1}$$

The pattern-specific mean incorporates receiver and server offsets around a pattern-specific set of covariate effects. Specifically,

$$\boldsymbol{\mu}_{m_{ij}} = (\boldsymbol{\alpha}_{m_{ij}} + \boldsymbol{\eta}_{r(ij)} - \boldsymbol{\delta}_{s(ij)})\mathbf{x}_{ij} \tag{2}$$

where $\boldsymbol{\alpha}_m$ is a $D \times P$ set of population effects, $\boldsymbol{\eta}_r$ is a matrix of equal dimension having receiver effects, and $\boldsymbol{\delta}_s$ a matrix of equal dimension with server effects. To illustrate the mean structure, suppose the $\mathbf{x}$ includes an intercept and indicator for a clay court match. The mean for a hard court match would then be represented as $\mathbf{x} = (1,0)$ and, writing $(\boldsymbol{\alpha}_m + \boldsymbol{\eta}_r - \boldsymbol{\delta}_s) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, the MVN mean becomes,

$$\boldsymbol{\mu} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}$$

All of the $\boldsymbol{\alpha}_m$, $\boldsymbol{\eta}_r$, and $\boldsymbol{\delta}_s$ are given a multivariate-normal prior with zero mean and covariance that has a non-informative LKJ Cholesky prior, a common choice for more efficient Bayesian hierarchical models [17]. The remaining parameter of the pattern-conditional MVN model is the observation-level covariance, $\boldsymbol{\Sigma}_m$. For the observation-level covariance, we modify the standard LKJ Cholesky prior to allow for possibly heavy-tailed outcomes. This involves drawing the scaling factors from a Student T distribution with 1 degree of freedom, truncated to the positive real line.

Key to the specification of the observation-level model in Eq. (1) is a strategy for identifying the latent pattern types. A popular class of models for modeling a latent categorical variable with grouped data are mixed membership models. The generic generating process for the mixed membership model is the following,

$$\boldsymbol{\pi}_i \sim Dirichlet(\boldsymbol{\alpha}_0) \tag{3}$$

$$m_{ij}|\boldsymbol{\pi}_i \sim Categorical(\boldsymbol{\pi}_i) \tag{4}$$

Each group entity is assigned a simplex from a simplex-generating process, such as a Dirichlet distribution as shown above. Conditional on the group-specific simplex, the ijth pattern is a random draw from the multinomial distribution with probabilities for each pattern determined by $\boldsymbol{\pi}_i$. As a result, the marginal generative model for each group i is a mixture distribution.

For the present application, where the latent factor of the Gaussian mixed membership (GMM) model represents patterns of play, the simplex parameters $\boldsymbol{\pi}_i$ can be interpreted as the ith receiver's style. This view of the GMM raises a clear drawback—the lack of sharing of styles between players. The nature of sport implies a sharing of information on all facets of the game, as players regularly train with and compete against each other. Consequently, a more plausible description of playing style would allow for information-sharing in the latent style factor.

The latent style allocation model naturally extends the GMM by placing a hierarchical generating process on the style parameters. In this process, we assign a population-level pattern distribution for each of K unknown style types. Patterns are drawn conditional on the style category, as detailed below.

$$\begin{aligned}
\boldsymbol{\theta}_k &\sim G(.) && (5)\\
\boldsymbol{\pi}_i &\sim Dirichlet(\boldsymbol{\alpha}_0) && (6)\\
k_{ij}|\boldsymbol{\pi}_i &\sim Categorical(\boldsymbol{\pi}_i) && (7)\\
m_{ij}|k_{ij} &\sim Categorical(\boldsymbol{\theta}_{k_{ij}}) && (8)
\end{aligned}$$

In this way, the latent style allocation model has two latent group variables: one defining playing style and the other defining the return impact patterns that make up styles of play.

Whereas the GMM had the grouped player data as a mixture of patterns, the latent style allocation treats playing styles as a mixture of patterns and player outcomes as a mixture of style types. A direct consequence of this key distinction between the GMM and the latent style allocation model is that players within the same style group share information, each style being defined by its pattern simplex. A player-specific marginal distribution is achieved by giving each player his own probability distribution over the style groups.

The process $G(.)$ is a parsimonious ordered stick-breaking procedure that ensures the identifiability of the latent style groups. The stick-breaking procedure begins with K ordered standard normal random variables. An inverse-logit, denoted $g(.)$ below, transforms these into an ordered set of probabilities. A stick-breaking procedure is then sequentially applied as detailed below,

$$\boldsymbol{\beta}_{km} \sim N(0,1) \text{ and } \beta_{1m} << \beta_{2m} << \ldots << \beta_{Km} \text{ for } m = 1,\ldots,M-1$$

$$\boldsymbol{\theta}_{km} = \begin{cases} g(\beta_{km}) & m = 1 \\ g(\beta_{km})\prod_{j<m}(1-g(\beta_{kj})) & m > 1 \text{ and } m < M \\ 1-\sum_{j<M}\boldsymbol{\theta}_{kj} & m = M \end{cases} \qquad (9)$$

With this process, a total of $K(M-1)$ parameters are used to define the pattern simplexes across the style groups. With the ordering constraints, the prior will encourage the first style group ($k = 1$) to have less weight on the first component mixture while the final style group ($k = K$) will tend to put greater weight on the first component. In the simplest case of a two-component Gaussian mixture ($M = 2$), the first component mixture weight would increase monotonically from the first to last latent style group. Neither the ordering of the style groups nor the ordering of the pattern groups are directly influenced by the mean or covariance parameters of the observation-level multivariate normal.

The latent group modeling presented here has close parallels to topic models. In fact, the GMM is a direct analog of latent Dirichlet allocation for a continuous multivariate outcome. A flexible extension of the standard LDA adds a hierarchical Dirichlet process (HDP) to the generating distribution of topics, in which topics are a mixture of an unknown set of categories [27]. This results in partial pooling between the known groups (i.e. documents) sharing the same category. When latent parameters follow a Dirichlet process, the number of latent groups is non-parametric. While this provides appealing flexibility, the countably-infinite set of latent categories presents difficulties for implementation in modern probabilistic programming languages, like Stan, where sampling of discrete parameters is not supported.

Like the HDP, the latent style allocation model allows information sharing between the known groups. A key advantage of the latent style allocation model, however, is that is can be readily implemented in modern Bayesian programming languages. The reason for this stems from the form of the marginal likelihood. Focusing just on the likelihood contribution to the target posterior and letting $L_{ij}(\boldsymbol{\Theta})$ be the term for one return impact location given model parameters $\boldsymbol{\Theta}$, the marginal likelihood for a single data point is the following sum

$$L_{ij}(\boldsymbol{\Theta}) = \sum_{k=1}^{K} \sum_{m=1}^{M} \pi_{ik} \theta_{km} MVN(\mathbf{Y}_{ij}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{10}$$

where π_{ik} is the ith receiver's probability weight for the kth style style and θ_{km} the probability weight for the mth pattern type within the kth style category. When K and M are fixed, it is possible to sum over the discrete variables corresponding to the latent style and pattern groups. There is a loss in flexibility when the number of latent groups are fixed but the trade-off, in this case, is a more computationally tractable posterior.

With two levels of latent grouping, the latent style allocation model can introduce greater complexity by increasing the number of style components, the number of pattern components, or both. While the choice of components at either level can be guided by standard model selection practices for Bayesian inference, some intuition about the comparative impact of increasing the number of style versus the number of pattern components will be a useful guide for selection procedures. We note that the latent style group describes the behavior of a collection of players while the latent pattern group describes within-player behavior. Thus, a greater number of patterns would be appropriate when there is considerable within player heterogeneity in return impact locations, while a greater number of styles would be appropriate when there are greater between-player differences in return impact locations.

4 Application

Using the ATP data sample, we investigate the choice of style and pattern components in the latent style allocation model that provide the best trade-off between predictive performance and model complexity. This is done by evaluating the expected log pointwise predictive density for all combinations of style and pattern components from 2 to 8. ELPD is a Bayesian leave-one-out cross-validation measure that assesses how well a model is able to describe future observations [29]. Between two models fit to the same data, the model with ELPD statistically closer to zero should be preferred. The ELPD can be efficiently approximated with Pareto-smoothed importance sampling [28].

Each latent style allocation model includes covariate effects for three serve directions on each court side and surface indicators for each of the three major surface types. First and second return points are fit separately, owing to the known heterogeneity between these two point types. In addition to identifiers for the receiver and serving player, these were the only contextual variables available from the source data.

The performance of the selected latent style allocation model is benchmarked against three other related approaches: a multivariate normal (MVN), a finite mixture, and mixed membership model. The MVN model has no latent group variables and all Y_{ij} are generated from an MVN with the same population effects and covariance. The finite mixture model is a simplification of the Gaussian mixed membership model where there is a constant pattern distribution for all players, $\boldsymbol{\pi}_i = \boldsymbol{\pi}$. In contrast to the finite mixture, the Gaussian

Table 1: Comparison of model expected log pointwise predictive density.

Model	First Serve Return	Second Serve Return
Multivariate Normal	-181748	-146143
Finite Mixture ($M = 6$)	-157105	-124560
Mixed Membership ($M = 6$)	-146365	-123107
Latent Style Allocation ($M = 6, K = 6$)	-142358	-118955

mixed membership model draws a receiver-specific pattern distribution using a Dirichlet prior with no pooling of information between players. For these benchmarks, the number of mixture components in these alternative models was equal to the number of pattern components in the selected latent style allocation model.

All models were fit using the Stan language [3] and its variational inference algorithm [16].

5 Results

Over all possible combinations of style groups and mixture components, the first serve and second serve return impact models with optimal ELPD had 6 style groups and 6 mixture components. Table 1 benchmarks the ELPD of the selected latent style allocation models against the multivariate normal, finite Gaussian mixture and Gaussian mixed membership alternatives, where the number of mixture components was set to 6 across all mixture models. The latent style allocation model improved predictive performance over both mixture model alternatives. For first serve returns, we observed a 2.8% improvement in predictive performance over a mixed membership model with the same number of mixture components; whereas a 3.5% improvement was observed for second serve returns.

For simplicity of presentation, we show detailed summaries of second serve return impact only. The tour-level posterior summaries represent the spatial distribution of return impact marginalized over all player effects. In this way, the tour-level posterior can be interpreted as the predicted impact locations for an average receiver against an average server. The second serve return tour-level posteriors reveal several modes along the lateral dimension of the court, with returns of centrally positioned serves (‘body’ serves) being more common (Figure 3). Two components, components 2 and 3, have very similar distributions on hard courts, each with probable depths that are 1 to 0 meters inside the baseline and similar high-density regions along the width of the court at 1, 3 and 5 meters from the center. We note that there is a marked difference in depth on clay court for these component densities, and that is the main difference between them.

Component 4 is the other low-dispersion component mixture (Figure 3). This mixture is distinctive for having probable depth of impact in the range of 1 meter inside to 1 meter behind the baseline. The most probable lateral positions are at 2 meters and 5 meters from the center. Component 5 has similar lateral modes but higher variance in depth, with depth locations ranging from 1 meter inside to 4 meters behind the baseline. Component 6 is unique for having a concentration of depth of position in the range of 0 to 2 meters behind the baseline across the possible lateral positions. Finally, component 1 stands out in having high dispersion in both dimensions and covering the most range for impact locations beyond the baseline and beyond the sidelines of the court.

Differences among the style types is best summarized by comparisons of their mixture probability

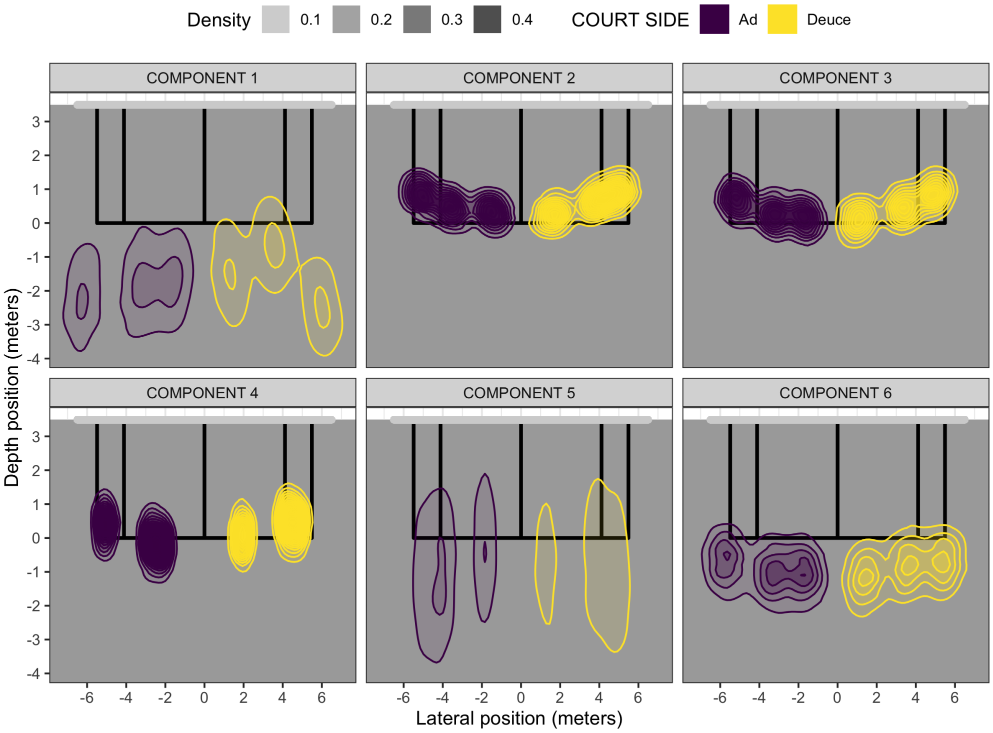


Figure 3: Tour-level posterior summaries for hard court second serve return impact locations for each mixture component of the latent style allocation model.

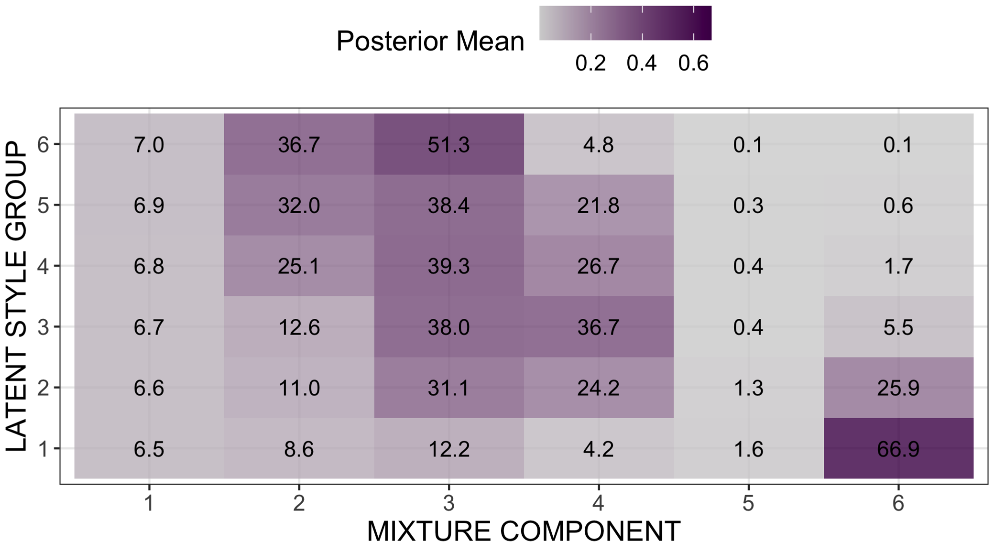


Figure 4: Posterior mean of mixture component weights for second serve return impacts by latent style group.

distributions. Figure 4 shows the posterior mean weight for each mixture and style group for second serve returns. Style group 1 has the highest weight on any single mixture, which is allocated to component 6. Style groups 2, 4 and 5 each place 25 to 30% weight on three components. Style groups 3 and 6 allocate the majority of weight to two mixtures, for style type 3, 74% is split between components 3 and 4; whereas style type 6 allocates 87% of its mixture density to components 2 and 3.

To illustrate how the latent style allocation model can assist with player-specific evaluation, we examine the posterior distributions for a selected group of top male players including 4 of the most successful players in tennis history: Andy Murray, Novak Djokovic, Roger Federer and Rafael Nadal. Figure 5 summarizes the posterior mean weights for each player and style group. Nadal and Medvedev have the most similar style distributions, with 74-92% weight allocated to the first style group on second serve return. From the predictive posteriors on hard court shown in Figure 6, we see that these styles stand out as having two distinct modes—one at 3-4 meters behind the baseline and the other 0 to 1 meter behind the baseline—on second serve return.

The remaining four players stand out in unique ways in their return impact patterns. Novak Djokovic's patterns are best represented by style group 6, where the most frequent locations at the center of the court are at the baseline and 1 meter inside the baseline when further out wide (Figure 5). Federer and and Djokovic share similar characteristics in the lateral dimension of their second serve return, however, the probability that Federer makes a return impact beyond the baseline is highly unlikely.

Andy Murray and Andrey Rublev are two players with stark differences in their second return styles. Murray's style weights are split fairly uniformly across 4 style categories and his predictive posterior shows impact locations that are 1 to 2 meters inside the court (Figure 6). Rublev has the majority of his style weights allocated to three groups and his predictive posterior shows a more conservative depth of impact, especially for centrally located serves to the Ad side.

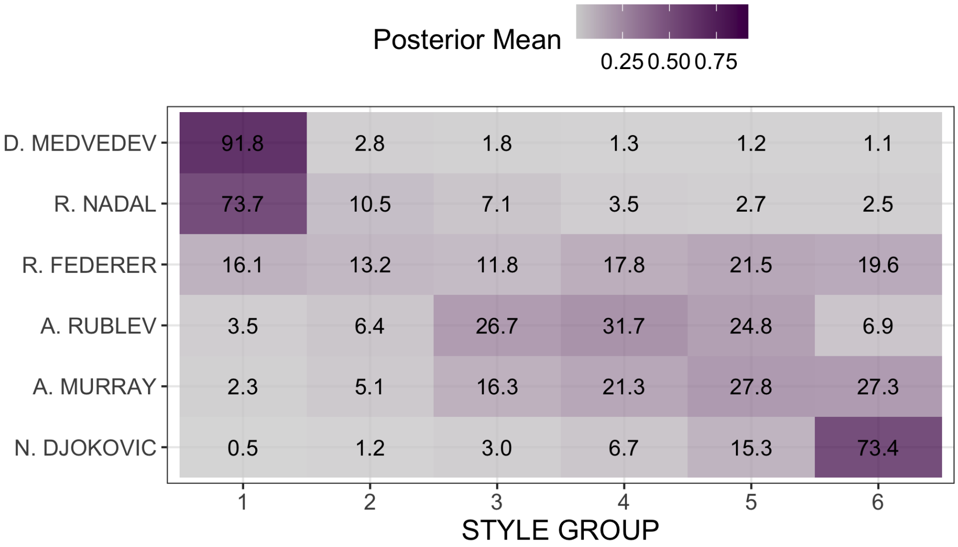


Figure 5: Posterior mean of second serve return impact latent style probability distribution for selected top male players.

6 Discussion

This paper has introduced a Gaussian mixture model with latent style allocation that provides a strategy for grouping multidimensional Gaussian mixtures across a finite number of latent groups. The model was motivated by our effort to describe patterns in the spatial distribution of return impact locations in men's professional tennis. In applying this model to thousands of points of top male players, we demonstrated that player return impact styles are better described by a mixture distribution than standard parametric distributions. We also showed that the latent style factor provides a highly interpretable method for identifying and summarising more or less similar spatial patterns among players.

It is well known that players have distinct patterns when receiving serve. Jimmy Connors and Andre Agassi, for example, were both considered 'aggressive returners' because they would take more shallow positions and make impact with the ball earlier than many of their contemporaries [25]. Although it is generally acknowledged that some top players prefer a deeper position on the serve return, detailed study of these preferences has been limited. To our knowledge, this is the first examination of return impact location of top tennis players in 2D space. Our application confirms that depth is an important property of serve return patterns, but it is not sufficient to describe the different impact styles of elite players. On first serve return, we find styles that overlap along the depth dimension but differ in their lateral positions. This study also identified styles with clear asymmetries in the lateral position on Ad and Deuce court that we have not seen discussed in the tennis coaching literature. We also observe that styles with deeper positions tend to be more diffuse, in general, suggesting that there is more range in the positioning and movement along the length of the court among top players.

On second serve returns, players are often encouraged to take a more aggressive position [1, p. 76].

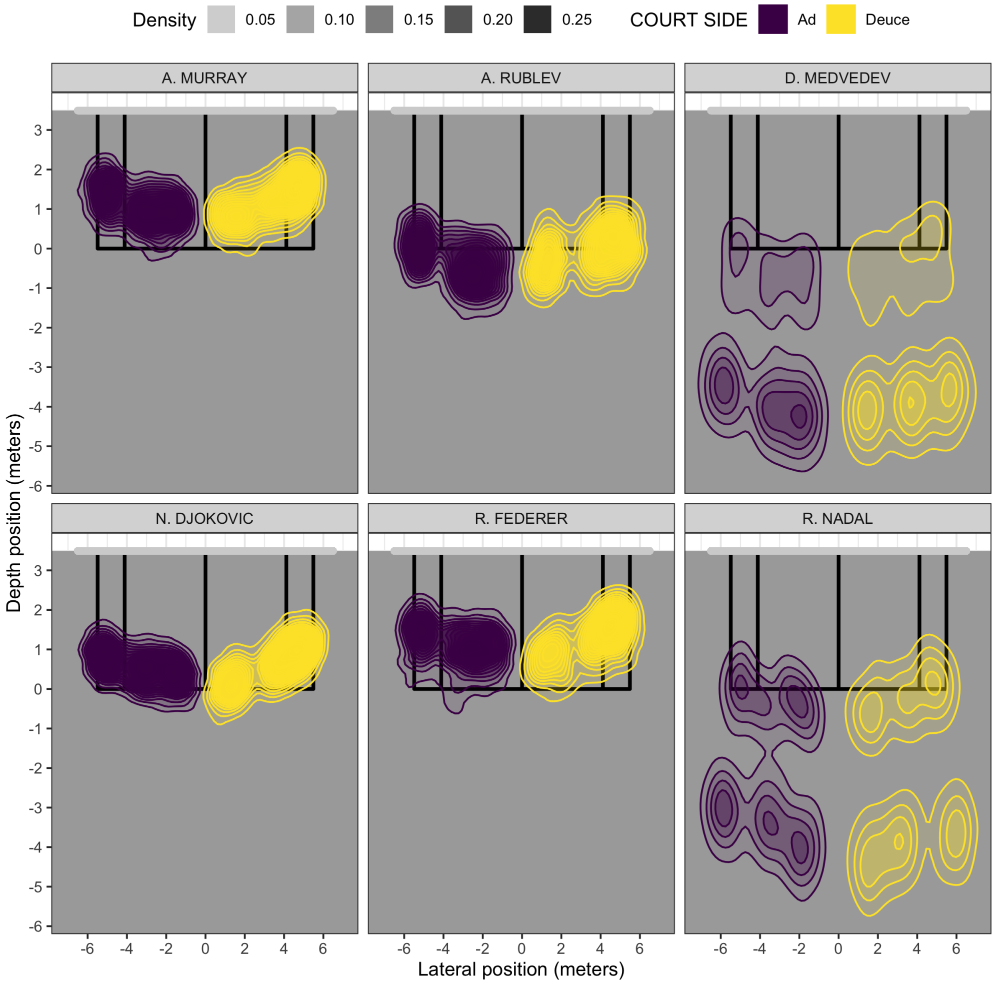


Figure 6: Predictive posterior distributions of second serve return impact locations on hard court for selected top male players.

While we observe a general shift forward in depth on second returns, this tendency is not universal. We find aggressive and defensive styles on both the first and second return that suggest that any combination among these is possible. A more general distinction we found in second return styles is the presence of distributions with multiple depth modes, with modes separated by several meters. Deeper exploration of the match data for some of the players in these style categories revealed that this kind of switching in position can happen within the same match, seemingly haphazardly. Some instructional texts recommend that players use different forward and backward positions on the serve return to 'upset your opponent's rhythm' [26, p. 299], which could explain the pattern we observed. This is one of the more surprising results of the present study and worthy of further investigation.

Our in-depth summary of a selected group of some of the most successful male tennis players in the past decade showed a wide range of return styles among the best in the game. Whether this is evidence that top players can be highly effective with any one of a number of return styles or that there is still room for even the best players to improve their return tactics remains an open question.

A key finding in this paper was the improved performance of the latent style allocation model over finite mixture alternatives. This result suggests two main conclusions. First, we observed multimodality in the lateral and depth dimensions that could not be explained by serve type or any other contextual variable available to us. This was most dramatic on second serve returns where a subset of players are observed to have high-density regions in the depth dimension that were separated by several meters. In this case, the latent style spatial descriptions required a more flexible distribution than what a standard parametric density could provide. Secondly, even among the top 150 male players, there is a wide range in the sample sizes of tracking data, as tracking systems are only available at the best events and marquee courts where lower-ranked players are less often observed. With sparse observations for some players, partial pooling of player distributions through the latent style groups can yield better predictive performance than a mixed membership model with no pooling in the assignment of player mixture weights.

More flexible frameworks for mixture models could also provide similar performance gains in modeling return impact locations over parametric or finite mixture alternatives. An HDP would be one of the most flexible options for a latent group factor variable that would allow for partial pooling and an unspecified number of group categories. The main advantage of the latent style allocation over the HDP is its ease of implementation. With modern Bayesian probabilistic programming languages, like Stan, sampling of an infinite process is often not possible without approximations [22].

While the spatial distribution of return impact locations in tennis was the primary motivation of the latent style allocation model, we believe this model can be useful for more general spatial problems in sport. As the capture of spatiotemporal data in sport continues to grow, there will be a growing interest in models that can describe complex patterns of athlete movement in space and categorize styles of movement in a model-based way. The flexibility, interpretability, and accessibility of the latent style allocation model all make it a practically meaningful tool for this emerging area of performance analysis in sport.

References

[1] R. Antoun. *Women's Tennis Tactics*. Human Kinetics, 2007. ISBN 9780736065726. URL `https://books.google.com.au/books?id=l5BtfjEd7oYC`.

[2] Iavor Bojinov and Luke Bornn. The pressing game: Optimal defensive disruption in soccer. In *10th MIT Sloan Sports Analytics Conference*, 2016.

[3] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32, 2017.

[4] Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514):585–599, 2016.

[5] Christina Chase. The data revolution: Cloud computing, artificial intelligence, and machine learning in the future of sports. In *21st Century Sports*, pages 175–189. Springer, 2020.

[6] Daniele Durante and David Dunson. Bayesian logistic gaussian process models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 194–201. PMLR, 2014.

[7] Rishav Dutta, Ronald Yurko, and Samuel L Ventura. Unsupervised methods for identifying pass coverage among defensive backs with nfl player tracking data. *Journal of Quantitative Analysis in Sports*, 1 (ahead-of-print), 2020.

[8] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Memory augmented deep generative models for forecasting the next shot location in tennis. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1785–1797, 2019.

[9] Mark Fischetti. In or out? *Scientific American*, 297(1):96–97, 2007.

[10] Bill Gerrard. Analytics, technology and high performance sport. *Critical issues in global sport management*, 205, 2016.

[11] Hazuan Hizan, Peter Whipp, Machar Reid, and Jon Wheat. A comparative analysis of the spatial distributions of the serve return. *International Journal of Performance Analysis in Sport*, 14(3):884–893, 2014.

[12] Hazuan Hizan, Peter Whipp, and Machar Reid. Gender differences in the spatial distributions of the tennis serve. *International Journal of Sports Science & Coaching*, 10(1):87–96, 2015.

[13] Guanyu Hu, Hou-Cheng Yang, and Yishu Xue. Bayesian group learning for shot selection of professional basketball players. *Stat*, page e324, 2020.

[14] Stephanie Kovalchik and Machar Reid. A shot taxonomy in the era of tracking data in professional tennis. *Journal of sports sciences*, 36(18):2096–2104, 2018.

[15] Stephanie Kovalchik, Martin Ingram, Kokum Weeratunga, and Cagatay Goncu. Space-time von cramm: Evaluating decision-making in tennis with variational generation of complete resolution arcs via mixture modeling. *arXiv preprint arXiv:2005.12853*, 2020.

[16] Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic variational inference in stan. *Advances in Neural Information Processing Systems*, 2015:568–576, 2015.

[17] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[18] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

[19] Sami Mecheri, François Rioult, Bruno Mantel, François Kauffmann, and Nicolas Benguigui. The serve impact in tennis: First large-scale study of big hawk-eye data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(5):310–325, 2016.

[20] Andrew Miller, Luke Bornn, Ryan Adams, and Kirk Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *International conference on machine learning*, pages 235–243. PMLR, 2014.

[21] Elia Morgulev, Ofer H Azar, and Ronnie Lidor. Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4):213–222, 2018.

[22] Akihiko Nishimura, David Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for sampling discrete parameters. *arXiv preprint arXiv:1705.08510*, 853, 2017.

[23] Akhil Nistala and John Guttag. Using deep learning to understand patterns of player movement in the nba. In *In Proceedings of the MIT Sloan Sports Analytics Conference*, pages 1–14, 2019.

[24] Machar Reid, Stuart Morgan, and David Whiteside. Matchplay characteristics of grand slam tennis: implications for training and conditioning. *Journal of sports sciences*, 34(19):1791–1798, 2016.

[25] J. Rutherford. *Skills, Drills & Strategies for Tennis*. Taylor & Francis, 2017. ISBN 9781351817288. URL `https://books.google.com.au/books?id=nkQrDwAAQBAJ`.

[26] M. Smith and F. Stolle. *Absolute Tennis*. New Chapter Press, Incorporated, 2017. ISBN 9781937559748. URL `https://books.google.com.au/books?id=vHxbAQAACAAJ`.

[27] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[28] Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther. Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1):3581–3618, 2016.

[29] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

[30] Xinyu Wei, Patrick Lucey, Stuart Morgan, Peter Carr, Machar Reid, and Sridha Sridharan. Predicting serves in tennis using style priors. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2207–2215, 2015.

[31] Yisong Yue, Patrick Lucey, Peter Carr, Alina Bialkowski, and Iain Matthews. Learning fine-grained spatial models for dynamic sports play prediction. In *2014 IEEE international conference on data mining*, pages 670–679. IEEE, 2014.

Supervised sequential pattern mining for identifying important patterns of play in rugby

R. Bunker*, K. Fujii**, H. Hanada***, I. Takeuchi****

* Graduate School of Informatics, Nagoya University, Japan
(research partly undertaken while employed by RIKEN).
rory.bunker@g.sp.m.is.nagoya-u.ac.jp
** Graduate School of Informatics, Nagoya University, Japan. fujii@i.nagoya-u.ac.jp
*** RIKEN Center for Advanced Intelligence Project, Japan. hiroyuki.hanada@riken.jp
*** Department of Computer Science, Nagoya Institute of Technology, Japan and
RIKEN Center for Advanced Intelligence Project, Japan. takeuchi.ichiro@nitech.ac.jp

[Note: this paper has considerable overlap with the full version, which is available at `https://arxiv.org/abs/2010.15377`, and as of 24 May 2021 is still under review at PLOS One and is not yet published.]

Abstract

In sport, despite being useful for identifying frequently occurring patterns of play, unsupervised sequential pattern mining techniques, which are applied to datasets consisting of unlabeled sequences of events, are limited in that they are unable to identify specific patterns that discriminate between good or bad outcomes, which is of considerable interest to coaches and performance analysts. In this study, Safe Pattern Pruning (SPP), a recently proposed supervised sequential pattern mining algorithm, was applied to data consisting of 490 labelled event sequences from one professional rugby team's matches from the 2018 Japan Top League competition. The SPP-obtained patterns of play that discriminated the most between scoring and non-scoring outcomes from both the team's and opposition teams' perspectives were compared with the most frequent patterns obtained with well-known unsupervised sequential pattern mining algorithms applied to subsets of the original dataset based on the label. Compared to the unsupervised methods, SPP obtained more sophisticated and useful patterns for performance analysis.

1 Introduction

Large amounts of data are captured in sport as a result of the advent of GPS tracking, optical and video analysis systems, and enhancements in computing power and storage. There is great interest in using such data to identify tactics, behavior and performance using methods from, e.g., statistics, data mining, machine learning, and deep learning. In this study, we consider sequences of events derived from event logs exported from video analysis systems, and aim to identify important patterns (sub-sequences).

Matches are made up of sequences of play, which are in turn made up of individual events. In practice, event logs exported from video analysis systems are commonly converted into frequencies, averages or ratios for use as performance indicators [1]; however, this results in the loss of the information associated with the order of the events. The occurrence of specific events in a particular order can influence outcomes.

Invasion sports, e.g., soccer, hockey, basketball and rugby, have many events and patterns that occur frequently and repeatedly, but may not necessarily be important for positive outcomes, e.g., scoring. One event log is generally for one specific match and can be split (delimited) into sequences of play based on some specificed rules, and sequences across a whole season or multiple season can then be analysed. If the sequences of play contain scoring events, these events can be extracted and used as a label, and supervised methods can be employed. In this paper, we consider ordered sequences of events and use supervised sequential pattern mining to identify important sub-sequences of events.

Sequential pattern mining [2] involves discovering frequent sub-sequences as patterns from databases that consists of ordered event sequences, which may or may not have strict notions of time [3] ([4] provides a review). Safe pattern pruning (SPP) [5, 6] is a supervised method (i.e., uses labeled data) that combines a convex optimisation technique called safe screening [7] with sequential pattern mining. Unsupervised sequential pattern mining techniques have been applied in sport, e.g, for technical/tactical analysis in judo [8], and to test for significant trends and interesting sequential patterns in cycling training [9]. Decroos, Van Haaren & Davis [10] combined clustering and the CM-SPADE method to data from soccer, with the method allowing for greater importance to be placed on relevant (crosses, shots) rather than merely frequent events (normal passes). We also aim to identify important rather than merely frequent events in this study, but by using a supervised rather than unsupervised approach. In rugby, previous studies have considered sequences of play in matches by analyzing their duration. The durations of sequences of plays that led to tries at the 1995 Rugby World Cup (RWC) were studied by [11], and [12] found that teams at the 2003 RWC that created movements exceeding 80 seconds in duration were more successful. Recently, [13] applied K-modes clustering to sequences of play in rugby, and [14] used convolutional and recurrent neural networks to predict the outcomes of sequences of play in rugby based on the order and on-field locations of events.

In this study, we apply SPP to event sequences from all matches played by a professional rugby team in the 2018 Japan Top League season. As a basis for comparison, we compare the SPP-obtained sub-sequences with those obtained by well-known unsupervised sequential pattern mining methods (PrefixSpan [15], GSP [16], Fast [17], and CM-SPADE and CM-SPAM [18]) when they are applied to subsets of the original labelled data, partitioned based on the label. The present study is motivated by the fact that, although sequential pattern mining techniques have been applied to sport, only unsupervised methods appear to have been used to date, and a comparison with a supervised method has not been conducted. Also, sequential pattern mining (unsupervised or supervised) has not yet been applied to sequences of play in rugby.

2 Materials and Methods

2.1 Data

XML event log files, generated from video tagged in SportsCode by the team's performance analyst, were obtained for all of the team's matches in the 2018 season.[1] Matches consist of passages of play (i.e., sequences of events), which are in turn made up of individual events. The original dataset consisted of a long sequence of events for each match (for each XML file). One approach we considered initially was to label the match sequences with win/loss outcomes, however, this did not yield interesting results since it is obvious that sequences containing a greater number of scoring events will occur in sequences of events for winning

[1] The team is not named for reasons of confidentiality. Written consent was obtained to use the data for research purposes.

matches. We then generated more granular data by specifying rules that delimit matches into sequences of play. There were 24 unique events (12 unique events for the team and opposition teams) in our dataset (Table 2, [19]).

We specified rules to split match sequences into sequences of play (Figure 2, [19]). Sequences of play should be of appropriate length, and begin and end at logical points, e.g., when play stops, or possession changes (e.g., [20]). We defined a sequence of play to start with either a kick restart, scrum, or lineout, which are events that result in play temporarily stopping. The rules specified that the occurance of a kick restart, scrum (except for a scrum reset where a scrum follows a previous scrum), or lineout, results in this event becoming the first event in a new passage of play event sequence; otherwise, if a try is scored or a kick at goal occurs, a new passage of play also begins. This resulted in a delimited dataset consisting of 490 sequences of play.[2] At this stage, the sequences of play were unlabelled and still contained scoring events (try scored, kick at goal) by the team and opposition teams.

The dataset was then divided into "scoring" and "conceding" datasets, where the sequences were from the team's scoring and conceding perspectives, respectively. Each sequence in the scoring and conceding datasets was labelled based on whether or not points (try or kick at goal) were scored or attempted by the team or opposition team, respectively.[3]. Since the label now identified scoring/not scoring or conceding/not conceding, the try scored and kick at goal events were removed from the event sequences (the list of events for the original delimited, scoring and conceding datasets are presented in Table 3, [19]). The process to create the scoring and conceding datasets from the original delimited dataset is shown in the upper half of Figure 3 in [19].

The SPP method[4] takes as input a dataset $\{(\boldsymbol{g}_i, y_i)\}_{i\in[n]}$, where $\boldsymbol{g}_i$ represents the i-th sequence of play and $[n]$ is the number of sequences of play in the dataset; each sequence $\boldsymbol{g}_i$ is assigned a label from $y_i \in \{\pm 1\}$. SPP constructs a sparse linear combinations of patterns, $f(\boldsymbol{g}_i;\mathscr{Q}) = \sum_{\boldsymbol{q}_j\in\mathscr{Q}} w_j I(\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i) + b$, where $I(\cdot)$ is an indicator function that takes the value 1 if sequence $\boldsymbol{g}_i$ contains sub-sequence $\boldsymbol{q}_i$ and 0 other otherwise, $\mathscr{Q}$ is the set of all possible patterns, and $w_j \in \mathbb{R}$ and $b \in \mathbb{R}$ are linear model parameters, which are estimated by solving the following minimisation problem: $\min_{\boldsymbol{w},b} \sum_{i\in[n]} \ell(y_i, f(\boldsymbol{g}_i;\mathscr{Q})) + \lambda\|\boldsymbol{w}\|_1 = \min_{\boldsymbol{w},b} \sum_{i\in[n]} \max\left\{0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\right\}^2 + \lambda\|\boldsymbol{w}\|_1$, where $\boldsymbol{w} = [w_1, \ldots, w_d]^\top$ is a vector of weights, ℓ is a loss function (the squared hinge-loss function $\ell(y, f(\boldsymbol{x}_i)) = \max\{0, 1 - yf(\boldsymbol{x}_i)\}^2$ is used for a two-class problem like ours), and $\lambda > 0$ is a regularization parameter tuned by cross-validation.[5] The feature vector $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ is defined for the ith sequence $\boldsymbol{g}_i$ as $x_{ij} = I(\boldsymbol{q}_j \sqsubseteq \boldsymbol{g}_i)$ for $j = 1, \ldots, |\mathscr{Q}|$, i.e., the feature vectors $\boldsymbol{x}_i = [I(\boldsymbol{q}_1 \sqsubseteq \boldsymbol{g}_i), I(\boldsymbol{q}_2 \sqsubseteq \boldsymbol{g}_i), \ldots, I(\boldsymbol{q}_d \sqsubseteq \boldsymbol{g}_i)]$ are binary variables that take the respective values 1 or 0 based on the presence or absence of sub-sequence $\boldsymbol{q}_j$ in sequence $\boldsymbol{g}_i$. Discriminative patterns are those that

[2] The passage of play event sequences are available on GitHub: `https://github.com/rorybunker/rugby-sequences`)

[3] Note that while a try scored is certain in points being scored, a kick at goal is not always successful. In our data, only the kick at goal being attempted was available, not whether the goal was actually successful. Since it was deemed more important to identify points-scoring opportunities than whether the kick at goal was ultimately successful (determined by the goal kicker accuracy/skill), we assumed that 100% of kicks at goal resulted in points being scored.

[4] The SPP code is available at `https://github.com/takeuchi-lab/SafePatternPruning`

[5] The regularization term means that many of the weights in the optimal solution to the optimisation problem will be zero. In addition, prior to solving the optimisation problem, SPP reduces the size of $\mathscr{Q}$ (generally large) by removing unnecessary patterns from the entire pattern-tree grown by PrefixSpan according to the SPP pruning criterion [6], and some weights are also removed prior to solving the optimisation problem by using safe screening.

have positive weights (in absolute terms) in the optimal solution to the minimization problem.[6]

The SPP-obtained sub-sequences were compared with those obtained by the unsupervised methods. The SPMF package [21] (v2.42c) was used to apply the five unsupervised sequential pattern mining methods. Since the unsupervised methods use unlabelled data, although support values for the patterns of play can be obtained (i.e., the number of times a particular pattern occurred), unlike SPP, we cannot obtain pattern weights. For a more fair comparison between the unsupervised methods and the supervised method (SPP), we assume prior knowledge of the sequence labels in the case of the unsupervised methods. The unsupervised methods were applied to "scoring+1" and "conceding+1" datasets (subsets of the scoring and conceding datasets with label +1), containing the sequences in which the team (respectively, opposition team) actually scored.

To exclude patterns that may have occurred by chance, patterns (q_js) with support values of five or more were considered. In the case of the patterns obtained by the unsupervised models, the top five patterns with the largest support values were recorded. In the case of the SPP-obtained patterns, the top five patterns with the largest positive w_j values were recorded. In addition, we restricted our analysis to patterns of play that had the highest positive weights. For the scoring dataset, this means the patterns that had a positive contribution to the team scoring, and for the conceding dataset, the patterns that had a positive contribution to opposition teams scoring.[7]

3 Results

The SPP algorithm was firstly applied to the scoring and conceding datasets. SPP obtained 93 patterns when applied to the scoring dataset, of which 75 had support of 5 or higher. Of these 75 patterns of play, 38 were discriminative (i.e., had positive weight, $w_j > 0$, in the solution to the regularised optimisation problem). SPP obtained 72 patterns when applied to the scoring dataset, of which 51 had support of 5 or higher. Of these 51 patterns, 31 were discriminative.

The five most discriminative patterns between scoring and non-scoring outcomes (i.e., patterns with the largest w_j values), obtained by applying SPP to the scoring dataset, are listed along with their corresponding weight values and odds ratios (ORs) in Table 1. The ORs for these patterns (simply the exponential of the weights) are included to aid in interpretation by providing a value that compares the cases where a sequence contains a particular pattern and when it does not.

The pattern with the highest weight value (0.919), which discriminated the most between scoring and non-scoring sequences, was a pattern consisting of a single line break event. The OR for the linebreak pattern is exp(0.919)=2.506, meaning that the team is 2.5 times more likely to score when a line break occurs in a sequence of play than if a line break is not made in a sequence of play (line breaks involve breaking through an opposition team's line of defense, advancing the attacking team forward and creating potential scoring opportunities). A lineout followed by phase play was the second most discriminative pattern between scoring and not scoring (w=0.808, OR=2.242) (set pieces such as lineouts can provide stable platforms from which to attack). The third most discriminative pattern (w=0.796, OR=2.217) can be interpreted as a kick in play made by the team being re-gathered, resulting in retained possession. The fourth most discriminative pattern

[6] The SPP command line options were set to -c 1 -M 1 -L 20. The c and M options mean that 10-times 10-fold cross-validation is used to tune λ, and L being set to 20 means the maximimum possible pattern length is 20. The default L1-regularised L2-SVM option was used to solve the minimization problem.

[7] For the sake of brevity, we did not consider the patterns that had the highest contribution to "not scoring" and "not conceding."

dataset	pattern	support	weight	OR
scoring	linebreak	77	0.919	2.506
	lineout, phase	71	0.808	2.242
	phase, breakdown, kick in play, phase, breakdown	9	0.796	2.217
	[phase, breakdown]x4, kick in play	9	0.732	2.079
	O-restart received, [O-phase, O-breakdown]x2, O-kick in play, phase, breakdown	6	0.710	2.033
conceding	O-linebreak	32	0.613	1.846
	O-phase, error, O-breakdown	10	0.392	1.479
	O-lineout	86	0.357	1.428
	O-breakdown, O-breakdown, O-phase, O-breakdown	5	0.339	1.403
	[O-breakdown, O-phase]x6, O-breakdown	16	0.261	1.299

Table 1: Top five SPP-obtained patterns of play with the largest weights (odds ratios) for the scoring and conceding datasets. The notation $[p]$ x n denotes that pattern p is repeated n times.

(w=0.732, OR=2.079) represents four repeated phase-breakdown plays by the team, followed by them making a kick in play, which indicates repeated retaining of possession before presumably attempting to gain territory in the form of a kick. The fifth most discriminative pattern (w=0.710, OR=2.033) can be interpreted as the opposition team receiving a kick restart made by the team, attempting to exit their territory via a kick but not finding touch, thus giving the ball back to the team from which they can potentially launch an attack.

The five most discriminative patterns between conceding and non-conceding outcomes, i.e., patterns with the highest positive weights when applying SPP to the conceding dataset, are listed along with their weight values and ORs in Table 1. A linebreak (w=0.613, OR=1.846) made by the opposition team was the most discriminative pattern between sequences in which the team conceded and did not concede, or equivalently, between sequences in which the opposition team scored and did not score. The weight value was not as large as for the team scoring from a linebreak against the opposition team (w=0.919 vs. w=0.613), suggesting that the team considered in this study appears to have strong defence, since linebreaks made by the opposition team were less likely to result in the opposition team scoring compared to the likelihood of linebreaks made by the team through the opposition defensive line resulting in the team scoring. The OR of 1.8 indicates that the opposition team is 1.8 times more likely to score when they make a linebreak in a sequence of play than if they do not. The second most discriminative pattern (w=0.392, OR=1.479) between conceding and non-conceding outcomes can be interpreted as the opposition team being in possession of the ball, the team making some form of error, resulting in the opposition team regaining possession. The third most discriminative pattern between conceding and non-conceding outcomes was an opposition team lineout (w=0.357, OR=1.428). The fourth (w=0.339, OR=1.403) and fifth (w=0.261, 1.299) most discriminative patterns between conceding and non-conceding outcomes represent repeated phase and breakdown play, with the fifth sub-sequence, for example, indicating the opposition team making over six repeated consecutive phases and breakdowns, suggesting the retaining of possession and building of pressure by the opposition team.

Table 2 shows the top five sub-sequences in terms of their support (frequency of occurence) from the scoring+1 and conceding+1 datasets. The obtained results show that only common, repeated patterns (breakdowns/phases) were detected by the unsupervised methods. By contrast, SPP not only provides a measure of the importance of the obtained patterns from the calculated weights and ORs, it also obtained more sophisticated and important patterns of play for use by coaches and/or performance analysts.

dataset	PrefixSpan	CM-SPAM	CM-SPADE	GSP	Fast	support
scoring+1	phase	phase	phase	phase	phase	84
	phase, breakdown	breakdown	breakdown	breakdown	breakdown	60
	breakdown	phase, breakdown	phase breakdown	phase breakdown	phase breakdown	60
	phase, phase	phase, phase	breakdown, phase	phase, phase	phase, phase	59
	phase, breakdown, phase	phase, breakdown, phase	phase, phase	breakdown, phase	breakdown, phase	59
conceding+1	O-phase	O-phase	O-phase	O-phase	O-phase	39
	O-phase, O-breakdown	O-breakdown	O-breakdown	O-breakdown	O-breakdown	33
	O-breakdown	O-phase, O-breakdown	O-phase, O-breakdown	O-phase, O-breakdown	O-phase, O-breakdown	33
	O-phase, O-phase	O-phase, O-phase	O-breakdown, O-phase	O-phase, O-phase	O-phase, O-phase	29
	O-phase, O-breakdown, O-phase	O-phase, O-breakdown, O-phase	O-phase, O-phase	O-breakdown, O-phase	O-breakdown, O-phase	29

Table 2: Top five patterns of play obtained by the unsupervised methods with the largest support for the scoring+1 and conceding+1 datasets.

4 Discussion

In this study, a supervised sequential pattern mining method called safe pattern pruning (SPP) was applied to sequences of play from the matches of a professional rugby union in Japan in the 2018 Top League season, which were labelled with points scoring/conceding outcomes, in order to identify patterns that discriminate between these outcomes. SPP was compared with a number of well-known unsupervised sequential pattern mining methods.

The obtained results suggest that SPP could detect sophisticated and important patterns of play that, when interpreted, would be of potential use for coaches and/or performance analysts for own- and opposition-team analysis to identify opportunities and to devise tactical strategies. The gained insights would be useful to both the team as well as opposition teams that are due to play the team. For both the team and their opposition teams, linebreaks were found to be most associated with scoring, and lineouts were found to be more associated with the creation of scoring opportunities than scrums. The latter result is consistent with [13], who found that lineouts followed by a driving maul are common approaches to scoring tries (albeit in a different competition, Super Rugby), and with [22], who found that around one-third of tries in the Japan Top League from 2003 to 2005 came from lineouts, the highest of any try source. As well as creating lineouts or perhaps prioritising them over scrums, our results suggest that effective strategies for opposition teams to employ may have included maintaining possession with repeated phase-breakdown play (aiming for over six repetitions), shutting down the team's ability to regain kicks in play, and making sure to find touch on exit plays from kick restarts made by the team.

The approach highlighted the potential utility of supervised sequential pattern mining as an analytical framework for performance analysis in sport, and more specifically, the potential utility of sequential pattern mining methods for performance analysis in rugby. Although the results obtained are encouraging, a limited amount of data from one sport was used, spatial information such as field position was not available in the data, only the team (not the specific player) that performed particular events was considered, and while SPP considers the order of events within sequences, it does not consider the order of the sequences within matches. In future work, we intend to apply the method to more data and to other sports in order to confirm its efficacy.

Acknowledgments

This work was partially supported by MEXT KAKENHI (20H00601, 16H06538), JST CREST (JPMJCR1502), and RIKEN Center for Advanced Intelligence Project.

References

[1] Hughes MD, Bartlett RM. The use of performance indicators in performance analysis. Journal of sports sciences. 2002;20(10): 739–754.

[2] Agrawal R, Srikant R. Mining sequential patterns. In: Proceedings of the eleventh international conference on data engineering. IEEE; 1995. p. 3–14.

[3] Mabroukeh NR, Ezeife CI. A taxonomy of sequential pattern mining algorithms. ACM Computing Surveys (CSUR). 2010 Dec 3;43(1): 1-41.

[4] Fournier-Viger P, Lin JCW, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition. 2017;1(1):54–77.

[5] Nakagawa K, Suzumura S, Karasuyama M, Tsuda K, Takeuchi I. Safe pattern pruning: An efficient approach for predictive pattern mining. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 1785–1794.

[6] Sakuma T, Nishi K, Kishimoto K, Nakagawa K, Karasuyama M, Umezu Y, et al. Efficient learning algorithm for sparse subsequence pattern-based classification and applications to comparative animal trajectory data analysis. Advanced Robotics. 2019;33(3-4): 134–152.

[7] Ghaoui LE, Viallon V, Rabbani T. Safe feature elimination for the lasso and sparse supervised learning problems. arXiv preprint arXiv:10094219. 2010;.

[8] La Puma I, de Castro Giorno FA. Ontology-Based Data Mining Approach for Judo Technical Tactical Analysis. In: The Third International Conference on Computing Technology and Information Management (ICCTIM2017); 2017. p. 90.

[9] Hrovat G, Fister Jr I, Yermak K, Stiglic G, Fister I. Interestingness measure for mining sequential patterns in sports. Journal of Intelligent & Fuzzy Systems. 2015;29(5): 1981–1994.

[10] Decroos T, Van Haaren J, Davis J. Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 223–232.

[11] Carter A, Potter G. The 1995 rugby world cup finals. 187 tries. 2001) Notational Analysis of Sport III University of Wales Institute Cardiff Cardiff, Wales UWIC. 2001; p. 224–229.

[12] van Rooyen KM, Noakes DT. Movement time as a predictor of success in the 2003 Rugby World Cup Tournament. International journal of Performance analysis in Sport. 2006;6(1):30–39.

[13] Coughlan M, Mountifield C, Sharpe S, Mara JK. How they scored the tries: applying cluster analysis to identify playing patterns that lead to tries in super rugby. International Journal of Performance Analysis in Sport. 2019;19(3):435–451.

[14] Watson N, Hendricks S, Stewart T, Durbach I. Integrating machine learning and decision support in tactical decision-making in rugby union. Journal of the Operational Research Society. 2020; p. 1–12.

[15] Wang K, Xu Y, Yu JX. Scalable sequential pattern mining for biological sequences Proceedings of the thirteenth ACM international conference on Information and knowledge management. 2004;178–187

[16] Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. In International Conference on Extending Database Technology 1996 Mar 25 (pp. 1-17). Springer, Berlin, Heidelberg.

[17] Salvemini E, Fumarola F, Malerba D, Han J. Fast sequence mining based on sparse id-lists. In International Symposium on Methodologies for Intelligent Systems. 2011;316-325.

[18] Fournier-Viger P, Gomariz A, Campos M, Thomas R. Fast vertical mining of sequential patterns using co-occurrence information. In Pacific-Asia Conference on Knowledge Discovery and Data Mining 2014;40-52.

[19] Bunker, R., Fujii, K., Hanada, H., & Takeuchi, I. Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. arXiv preprint arXiv:2010.15377. 2020. `https://arxiv.org/abs/2010.15377`

[20] Liu T, Fournier-Viger P, Hohmann A. Using diagnostic analysis to discover offensive patterns in a football game. In: Recent Developments in Data Science and Business Analytics. Springer; 2018. p. 381–386.

[21] Fournier-Viger P, Gomariz A, Gueniche T, Soltani A, Wu CW, Tseng VS sequential pattern miningF: a Java open-source pattern mining library. The Journal of Machine Learning Research. 2014 Jan 1;15(1): 3389-93.

[22] Sasaki K, Furukawa T, Murakami J, Shimozono H, Nagamatsu M, Miyao M, Yamamoto T, Watanabe I, Yasugahira H, Saito T, Ueno Y. Scoring profiles and defense performance analysis in Rugby Union. International Journal of Performance Analysis in Sport. 2007 Oct 1;7(3): 46-53.

Not all goals are equally important — a study for the NHL

Jon Vik*, Min-Chun Shih*, Rabnawaz Jansher*, Niklas Carlsson*, Patrick Lambrix*

Linköping University, Sweden + email address: niklas.carlsson@liu.se, patrick.lambrix@liu.se

Abstract

The evaluation of player performance is an important topic in sport analytics and is used by coaches for team management, in scouting and in sports broadcasts. When evaluating the performance of ice hockey players many metrics are used, including traditional metrics such as goals, assists, points and modern metrics such as Corsi. One weakness of such metrics is that they do not take into consideration the context in which the value for the metric was assigned. For instance, when a player scores a goal, then the value of the goals metric for that player is raised by one, regardless of the importance of the goal. In this paper, we introduce new variants of classical metrics based on the importance of the goals regarding their contribution to team wins and ties. Further, we investigate using play-by-play data from the 2013-2014 NHL season how these new metrics relate to the classical metrics and which players stand out with respect to important goals.

1 Introduction

When evaluating the performance of ice hockey players, it is most common to use metrics that attribute a value to the actions the player performs (e.g., scoring a goal for the goals metric or giving a pass that leads to a goal for the assists metric) and then compute a sum over all those actions. Some extensions to these traditional metrics have been proposed, e.g., for the +/- metric [1,7]. There is also work on combining metrics such as in [2]. Some of the approaches for player performance metrics take game context into account such as event impacts [11]. Other works model the dynamics of an ice hockey game using Markov games where two opposing sides (e.g., the home team and the away team) try to reach states in which they are rewarded (e.g., scoring a goal) [3,5,6,9,10,12–14]. An approach to predict the tier (e.g., top 10%, 25% or 50%) to which a player belongs is presented in [4].

Although some metrics take context into account for goals, e.g., the location of the shot, few take into account the importance of goals. For instance, a goal scored when the team is in the lead with 9–2 at the end of the game is most likely not crucial for winning. In contrast, scoring a goal when the score is tied at 2–2 with fifteen seconds left of the game is of more importance for winning.

Furthermore, some players have a reputation to often make important goals, while others may have the reputation to mainly score when the team is playing 'easier' games. For instance, during the 2013-2014 season the Washington Capitals' Alexander Ovechkin ranked the highest regarding game-tying and lead-taking goals while he only ranked 29th regarding goals scored when the team is already in the lead.

The importance of goals was taking into account in the added goal value metric in [8] and in this paper we introduce variants of the classical goals, points[1], assists and +/- metrics that take into account the importance of the goals.

2 Game points importance value

As a basis for our new metrics we need to formally define the importance of a goal. Our intuition is that the importance of the goal represents the change in probability of the team taking points for the game before and after the goal has been scored.[2] As we only look at regulation time, in the NHL the team can earn 2 points for a win, 1 for a tie and 0 for a loss.[3]

First, we define the probability of an outcome given a context, where outcome is one of win, tie, or loss, as the ratio of the number of occurrences of the context given the outcome and the number of occurrences of the context in our data set.

$$P(outcome \mid context) = \frac{Occ(context \mid outcome)}{Occ(context)} \tag{11}$$

In our experiments the context is defined by time (t) in one second intervals, goal differential (GD) and manpower differential (MD).

We attribute a game points importance value (GPIV) to a context. Intuitively, the GPIV represents how much a goal in a particular context increases or decreases the expected game points taking into account that a win gives 2 points, while a tie gives 1 point. When a goal is scored the context after the goal (context AG) has the same time as the context before the goal (context BG), but the GD is changed by one and the MD may (minor penalty power-play goal) or may not change (even strength, short-handed, or major penalty power-play goal).

$$\begin{aligned} GPIV(context) = 2 * [P(win \mid contextAG) - P(win \mid contextBG)] \\ +1 * [P(tie \mid contextAG) - P(tie \mid contextBG)] \end{aligned} \tag{12}$$

From Fig. 1 we note that the value of GPIV is high when the GD is between -1 and 1 at the end of the third period, as scoring then will tie the game (going from 0 to 1 game point) or result in a 1 or 2 goals lead (going from 1 to 2 points for GD = 0, or strengthening the probability of the win for GD = 1). However, as the scoring frequency in the last minute is three times higher than at any other arbitrary minute in the game (see Fig. 2), this increase in GPIV may not be as high as expected.

Scoring goals is not always positive for the probability of taking game points. We noted that taking a 2 or 3 goal lead early in the game may have negative consequences. This could be explained by the possibility of the leading team becoming too complacent with a comfortable lead. In general, negative consequences were limited to the first period or special MD cases in the beginning of the second period.

[1] Defined as the number of goals plus the number of assists for the player and often denoted by P. In this paper we also use the points a team receives for a win or a tie, which are used to produce a ranking of the teams, often denoted by PTS. To avoid confusion, we call this latter kind of points 'game points'.

[2] In [8] only the change in win probability is considered.

[3] When taking overtime into account, an extra point will be distributed to the winner in overtime for a game that was tied in regulation time. Therefore, in the NHL a team is awarded 2 points for a win (in regulation time or overtime), 1 point for a loss in overtime, and 0 points for a loss in regulation time. The distribution of points can be different in other leagues, e.g., in the SHL (Sweden) 3 points are always awarded for each game.

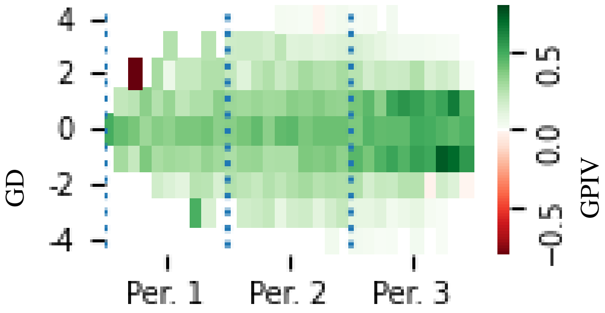


Figure 1: GPIV versus GD. Each bin is two minutes. Less than three observations for each bin are left out.

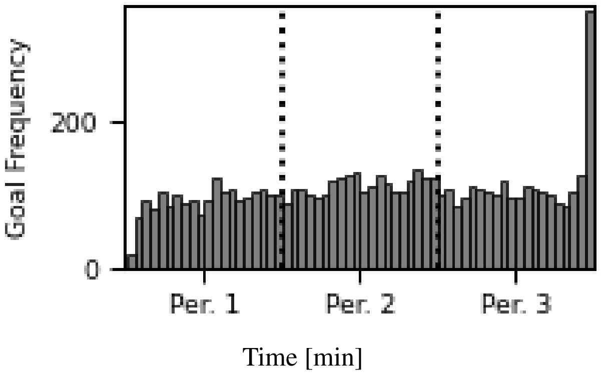


Figure 2: Goal frequency for each minute of the first three periods in the NHL during the 2013-2014 season.

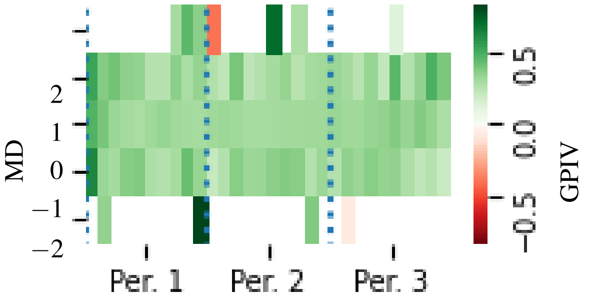


Figure 3: GPIV versus MD. Each bin is two minutes. Less than three observations for each bin are left out.

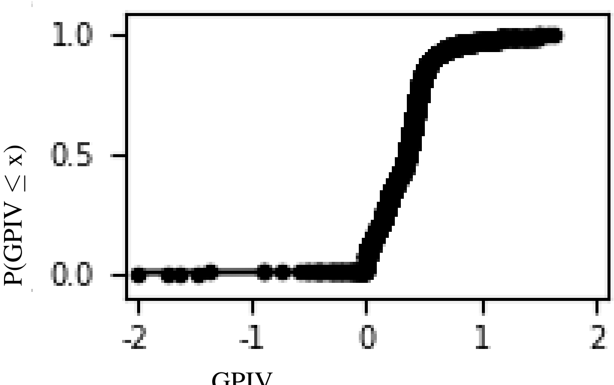


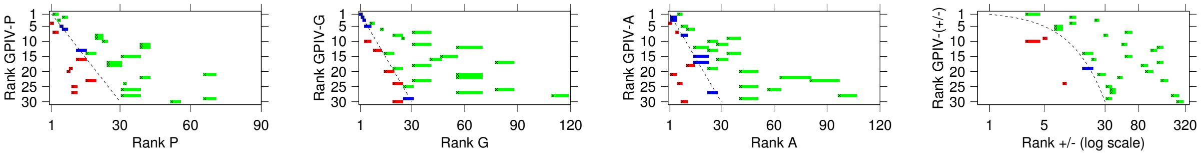
Figure 4: Cumulative distribution function of GPIV.

In contrast to GD, MD does not seem to have as much influence on the GPIV, except for some goal scoring with MD = 2 or -2 (See Fig. 3).

In Fig. 4 we see that the probability of a negative GPIV is 0.02. Nearly 82% of all GPIV range between 0 and 0.5. Further, 18% of the GPIVs range from 0.5 to 1.64. What is interesting with this last group is that they have the same or greater GPIV (0.5) as typical game deciding goals scored in overtime (which results in the team directly being awarded an extra point instead of - on average - getting the extra point with probability 0.5).

3 GPIV-weighted performance metrics

We define new variants of the classical metrics goals (G), assists (A), points (P) and +/- which we call GPIV-G, GPIV-A, GPIV-P and GPIV-+/-, respectively. In the classical metrics the value is raised by 1 when a goal is scored (for G and P), an assist is giving to a goal (for A and P) or the player is on the ice when a goal is scored (for +/-). For the latter when a goal is scored by the opposing team the value is decreased by 1. For the



(a) Points (P)

(b) Goals (G)

(c) Assist (A)

(d) Plus-minus (+/-)

Figure 5: Rank comparisons. Here, colors are used to show players that see improved (green), similar (blue), and reduced (red) ranking when using the weighted metrics.

variants of the metrics, instead of raising or decreasing by 1, we raise or decrease the value by the GPIV of the goal. The new metrics value the amount of goals as well as the importance of goals. Some of the highest ranked players are involved in many goals, while others may be involved in fewer goals, but with higher importance.

One way to compare the classical metrics and their new variants is to compute their correlations. For P and GPIV-P the maximal information coefficient is 0.765, the Pearson correlation coefficient is 0.944 and Spearman's rank correlation coefficient is 0.949. For the correlations between +/- and GPIV-+/- the values were 0.384, 0.769, and 0.750, respectively. The much weaker correlation for the +/- metrics is also illustrated in Fig. 5. Here, we use colors to show the top 30 players according to the GPIV-based metrics that see increased, same, or reduced rank with the GPIV-based metrics compared to the classical metrics.

Another way to check whether metrics are reasonable is to perform the eye test. Looking closer at the results[4], several players stand out. First, Alex Ovechkin went from a rank of tied for 6-7 (P) to being ranked 2nd (GPIV-P) when using the weighted points. This is a considerable difference in rank, but can be explained by the many important goals he scored that season. For example, as mentioned already in the introduction, Alexander Ovechkin had the most game-tying and lead-taking goals while he only ranked 29th regarding goals scored when the team is already in the lead.

Other players on the top-10 list that saw significant increases in their relative point-based rankings where Blake Wheeler (Winnipeg Jets), Anze Kopitar (LA Kings), and Eric Staal (Caroline Hurricanes). Similar to Alexander Ovechkin, the last two of these are players that have proven they can take their game to the next level during the playoffs (when goals are tougher to get by and each goal is typically considered of greater value). For example, these three players have all won the Stanley Cup (Ovechkin 2018, Kopitar 2012 and 2014, and Staal 2006) and all had the most points or goals during the playoffs of all players in the league during the years they won the Stanley Cup. Furthermore, all four these players are or have been captains of their respective teams (including Wheeler).

In general, we see many Stanley Cup winners on the top-10 list (8 out of 10), as only Pavelski (rank 3) and Wheeler (rank 8) have not won the Stanley Cup. However, both these players have been known for their high compete level and are both considered game changing players.

A closer look at the top-30 lists for the GPIV-based points, goals, assist, and +/- metrics reveals many other names that saw substantial increases in their relative rankings. In most cases, these players can typically

[4] Tab. 1 shows the top 10 players with respect to GPIV-P. The complete results for the 2013-2014 season for GPIV-G, GPIV-A, GPIV-P and GPIV-+/- are available at `https://www.ida.liu.se/research/sportsanalytics/projects/conferences/MathSport-21/`.

Table 1: Top-10 players according to GPIV-P with rank according to traditional points (P-rank), rank according to GPIV-weighted points (GPIV-P-Rank), the difference between these ranks (Rank-diff), the player name (Player) and position (Position), the points (P) and the GPIV-weighted points (GPIV-P).

P-Rank	GPIV-P-Rank	Rank-diff	Player	Position	P	GPIV-P
2-3	1	1	Sidney Crosby	C	69	25.734
6-7	2	4	Alex Ovechkin	R	64	25.085
4	3	1	Joe Pavelski	C	67	23.467
1	4	-3	Tyler Seguin	C	70	22.259
5	5	0	Phil Kessel	R	66	22.006
6-7	6	0	Ryan Getzlaf	C	64	21.366
2-3	7	-5	Corey Perry	R	69	20.803
20-22	8	12	Blake Wheeler	R	51	20.295
20-22	9	11	Anze Kopitar	C	51	19.812
23-24	10	13	Eric Staal	C	50	19.791

be labeled as players known to have seen great success in the playoffs, for being strong two-way players, or that are remembered for having been game changers for at least part of their career.

4 Conclusions

In this paper we introduced new variants of the classical metrics goals, assists, points and +/- by taking into account the context in which goals are scored. The new metrics weigh goals regarding the change in probability of obtaining game points. The new metrics pass the eye test.

For future work we will compute the newly introduced metrics for other NHL seasons. It will be interesting to see whether the observations of the 2013-2014 season regarding the new metrics will also be observed in the other seasons. We also want to see whether trends for players in the classic metrics will be followed by the new metrics.

References

[1] R Gramacy, S Jensen, and M Taddy. Estimating player contribution in hockey with regularized logistic regression. *JQAS*, 9:97–111, 2013.

[2] W Gu, K Foster, J Shang, and L Wei. A game-predicting expert system using big data and machine learning. *Expert Sys with Appl*, 130:293–305, 2019.

[3] E Kaplan, K Mongeon, and J Ryan. A Markov Model for Hockey: Manpower Differential and Win Probability Added. *INFOR*, 52(2):39–50, 2014.

[4] T Lehmus Persson, H Kozlica, N Carlsson, and P Lambrix. Prediction of tiers in the ranking of ice hockey players. In *MLSA 2020*, pages 89–100, 2020.

[5] G Liu and O Schulte. Deep reinforcement learning in ice hockey for context-aware player evaluation. In *IJCAI*, pages 3442–3448, 2018.

[6] D Ljung, N Carlsson, and P Lambrix. Player pairs valuation in ice hockey. In *MLSA 2018*, pages 82–92, 2019.

[7] B Macdonald. A Regression-Based Adjusted Plus-Minus Statistic for NHL Players. *JQAS*, 7(3):Article 4, 2011.

[8] S Pettigrew. Assessing the offensive productivity of NHL players using in-game win probabilities. In *MIT Sloan Sports Analytics Conference*, 2015.

[9] K Routley and O Schulte. A Markov Game Model for Valuing Player Actions in Ice Hockey. In *UAI*, pages 782–791, 2015.

[10] C Sans Fuentes, N Carlsson, and P Lambrix. Player impact measures for scoring in ice hockey. In *MathSport*, pages 307–317, 2019.

[11] M Schuckers and J Curro. Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. In *MIT Sloan Sports Analytics Conference*, 2013.

[12] O Schulte, M Khademi, S Gholami, Z Zhao, M Javan, and P Desaulniers. A Markov Game model for valuing actions, locations, and team performance in ice hockey. *Data Min Knowl Disc*, 31(6):1735–1757, 2017.

[13] O Schulte, Z Zhao, M Javan, and P Desaulniers. Apples-to-apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact. In *MIT Sloan Sports Analytics Conference*, 2017.

[14] A Thomas, S Ventura, S Jensen, and S Ma. Competing Process Hazard Function Models for Player Ratings in Ice Hockey. *Ann Appl Stat*, 7(3):1497–1524, 2013.

Sports scheduling and managerial aspects: insights for Argentina's National Basketball League

Nicolás García Aramouni*, Juan José Miranda Bront**
*Universidad Torcuato Di Tella, Buenos Aires, Argentina
**Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

Abstract

A competition's structure and a league's schedule represent key strategic decisions from a managerial standpoint. Argentina's National Basketball League (LNB) has undergone a major transformation since 2014, following a schedule design resembling the National Basketball Association (NBA), reducing total distance traveled using a variant of the Traveling Tournament Problem (TTP). In this work, we consider an alternative league design, with a time-constrained schedule similar to the existing one before 2014, and revisit some well known model that incorporates the matches interest distribution throughout the tournament. For six LNB's seasons, we observe that our approach reduces the overall distance traveled in several seasons and that under moderate assumptions regarding stadium attendance our model translates into higher revenue.
Keywords: sports scheduling; integer programming; tournament design; game importance

1 Introduction

From a managerial standpoint, a competition's structure and the design of a league's schedule represent key strategic decisions with a direct impact in terms of revenue and other important indicators. Current research is mainly devoted to tackle specific real-world cases and to provide methodological improvements for these particular problems. As a consequence, the literature providing algorithmic comparisons or evaluating the impact of different league structures is rather scarce (see, e.g. [3, 6]). Argentina's National Basketball League (LNB) is one of such cases, as it has undergone a sequence of major transformations since 2014. Briefly, the LNB's season moved from a 16 teams tournament divided in two conferences, playing first a regional and then a national phase in 2013-14, to a touring system in 2014-15. The number of teams has been increased to 18 and 20 teams in seasons 2014-15 and 2015-16, respectively. Due to the large number of games, starting in 2017-18 the season is a single-conference double round-robin system, where an auxiliary tournament is held before the season to compensate for the reduction in the number of games.

The process and the current tournament format are described in Duran et al. [4], resembling the National Basketball Association (NBA), and focusing on minimizing the total distance by a variant of the Traveling Tournament Problem (TTP). The change showed a positive impact in terms of distance traveled in a large country as Argentina. In this new problem, tours are not extensively explored and, instead, the teams are requested to provide for the optimization a set of both preferred as well as less desirable tours, from which essentially the schedule for each team is constructed.

After a few years implementing this new format, we highlight two observations that have been raised by some teams and relevant stakeholders. One refers to the lack of regularity in the tournament structure, as the new format moved from playing on weekends to a time-relaxed format, where games can be scheduled any day of the week. In addition, the stadium attendance seems to be lower during weekdays and during small periods with a high concentration of games. As one may expect, the relative local importance of the LNB is far from the NBA. Thus, focusing only on the distance minimization can overlook the business side of the different leagues, only concentrating on logistics efficiency.

By design, the efficiency of the NBA schedule relies not only on the touring system, but also on having more matches among teams from the the same division and conference with respect to the remaining teams. In this work, we build upon the experience by Duran et al. [4] by considering an alternative league design, with a time-constrained structure, concentrating matches in weekends and having a higher number of matchups among teams that are closer to each other. We consider some well known techniques, such as coupled-based tournaments, and revisit some known timetabling models with a new objective function that incorporates the matches interest distribution throughout the tournament and a re-organization of the league, similar to the existing one before 2014. We conduct extensive computational experiments comparing different structures for each season. Our work is somehow aligned with the ideas explored in Goosens et al. [5], although we follow a different methodology. Evaluating six LNB's seasons, we show that our approach reduces the overall distance traveled in four out of the six seasons, with an average reduction of about 12%. Furthermore, we show that under moderate assumptions regarding stadium attendance our model translates into higher revenue, with increments ranging from 18 to 45 percent depending on the season.

2 Model

We first introduce some basic definitions and notation used throughout the paper to model the problem. Let $N = \{1,\dots,2n\}$ bet the set of teams, and $P = \{1,2,...,n\}$ the set of couples defined for the tournament. To reflect the separation into conferences, we consider the partition $P = C_N \cup C_S$, where C_S and C_N denote the couples in the southern and northern conferences, respectively. For the ease of exposition, we assume n to be even as well, although this is not always the case in our experiments and the models are slightly adjusted. To model the tournament, let $WE = \{0,1,\dots,m,m+1\}$ be the set of time-slots (i.e., weekends) where the games take place, where 0 and $m+1$ are two artificial slots indicating the start and the end of the tournament.

For a double round-robin tournament, we simply set $m = n$ since two games are considered for every time-slot, with the only exception for the weekends assigned for the intra-couple matches. We further define $F = \{1,2,...,m\}$ as the subset of time-slots dedicated to the tournament, and let $F = F_1 \cup F_2$, with $F_1 = \{1,\dots,r\}$ and $F_2 = \{r+1,\dots,m\}$ indicating the two rounds of the tournament. For simplicity, intra-couple matches dates are assumed to be fixed and scheduled in time-slots $I = \{1,r+1\}$, i.e. at the beginning of each round, and the set of weekends with inter-pair matches, that is, $E = F/I$. In case n is odd, $I = \emptyset$ and the intra-couples games are not gropued and left as a decision.

2.1 Proposed league structure and metrics

Our proposal considers a 2-stage tournament organized as follows: (i) a regional stage where teams of the same conference play a double round-robin schedule; and (ii) a national stage where all teams play a single round-robin schedule. We re-introduce the couples system and generate a more regular schedule, restricting matches to take place only on weekends. Additionally, standings and the classification to playoffs could be organized by conference, or a unique standing could be also generated by resorting to more sophisticated indicators, but this aspect is out of scope from our study. Overall, the general structure would be almost identical to the one that existed prior to the 2014-15 season, specially for the regular season. However, for the seasons prior to 2017-18 we present a double-round robin for the national stage to keep the total number of matches the same and make a fair comparison.

If games are only held on weekends, it is reasonable to assume that a setting like the TTP may not suitable as the total traveled is fixed under the assumption that teams return home between consecutive weekends. For instance, if a game is held on a Sunday and the next one is held on a Friday, teams may return to their home in-between. In the context of the current touring system, we implicitly introduce tours of exactly two matches by design, not only reducing the computational cost but also affecting the total distance traveled, which will not be explicitly minimized in our model. As couples are composed by nearby teams, when a couple visits another one in a particular weekend, the distance traveled between matches should be close to the one that should be produced by a potential tour in the TTP model.[1] From our perspective, it is important to measure the quality of a schedule in several other dimensions besides the distance traveled, such as: business impact, reflected by expected level of profits; structure of the competition, measured by the distribution of interesting games throughout the tournament; fairness, measured by the number and the distribution of breaks among teams; among others. We focus on the first two explicitly, but also consider the distance in our analysis.

Let $\rho_{ij} \in \{0,1\}$ be a constant taking value 1 if teams i and j are considered *rivals*, and 0 otherwise. In addition, assume available as input a *ranking* $\nu \in \mathbb{N}^n$, providing an ordering for the teams. We define the *level of interest* for a game played by teams $i, j \in N$, denoted by ψ_{ij}, as

$$\psi_{ij} = \frac{1+\rho_{ij}}{1+|\nu_i - \nu_j|}. \tag{13}$$

The rationale for this metric is as follows. First, if the two teams are rivals the interest of the game increases. These rivalries could be defined by the so-called derby, i.e. teams belonging to the same city or region, or sometimes simply by a common history between the teams. Note that a team may have more than one rival. Second, the interest of a game is positively influenced by the closeness of the teams in the ranking ν. A first argument would be that teams with a similar ranking have a similar strength, and thus the game between the teams could be close when reaching the *clutch*. Assume ν represents a good estimation of the final position of the teams in the tournament. Clearly, the games between teams having the highest ranks are interesting regarding the definition of the tournament, and possibly their position regarding the playoffs. On the other hand, games between teams with the lowest ranks usually define who moves down to the second division. Finally, teams with mid-rankings usually struggle to get the last places into international competitions or the tournament playoffs. For simplicity, we consider the final standings of the previous season as a proxy for ν.

In order to incorporate this metric in our model, we extend the definition to couples. Let $\psi_{AB}^{\text{couple}}$ be the level of interest when couple $A = (i, j)$ plays against couple $B = (k,l)$, involving teams $i, j, k, l \in N$, which is simply aggregating the the interest of the four games, i.e.

$$\psi_{AB}^{\text{couple}} = \psi_{ik} + \psi_{il} + \psi_{jk} + \psi_{jl}. \tag{14}$$

When needed, we extend this idea for an intra-couple match by defining $\psi_{AA}^{\text{couple}} = 4 \times \psi_{ij}$, for $A = (i, j)$. This definition is to avoid defining the intra-couple weekend as the less interesting weekend will be the one with

[1] For the sake of consistency, we apply the same methodology for all seasons from 2012-13 to 2018-19, except for season 2014-15 where the number of resulting couples is odd.

only intra-couple matches, simply because it has a lower number of matches. In addition, this also avoids trivial solutions.

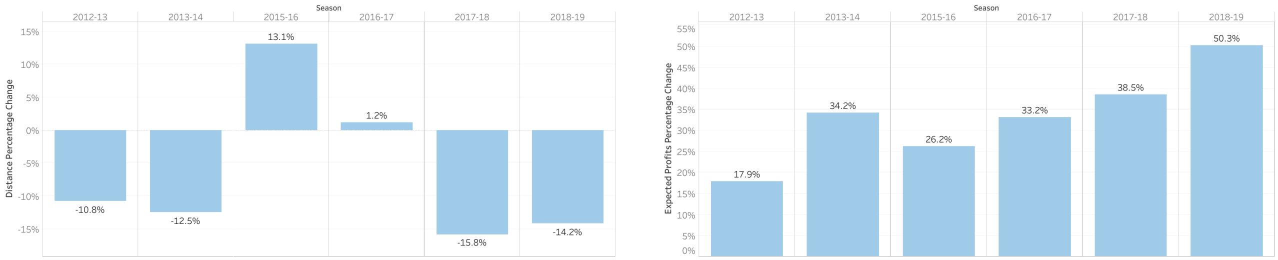
2.2 ILP formulation

We generate the schedule for the teams by simply formulating an ILP model for both stages simultaneously. We consider an objective function to maximize interest in the last matchdays of the tournament for the regional stage. In this setup, the optimization problem can be decomposed into two independent problems, one per conference. Thus, let $C = C_N, C_S$ be a generic conference.

Let c_{ijt} denote the interest when couple $i \in C$ faces couple $j \in C$ in time-slot t. Based on our definition of interest, we set $c_{ijt} = t \times \psi_{ij}^{\text{couple}}$. Let further $\hat{E} \subseteq E \times E \times E$ be the set of triplets of three consecutive increasing time-slots in E, which is used to limit the number of home and away games. We define binary variables x_{ijt} take value 1 if couple $i \in P$ plays at home against couple $j \in P$ at time-slot $t \in F$.

$$
\begin{aligned}
\max \; & \sum_{i \in P} \sum_{j \in P} \sum_{t \in F} c_{ijt} x_{ijt} && && (15)\\
\text{s.t. } & x_{iit} = 1 && \forall i \in P, \forall t \in I && (16)\\
& x_{iit} = 1 && \forall i \in P, \forall t = 0, m+1 && (17)\\
& \sum_{j \in P} x_{ijt} + \sum_{j \in P} x_{jit} = 1 && \forall i \in P, \forall t \in WE && (18)\\
& \sum_{j \in P} x_{ijt} + \sum_{j \in P} x_{ijt_1} + \sum_{j \in P} x_{ijt_2} \leq 2 && \forall i \in P, \forall (t, t_1, t_2) \in \hat{E} && (19)\\
& \sum_{j \in P} x_{jit} + \sum_{j \in P} x_{jit_1} + \sum_{j \in P} x_{jit_2} \leq 2 && \forall i \in P, \forall (t, t_1, t_2) \in \hat{E} && (20)\\
& \sum_{t \in F_1} x_{ijt} + \sum_{t \in F_1} x_{jit} = 1 && \forall i, j \in C, i \neq j && (21)\\
& \sum_{t \in F} x_{ijt} + \sum_{t \in F} x_{jit} = 2 && \forall i, j \in C, i \neq j && (22)\\
& \sum_{t \in WE} x_{ijt} = 1 && \forall i, j \in C, i \neq j && (23)\\
& x_{ijt} \in \{0, 1\} && i, j \in P, t \in W && (24)
\end{aligned}
$$

The objective function (15) maximizes the aggregated interest, assigning more weight to games scheduled closer to the end of the season. Constraints (16) fix intra-couple matches, while constraints (17) make teams to be home both at the beginning and the end of the tournament. Constraints (18) limit couples to have one assignment per weekend. Constraints (19) and (20) restrict the number of consecutive weekends playing at home or away matches to two for each team, respectively. These constraints are based on the ideas proposed by Bonomo et al. [1] for the National Volleyball League in Argentina. Note that they impose shared conditions between the two rounds. Moreover, equations (21) force teams play once against each other in the first round, while equations (22) make teams play twice against each other in the entire tournament. Constraints (23) force the home condition of a particular match is seen only once. Finally, (24) define the domain of the variables.



Perc. change in total distance compared to LNB. Perc. change in expected profits compared to LNB.

Figure 1: Perc. change in total distance and expected profits, relative to LNB, by season

This formulation resembles the one proposed by Briskorn and Drexl [2], with the same generic objective function with weights c_{ijk}. The key differences lie in the specific definition of these weights, as well as in the constraints. While Briskorn and Drexl [2] force the schedule to have the minimum number of breaks (i.e. $|C|-2$), we tackle this via constraints (19) and (20). Needless to say, this model generates a couple-level schedule, which can be easily translated into a team-level schedule. In addition, the model is slightly modified for the national stage (where the teams face each other once). Recall that the LNB's tournament format changed during the period considered, and thus the overall number matchups between teams depends on the season. Due to space limitations, we omit the details.

3 Preliminary experimental results

We conducted computational experiments to analyze the differences between the the tournament designed following the ideas presented in Section 2 and the original LNB fixture executed each year, designed following the method proposed in [4]. Regarding the metrics, we had access to the distance matrix for the instances reported in [4]. For the remaining seasons and new teams, distances are computed using Google Maps. The model is implemented using Python 3 and CPLEX as an ILP solver. To compute the couples within each conference, we resort to a minimum weight matching algorithm. The general structure proposed in Section 2 is slightly adjusted depending on the season to reflect the changes introduced each year, resulting in the following setup:

- Seasons 2012-13, 2013-14, 2015-16, 2016-17: four games against teams of the same conference and two against teams of the other conference
- Seasons 2017-18, 2018-19: four games against the team of the same couple, three games against teams of the same conference and one against teams of the other conference

Recall that the Season 2014-15 is not considered as the number of couples would be odd. Thus, in our analysis we only modify the number of matches per rival only for seasons 2017-18 and 2018-19, although the total number of games each team plays remains unchanged. However, all the 6 seasons analyzed share that there are more matches among teams from the same conference.

Regarding the profit analysis, we make some general assumptions based on news articles and private communications with some of the teams. We use as a reference values as of March, 2019, and consider a ticket price of 200 ARS and a fixed cost of ARS 50,000 per match related to the stadium and staff.[2]. To account for the inflation, we used price indexes and the dollar/peso exchange rate in order to compute the profits in nominal US Dollars for each season. The other cost considered relates to the trips incurred by each team, assuming a cost of USD 2 per kilometre as reported by [4].

To the best of our knowledge, there are no formal records for the attendance at each game in the LNB. Then, again we make some moderate assumptions based on informal conversations with several teams in the LNB. Approximately 15% of teams claimed to prefer the current format, and that the average occupancy in the stadium was about 75%. Thus, we randomly selected 15% of the teams and assumed this level of attendance. For the two most popular teams, we assumed a 100% occupancy at every match. For the remaining teams, we consider a 50% occupancy on weekends (Fridays, Saturdays and Sundays) and a 30% occupancy on weekdays, as reported by one of the teams consulted.

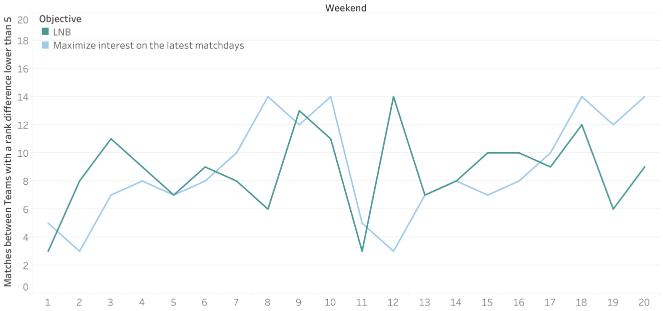


Figure 2: Games between teams with up to 5 positions of difference - National Stage, 2015-16 Season

Figure 1 shows the differences between our approach relative to the original LNB schedule, expressed in percentages, for distance (left) and profit (right). There are different drivers that explain the difference between LNB's and our distance. For seasons 2012-13 and 2013-14 the reduction in distance seems to come from constructing the couples using MIP instead of manually. For seasons 2015-16 and 2016-17, the weekly tours seem to have a positive impact in reducing the number of long trips, although it seems to be marginal in the latter. The most interesting results are for seasons 2017-18 and 2018-19, where the overall distance decreases by changing the mix of games each team plays. While LNB's schedule makes teams play the same number of against every other team, we propose that every team should play more games with rivals of the same conference and discard tours. Regarding the profits, our model brings up consistently more profits. The reduction in the logistic costs translates in significant savings, but we also observe an increase in the overall revenue due to a greater share of matches played on weekends, specially since the change in the schedule structure that began in the 2015-16 season.

Finally, we report in Figure 2 the number of matches between teams ranked close to each other (with less than 5 positions of difference) for every weekend of the National Stage of the 2015-16 season. We can observe that our approach, even when constructing the schedule using couples, is capable of obtaining a desirable distribution of games when compared with LNB's original season, with interesting matches between teams that should struggle to avoid relegation (teams in the bottom places of the standings), teams that compete to

[2] Cardone, Matías, "La devaluación llegó a la Liga Nacional de básquet". Last access: May 16th, 2021)

qualify to international competitions or the playoffs (teams in the middle of the standings) or teams within the top spots (first places of the standings). Despite the particular metric, we find interesting to include this kind of quality metrics explicitly in the design of the fixture to capture other business and organizational objectives besides the total distance traveled.

References

[1] F. Bonomo, A. Cardemil, G. Durán, J. Marenco, and D. Sabán. An application of the traveling tournament problem: the argentine volleyball league. *Interfaces*, 42(3):245–259, 2012.

[2] D. Briskorn and A. Drexl. A branch-and-price algorithm for scheduling sport leagues. *Journal of the Operational Research Society*, 60(1):84–93, 2009.

[3] G. Durán. Sports scheduling and other topics in sports analytics: a survey with special reference to latin america. *TOP*, 29(1):125–155, 2021.

[4] G. Durán, S. Durán, J. Marenco, F. Mascialino, and P. A. Rey. Scheduling Argentina's professional basketball leagues: A variation on the Travelling Tournament Problem. *European Journal of Operational Research*, 275(3):1126–1138, 2019.

[5] D. R. Goossens, J. Beliën, and F. C. Spieksma. Comparing league formats with respect to match importance in belgian football. *Annals of Operations Research*, 194(1):223–240, 2012.

[6] G. Kendall, S. Knust, C. C. Ribeiro, and S. Urrutia. Scheduling in sports: An annotated bibliography. *Computers and Operations Research*, 37(1):1–19, 2010.

Analysing the restricted assignment problem of the group draw in sports tournaments

László Csató*
*Institute for Computer Science and Control (SZTAKI),
Eötvös Loránd Research Network (ELKH),
Laboratory on Engineering and Management Intelligence,
Research Group of Operations Research and Decision Systems,
Corvinus University of Budapest (BCE),
Department of Operations Research and Actuarial Sciences
email address: laszlo.csato@sztaki.hu

Abstract

Many sports tournaments contain a group stage where the allocation of teams is subject to some constraints. The standard draw procedure extracts the teams from pots sequentially and places them in the first available group in alphabetical order such that at least one assignment of the teams still to be drawn remains acceptable. We show how this mechanism is connected to generating permutations and provide a backtracking algorithm to find the solution for any given sequence. The consequences of draw restrictions are investigated through the case study of the European Qualifiers for the 2022 FIFA World Cup. We quantify the departure of its draw procedure from even distribution and propose two alternative approaches to increase the excitement of the draw.

Keywords: Assignment problem; Backtracking algorithm; Football; Mechanism design; Permutation.

AMS: 05A05, 68U20, 68W40, 91B14

JEL codes: C44, C63, Z20

The full version of the paper is available at `https://arxiv.org/abs/2103.11353`.

1 Introduction

Mechanism design usually focuses on theoretical properties like efficiency, fairness, and incentive compatibility [Abdulkadiroğlu and Sönmez, 2003, Roth et al., 2004, Csató, 2021b]. On the other hand, institutions—like governing bodies in major sports—often emphasise simplicity and transparency, which calls for a comprehensive review of how these procedures that exist in the real world perform with respect to the above requirements.

We offer such an analysis of a mechanism used to solve a complex assignment problem. Several sports tournaments are organised with a group stage where the teams are assigned to groups subject to some rules. This is implemented by a draw system that satisfies the established criteria.

In particular, we analyse the draw procedure of the Union of European Football Associations (UEFA), applied for various competitions of national teams such as the UEFA Nations League [UEFA, 2020b], the UEFA Euro qualifying [UEFA, 2018], or the European Qualifiers for the FIFA World Cup [UEFA, 2020a]. The mechanism works as follows to generate a sense of excitement and to ensure transparency. First, the teams are divided into seeding pots based on an exogenous ranking. For each pot, balls representing the teams are placed in a bawl and drawn randomly. The teams are assigned to the groups in alphabetical order, i.e. the first team drawn from each pot is allocated to the first group, the second team to the second group, and so forth. However, there are draw constraints to provide an assignment "*that is fair for the participating teams, fulfils the expectations of commercial partners and ensures with a high degree of probability that the fixture can take place as scheduled*" [UEFA, 2020a]. Consequently, if a draw restriction applies or is anticipated to apply, the actual team is allotted to the first available group as indicated by a computer program in order to avoid any dead end, a situation when the teams still to be drawn cannot be assigned to the remaining empty slots.

This procedure is not so simple as intuition suggests.

Example 1. Consider the European Qualifiers for the 2022 FIFA World Cup. Assume that Pots 1–4 are already emptied and Group H consists of Portugal from Pot 1, Ukraine from Pot 2, Iceland from Pot 3, and Serbia from Pot 4. The draw continues with Pot 5. First, Armenia is drawn and allotted to Group A. Second, Cyprus is drawn and assigned to Group B. Third, Andorra—a country without any draw constraints—is drawn and placed in… Group H, the first available group according to the computer.

Example 1 uncovers that the number of options available to a team depends not only on its own attributes but also on the characteristics of the remaining teams: Andorra can be allocated only to Group H, otherwise, no feasible assignment exists.

The mechanism described above is used generally to draw groups in the presence of some constraints. Nonetheless, UEFA does provide neither an exact algorithm to determine the group allocation for a given random order of the teams, nor an analysis on the effects of the particular conditions. Our work aims to fill this research gap.

The main contributions can be summarised as follows:

- We highlight how the restricted group assignment problem is linked to generating all permutations of a sequence [Csató, 2021a, Section 3.1];
- We present a backtracking algorithm to produce the group allocation for teams drawn randomly, which also finds the first available group for the team drawn [Csató, 2021a, Section 3.2];

- We reveal the implications of the draw constraints in the case of the European Qualifiers for the 2022 FIFA World Cup [Csató, 2021a, Section 4.2];
- We quantify the departure of the UEFA draw procedure from the "evenly distributed" system in this particular tournament [Csató, 2021a, Section 4.3];
- We propose two alternative approaches for solving the group assignment problem to increase uncertainty during the draw.

Group allocation is an extensively discussed topic in the mainstream media. Several articles published in famous dailies such as *Le Monde* and *The New York Times* illustrate the significant public interest in the FIFA World Cup draw [Aisch and Leonhardt, 2014, Guyon, 2014, Guyon, 2017, McMahon, 2013], as well as in the UEFA Champions League group round draw [Guyon, 2020a] and the Champions League knockout stage draw [Guyon, 2020b]. Thus a better understanding of these draw procedures and their consequences is relevant not only for the academic community but for sports administrators and football fans around the world.

2 Literature review

Several scientific works focus on the FIFA World Cup draw. Before the 2018 edition, the host nation and the strongest teams were assigned to different groups, while the remaining teams were drawn randomly with maximising geographic separation: countries from the same continent (except for Europe) could not have played in the same group and at most two European teams could have been in the same group.

For the 1990 FIFA World Cup, [Jones, 1990] shows that the draw was not mathematically fair. For example, West Germany would be up against a South American team with a probability of 4/5 instead of 1/2—as it should have been—due to the incorrect consideration of the constraints. Similarly, the host Germany was likely to play in a difficult group in the 2006 edition, but other seeded teams, such as Italy, were not [Rathgeber and Rathgeber, 2007].

[Guyon, 2015] identifies severe shortcomings of the procedure used for the 2014 FIFA World Cup draw: imbalance (the eight groups are at different competitive levels), unfairness (certain teams have a greater chance to end up in a tough group), and uneven distribution (the feasible allocations are not equally likely). The paper also discusses alternative proposals to retain the practicalities of the draw but improve its outcome.

[Laliena and López, 2019] develop two evenly distributed designs for the group round draw with geographical restrictions that produce groups having similar (or equal) competitive levels.

[Cea et al., 2020] analyse the deficiencies of the 2014 FIFA World Cup draw and give a mixed integer linear programming model to create groups. The suggested method takes into account draw restrictions and aims to balance "quality" across the groups.

Other studies deal with the UEFA Champions League, the most prestigious association football (henceforth football) club competition around the world. [Klößner and Becker, 2013] investigate the procedure to determine the matches in the round of 16, where eight group winners should be paired with eight runners-up. There are $8! = 40{,}320$ possible outcomes depending on the order of runners-up, but clubs from the same group or country cannot face each other, and the group constraint reduces the number of feasible solutions to 14,833. The draw system is proved to inherently imply different probabilities for certain assignments, which are translated into more than ten thousand Euros in expected revenue due to the substantial amount of prize money. Finally, the authors propose a better suited mechanism for the draw.

Analogously, [Boczoń and Wilson, 2018] examine the matching problem in the knockout phase of this tournament. The number of valid assignments is found to be ranged from 2,988 (2008/09 season) through 6,304 (2010/11) to 9,200 (2005/06), determined by the same-nation exclusion that varies across the years. It is analysed how the UEFA procedure affects expected assignments and addresses the normative question of whether a fairer randomisation mechanism exists. They conclude that the current design comes quantitatively close to a constrained best in fairness terms.

To summarise, the previous academic literature of constrained matching mechanisms for sports tournaments mostly discusses either the FIFA World Cup draw or the UEFA Champions League knockout phase draw. Both problems are simpler than the one discussed here. The World Cup draw does not require backtracking as the group skipping policy could not lead to impossibility [Jones, 1990, Guyon, 2015]. Even though dead ends should be avoided in the knockout stage of the Champions League, only 16 teams need to be paired, thus the number of feasible solutions remains tractable and the complexity of backtracking is more limited compared to Example 1.

3 Discussion

The draw mechanism of the European Qualifiers for the 2022 FIFA World Cup is found to be somewhat biased but it remains close to the principle of equal treatment [Csató, 2021a]. Its sporting effects are ambiguous and insignificant with respect to qualification. Therefore, in contrast to the conclusion of [Klößner and Becker, 2013], the UEFA procedure does not lead to substantial financial differences here. The chosen implementation seems to be a reasonable compromise until the draw constraints do not exclude too many assignments.

Previous studies have made several recommendations to create (more) evenly distributed groups. However, these proposals usually use a fundamentally novel approach and/or are less interesting for fans to watch, hence they are unlikely to be applied soon.

Nonetheless, the current policy has another shortcoming as it might lead to a deterministic assignment too early, which is detrimental to the excitement of the draw.

Example 2. Assume that there are four groups A–D and four teams $T1$–$T4$ drawn in this order, while $T3$ cannot be assigned to group C. After team $T1$ is placed in group A and $T2$ in group B, uncertainty entirely disappears since $T3$ should play in group D and $T4$ in group C.

The problem of Example 2 is caused by the principle that the team drawn is reassigned from the first available group in alphabetical order only if no permutation of the remaining teams satisfies the draw constraints. Therefore, two alternative policies are given to increase uncertainty during the draw.

Definition 3.1. Mechanism A*: If the team drawn can be allocated to a group such that at least two feasible assignments of the remaining teams to the empty slots exist, then it is placed in the first available group with this property in alphabetical order. Otherwise, the team drawn is placed in the first available group in alphabetical order.*

Mechanism A retains at least two valid group assignments as long as possible.

Definition 3.2. Mechanism B*: The team drawn is allocated to the first available group in alphabetical order where the highest number of the remaining teams in its pot cannot play.*

Mechanism B aims to maximise the number of acceptable assignments for the teams still to be drawn.

Example 3. Consider the situation outlined in Example 2:

- Mechanism A assigns $T1$ to group A, $T2$ to group C (otherwise, only one feasible allocation remains), $T3$ to group B, and $T4$ to group D.
- Mechanism B assigns $T1$ to group C (since $T3$ cannot be placed here), $T2$ to group A, $T3$ to group B, and $T4$ to group D.

Note that after allocating the first team $T1$, six assignments satisfy all restrictions under Mechanism B but only four under Mechanism A as $T3$ cannot be allotted to group C.

In addition, mechanism B reduces the probability of a dead end situation by filling first the groups with many draw constraints. The comparison of these procedures will be the topic of future research.

Acknowledgements

The research was supported by the MTA Premium Postdoctoral Research Program grant PPD2019-9/2019.

References

[Abdulkadiroğlu and Sönmez, 2003] Abdulkadiroğlu, A. and Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747.

[Aisch and Leonhardt, 2014] Aisch, G. and Leonhardt, D. (2014). Mexico, the World Cup's Luckiest Country. *The New York Times*. 5 June. https://www.nytimes.com/2014/06/06/upshot/mexicos-run-of-world-cup-luck-has-continued.html.

[Boczoń and Wilson, 2018] Boczoń, M. and Wilson, A. J. (2018). Goals, constraints, and public assignment: A field study of the UEFA Champions League. Technical Report 18/016, University of Pittsburgh, Kenneth P. Dietrich School of Arts and Sciences, Department of Economics. https://www.econ.pitt.edu/sites/default/files/working_papers/Working%20Paper.18.16.pdf.

[Cea et al., 2020] Cea, S., Durán, G., Guajardo, M., Sauré, D., Siebert, J., and Zamorano, G. (2020). An analytics approach to the FIFA ranking procedure and the World Cup final draw. *Annals of Operations Research*, 286(1-2):119–146.

[Csató, 2021a] Csató, L. (2021a). Analysing the restricted assignment problem of the group draw in sports tournaments. Manuscript. arXiv: 2103.11353.

[Csató, 2021b] Csató, L. (2021b). *Tournament Design: How Operations Research Can Improve Sports Rules*. Palgrave Pivots in Sports Economics. Palgrave Macmillan, Cham, Switzerland.

[Guyon, 2014] Guyon, J. (2014). A Better Way to Rank Soccer Teams in a Fairer World Cup. *The New York Times*. 13 June. https://www.nytimes.com/2014/06/14/upshot/a-better-way-to-rank-soccer-teams-in-a-fairer-world-cup.html.

[Guyon, 2015] Guyon, J. (2015). Rethinking the FIFA World Cup$^{\text{TM}}$ final draw. *Journal of Quantitative Analysis in Sports*, 11(3):169–182.

[Guyon, 2017] Guyon, J. (2017). Tirage au sort de la Coupe du monde : comment ça marche et quelles probabilités pour la France. *Le Monde*. 30 November. https://www.lemonde.fr/mondial-2018/article/2017/11/30/tirage-au-sort-de-la-coupe-du-monde-comment-ca-marche-et-quelles-probabilites-pour-la-france_ 5222713_5193650.html.

[Guyon, 2020a] Guyon, J. (2020a). Ligue des champions : Barcelone et Atlético, adversaires les plus probables pour le PSG. *Le Monde*. 1 October. https://www.lemonde.fr/football/article/2020/10/01/ligue-des-champions-barcelone-et-atletico-adversaires-les-plus-probables-pour-le-psg_ 6054370_1616938.html.

[Guyon, 2020b] Guyon, J. (2020b). Ligue des champions : Borussia M'gladbach, adversaire le plus probable du PSG en huitiéme de finale. *Le Monde*. 12 December. https://www.lemonde.fr/football/article/2020/12/12/ligue-des-champions-borussia-m-gladbach-adversaire-le-plus-probable-du-psg-en-huitieme-de-fin 6063134_1616938.html.

[Jones, 1990] Jones, M. C. (1990). The World Cup draw's flaws. *The Mathematical Gazette*, 74(470):335–338.

[Klößner and Becker, 2013] Klößner, S. and Becker, M. (2013). Odd odds: The UEFA Champions League Round of 16 draw. *Journal of Quantitative Analysis in Sports*, 9(3):249–270.

[Laliena and López, 2019] Laliena, P. and López, F. J. (2019). Fair draws for group rounds in sport tournaments. *International Transactions in Operational Research*, 26(2):439–457.

[McMahon, 2013] McMahon, B. (2013). Why the FIFA 2014 World Cup Finals will be unique and very unfair. *Forbes Magazine*. 1 December. https://www.forbes.com/sites/bobbymcmahon/2013/12/01/why-the-fifa-2014-world-cup-finals-will-be-unique-and-very-unfair/?sh=76e61ea62dab.

[Rathgeber and Rathgeber, 2007] Rathgeber, A. and Rathgeber, H. (2007). Why Germany was supposed to be drawn in the group of death and why it escaped. *Chance*, 20(2):22–24.

[Roth et al., 2004] Roth, A. E., Sönmez, T., and Ünver, M. U. (2004). Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488.

[UEFA, 2018] UEFA (2018). UEFA EURO 2020 qualifying draw. 2 December. https://www.uefa.com/european-qualifiers/news/newsid=2573388.html.

[UEFA, 2020a] UEFA (2020a). FIFA World Cup 2022 Qualifying draw procedure. https://www.uefa.com/MultimediaFiles/Download/competitions/WorldCup/02/64/22/19/2642219_DOWNLOAD.pdf.

[UEFA, 2020b] UEFA (2020b). UEFA Nations League 2020/21 – league phase draw procedure. https://www.uefa.com/MultimediaFiles/Download/competitions/General/02/63/57/88/2635788_DOWNLOAD.pdf.

Analysis of soccer player's activity profiles using time-series data

Yuki Masui*, Nobuyoshi Hirotsu**, Yu Shimasaki***, and Masafumi Yoshimura****

*Juntendo University, Inzai, Chiba, Japan + email:yuki.m3240@gmail.com
**Juntendo University, Inzai, Chiba, Japan + email:nhirotsu@juntendo.ac.jp
***Juntendo University, Inzai, Chiba, Japan + email:yshimasa@juntendo.ac.jp
****Juntendo University, Inzai, Chiba, Japan = email:myoshi@juntendo.ac.jp

Abstract

The aim of this study was to classify the activity profiles of soccer players during a game based on the inertial measurement unit (IMU) time-series data measured by wearable devices utilizing a Fast Fourier Transform (FFT) method. The subject is 18 collegiate female soccer players. They were equipped with the wearable device (OptimEye S5 and G5, Catapult Sports, Australia) and measured the acceleration data during official matches. After measuring, the data was analyzed using the specific software and extracted the IMU data. The IMU data was converted by a FFT with Python to analyze the movement patterns of each player. As a result of analysis, this procedure could make it possible to identify each movement pattern appeared in the game by means of the IMA data only.

1 Introduction

Recently, with the development of technology, the measurement of physical data using wearable devices has become a common in the field of team sport coaching. Wearable devices are now widely used to understand the physical demands of soccer matches (Krustrup et al., 2005). The devices contain Global Positioning System (GPS) and can measure total distance, distances traveled within velocity bands, movement speed, trajectory, and sprinting of all players at once during training and matches. Inertial measurement units (IMU) with accelerometers, gyroscopes, and magnetometers are also included in wearable devices and can quantify the movements of non-running-based work such as acceleration, deceleration and change of direction. However, it is difficult to analyze the situations wherein quantified movements are performed in matches, because it takes a lot of time and effort. To conduct a qualitative analysis, a separate video-based analysis is required, which also takes a large amount of time and effort.

Activity profile (Suarez-Arrones et al., 2015; Varley et al., 2012) and time-motion analysis (Castellano et al., 2011) are the major research tools using wearable devices as they can measure a large number of players simultaneously and analyze them in a short time. Although the data taken with wearable devices has been used to measure the items, not many applications of machine learning methods such as deep learning to analyzing the raw time-series data has been reported so far in terms of soccer. The aim of this study was to classify the activity profiles of soccer players during a game based on the time-series data measured by wearable devices utilizing a Fast Fourier Transform (FFT) method. If detailed analysis using raw data can be conducted through this research, it will be possible to analyze the movement patterns of players from the data alone, which will enable analysis in a shorter time than the conventional method of analysis that requires considerable time and effort, and will contribute to the coaching field.

2 Method

2.1 Subjects and match data

Eighteen collegiate female soccer players (age: 20.3 ± 1.29 years; height: 161.1 ± 5.65 cm; body mass: 55.6 ± 6.24 kg) from the same college league team participated in this study. Goalkeepers were excluded from this study. Eight matches (weather: sunny or cloudy; temperature: 22.9 ± 4.45 °C; humidity: 62.7 ± 15.1 %) of the 33rd Kanto University Women Football League 2nd Division held from August 24 to November 10, 2019 were examined. The match times were a total of 90 minutes with 45-minute halves and a 15-minute interval between the halves.

2.2 Data measurement and analysis

Activity data from IMU were collected using wearable devices (OptimEye S5 and G5, Catapult Sports, Australia) operating at a sampling frequency of 100 Hz. By combining the obtained information, we could detect the inclination and direction. Thus, it was possible to measure not only the acceleration and deceleration frequencies but also the right and left movement frequencies. The IMU device has been validated in terms of high levels of reliability and low levels of measurement error (Castellano et al., 2011; Hoppe et al., 2018; Varley et al., 2012).

The players wore special harnesses that enabled these devices to be fitted to their upper backs. To extract and analyze the data of only those players who competed in the match, the match start time, the first half end time, the second half start time, and the full time were recorded based on the ratio clock. Moreover, the time of a player exiting and coming on as a substitute was regarded as her time of leaving and entering the soccer field, respectively. After recording, the data were uploaded into a PC and analyzed using the software package (Openfield version 2.2.0, Catapult Sports, Australia). After analyzing with the specific software, we extracted 100 Hz raw data and performed a FFT of the data with Python to analyze the movement patterns of each player. The items to be analyzed were the value of acceleration for the x-, y-, and z- axes. After analyzing the data, we compared the video to the data to see what movements were measured.

3 Result

In order to evaluate the analysis of players during matches based on the above models, we needed to divide 100 Hz raw data measured over 90 minutes into 3 seconds intervals. Figure 3 shows an example of the time-series acceleration data and its FFT when the player cleared the ball.

In Figure 3, it can be seen that the movement in the z-axis direction is larger than those in other directions. Figure 4 shows another example when the player sprinted.

As shown in Figures 3 and 4, we can see the pattern of the amplitude of frequency between directions. We can also see the difference of the patterns between plays appeared in the game. In addition to Figures 3

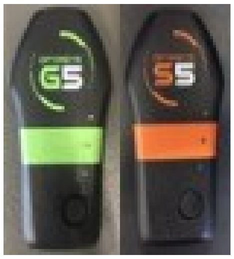


Figure 1: OptimEye G5(left), S5(right)

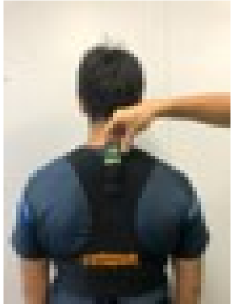


Figure 2: Attaching the wearable device

and 4, we were able to analyze four more types of the movement: contact with opposing players, running phase changing to fast speed, driving a cross and throw-in.

4 Conclusion and further study

In this paper, we have used the FFT method to analyze the movement patterns of each player from 100 Hz raw data. As a result of analysis, this procedure could make it possible to identify each movement pattern appeared in the game without watching a video.

In this paper, we have just analyzed the objective of play from 100Hz raw data using FFT. As this study is still in progress, we plan to present more in the conference. Further, in order to contribute to soccer teams that use wearable devices to measure data, it will be necessary to analyze data about every movement during a soccer game in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19K04911, Japanese Center for Research on Women in Sport of Juntendo University, and Institute of Health & Sports Science & Medicine, Juntendo University. We acknowledge the players and coaching staff of Juntendo Women's Football Club for their contribution to the study.

References

Castellano, J., Blanco-Villase?or, A., and Alvarez, D. (2011). Contextual variables and time-motion analysis in soccer. Int. J. Sports Med., 32: 415-421. doi:10.1055/s-0031-1271771.

Castellano, J., Casamichana, D., Calleja-Gonzalez, J., Roman, J. S., and Ostojic, S. M. (2011). Reliability and accuracy of 10 Hz GPS devices for short-distance exercise. J. Sports Sci. Med., 10: 233?234.

Hoppe, M. W., Baumgart, C., Polglaze, T., and Freiwald, J. (2018). Validity and reliability of GPS and LPS for measuring distances covered and sprint mechanical properties in team sports. PLOS ONE., 13: e0192708. doi:10.1371/journal.pone.0192708.

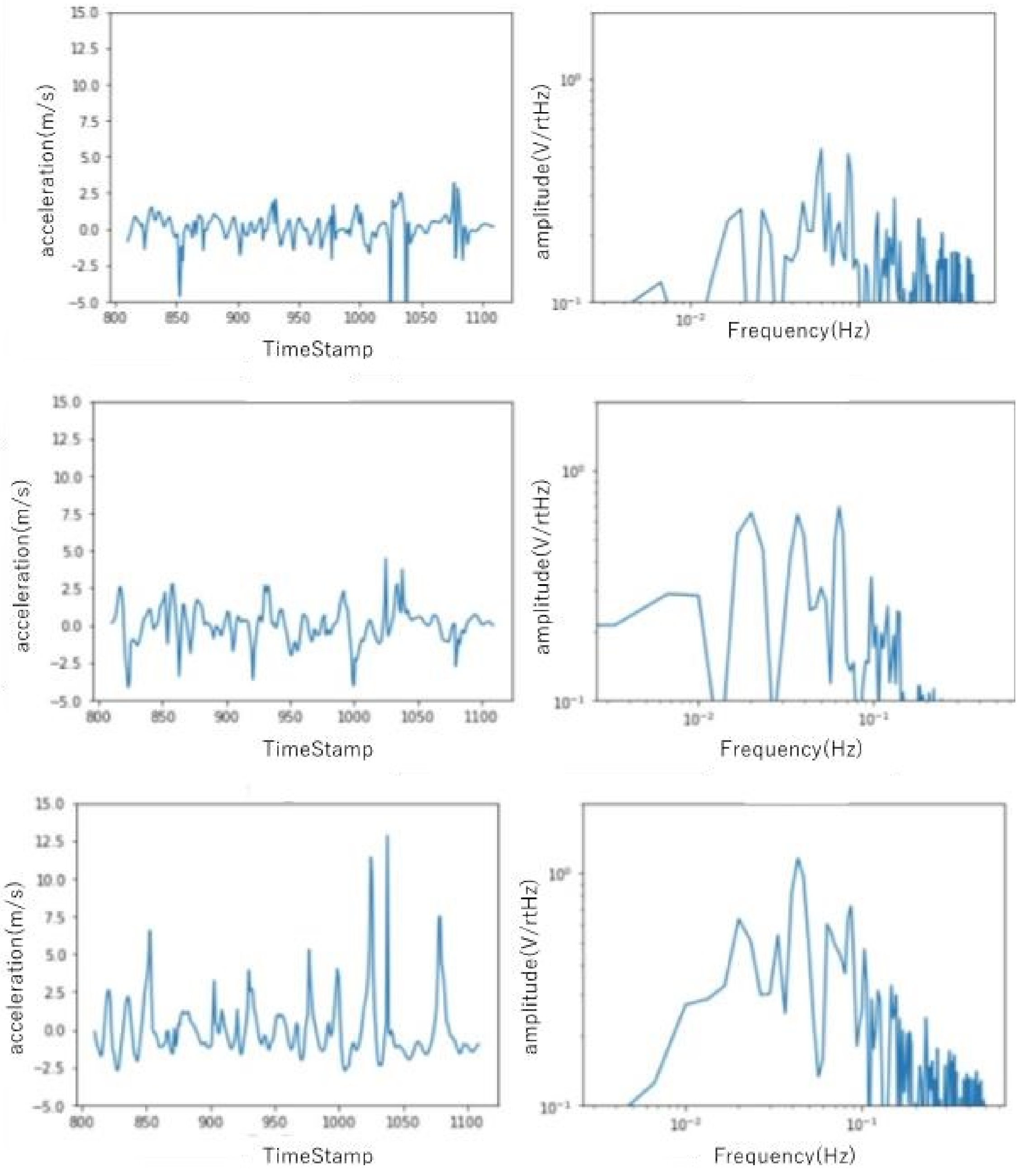


Figure 3: Acceleration of the player cleared the ball (Top: x-axes, Middle: y-axes, Bottom: z-axes)

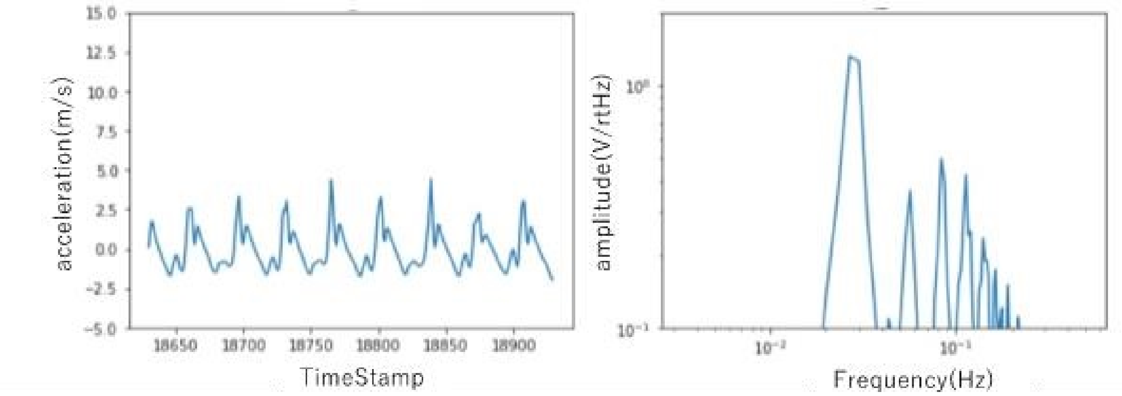


Figure 4: Acceleration of the player sprinted (z-axis direction)

Krustrup, P., Mohr, M., Ellingsgaard, H., and Bangsbo, J. (2005). Physical demands during an elite female soccer game: importance of training status. Med. Sci. Sports Exerc., 37: 1242-1248. doi:10.1249/01.mss.0000170062.73981.94.

Suarez-Arrones, L., Torreno, N., Requena, B., Saez De Villarreal, E., Casamichana, D., Barbero-Alvarez, J. C., and Munguia-Izquierdo, D. (2015). Match-play activity profile in professional soccer players during official games and the relationship between external and internal load. Journal of Sports Medicine and Physical Fitness., 55: 1417-1422.

Varley, M. C., Fairweather, I. H., and Aughey, R. J. (2012). Validity and reliability of GPS for measuring instantaneous velocity during acceleration, deceleration, and constant motion. J. Sports Sci., 30: 121?127. doi:10.1080/02640414.2011.627941.

Is there “home road” advantage in men’s professional cycling?

Michael A. Smith*

*email address: michaelsmithecon@gmail.com

Considerable research has been invested into calculating home field advantage across various sports, games, and competitions, however professional cycling has not yet come to a general consensus regarding the quantifiable advantage of riding on local roads. This article utilizes a data set of stage-by-stage results from editions of the three Grand Tours of men’s professional cycling (the Giro d’Italia, the Tour de France, and the Vuelta a España) between 1998 and 2019 to investigate the magnitude of the advantage, if any, for riders racing closer to home. Presented is an empirical model that uses the distance between the finish line of a given stage and the birthplace of a given Grand Tour racer that controls for other independent variables like race, stage number, stage distance, the number of riders crossing the finish line, and the rider themselves. In this model, birthplace is used as a proxy for what a rider might consider as their “home turf,” as birthplace is a more consistent and easily accessible variable than a rider’s current address or where their team is registered. There appears to be a sizable, statistically significant home field advantage across the Grand Tours for varying levels of proximity.

5 Literature Review

Home field advantage (also known as home court, home ice, or home team advantage, depending on the sport), in some way or another, has been found across nearly every major global sport, particularly team sports [2]. While home advantage has been decreasing over time across some of the more popular sports [5], the effect still persists in the modern era [6].

While there has been considerable focus and consistent findings regarding the presence home field advantage in team sports, there is more mixed evidence regarding home field advantage in individual sports. A literature review of home advantage in individual competitions [3] found that, with the exception of subjectively evaluated sports, the effect is much smaller than in team sports. In the few instances where home advantage may be present in objective individual sports, like the US Open golf championships or Wimbledon tennis tournament, much of the home field advantage can be explained by unequal access between foreign and domestic competitors [4]. When only the top foreign competitors are invited to participate, there is a greater chance that lower caliber domestic players can succeed beyond what their ranking might suggest.

While modern day road cycling is a team sport, where most riders work towards the success of team leaders under the direction of a team manager, racers are timed and ranked individually. Additionally, support riders who showcase their value may have a clear path to promotion to leader status on their current team or under a new contract with a different team [1] that is not always available in pure team sports.

It appears that the only empirical evidence of home field advantage within professional cycling comes from a presentation given by [7] at the MathSport Asia Conference. Using Cycling Quotient points (which are distributed to riders based on their rank in a given race, adjusted for the difficulty level and the type of race) from CQRanking.com for the years of 2011 to 2018, they found a statistically significant home country advantage for riders participating in World Tour cycling events.

This paper builds on this research on two main fronts: by adding several explanatory variables that strengthen the robustness of the underlying model and by having more granular location data that allows for proximity to be considered on the city-level instead of the country-level. However, this article is more limited in the fact that only data from Grand Tours are considered, whereas using Cycling Quotient points as a proxy for performance allows for all World Tour events to be considered.

6 Background

While the early editions of the Giro, the Tour, and the Vuelta were comprised primarily of Italian, French, and Spanish riders, nowadays the Grand Tours are truly international competitions. This development was fairly recent, as professional cycling only started to branch out from its four core roots (the three Grand Tour hosts and Belgium) in earnest in the 1980s, becoming a globalized sport in the 1990s [8]. This globalization of the sport, and the corresponding dilution of native-born riders in each race, allows for a better understanding of home field advantage by increasing the dispersion of where a rider calls home and by limiting the number of overlapping claims to "home field".

By rule of the Union Cycliste Internationale, Grand Tours must be no shorter than 15 days and no longer than 23 days from the first day of racing to the last. The days of racing can generally be broken up into six categories of stages, three of which are considered "mass start" stages (where riders all start concurrently) and three of which are considered variations of time trials (where teams/riders start at regular interval gaps from one another). See Table 1 for a breakdown of stage type by Grand Tour race between 1998 and 2019.

Unlike every other type of stage present in the Grand Tours, the time recorded for an individual rider in a team time trial is not their own, but instead the time of (usually) the fourth rider from their team to cross the finish line. Due to this disconnect, team time trials have been completely dropped from the analysis presented in this article.

7 Materials and Methods

The database provided by CyclingRanking.com contained two primary data sets. The first data set includes complete stage results for each Grand Tour race going back to 1998 or earlier. Observations are uniquely identified by a combination of race, year, stage number, stage rank, and rider ID number. Within each

Table 1 (does not include team time trials or stages without full results)
Race and Stage Type Breakdown - Grand Tours, 1998-2019

	Generic	Mountain	Prologue	Sprint	Time Trial	Total
Giro	144	140	7	124	36	451
Tour	143	138	9	123	36	449
Vuelta	151	158	3	102	36	450
Total	438	436	19	349	108	1,350

observation are variables for stage date, stage distance, stage origin, stage destination, stage type, and rider’s time. Modest data processing can create three dependent variables aside from rank: total elapsed time, time gap to stage winner, and time gap as percent of winner’s time.

Stages that were cancelled or neutralized were dropped. As noted earlier, team time trials have also been dropped from consideration. However, stages that were shortened prior to racing and that were raced to completion are included, with the updated distance.

The second data set contains the ID number, first name, last name, nationality, and place of birth for each rider who finished at least one stage of a Grand Tour. The two data sets can be easily merged together using the rider ID number. Aside from simple cleaning, this analysis takes the results and places of birth for each rider at face value.

Simply using the linear distance between a rider’s birthplace and the stage destination may undersell home field advantage. For example, it is doubtful that a rider who was born 1000 kilometers from the finish line of a stage can be considered more of a local than a rider born 1200 kilometers away. Additionally, there may be substantial returns to being a rider who was born only 25 kilometers away compared to 50 kilometers away that would not be accurately reflected in a model using linear distance alone.

Thus, this analysis uses binary interpretations of the distance between rider birthplace and stage destination. To achieve this end, this paper analyzes models with binary indicators with cutoffs of 10 and 25 kilometers. To investigate the spatial persistence of home field advantage, this paper also sparingly uses binary flags with kilometer cutoffs of 50, 100, 250, 500, 1000, 2500, and 5000. A breakdown of the number and proportion of rider-stages that fall into each of those respective buckets can be found in Table 2.

8 Model and Empirical Results

This paper employs several ways of estimating home field advantage in men’s road cycling. Since there are numerous ways to categorize success, especially during multi-stage races like the Grand Tours, the following analysis considers multiple performance indicators. As previously mentioned, the models in this article make use of four dependent variables: stage rank, time gap to the winner (in minutes), time gap as percent of the stage winner’s time, and speed (in kilometers per hour). This analysis could also be extended to include total elapsed time as a regressand, however, the lack of variability of elapsed time between riders intra-stage combined with its incredible inter-stage volatility undermines its usefulness.

Table 2
Rider-Stages Flagged as Home Field Advantage - Grand Tours, 1998-2019

Race	Count	Total	Share (%)
10 km Dummy	319	227,043	0.1
25 km Dummy	710	227,043	0.3
50 km Dummy	1,851	227,043	0.8
100 km Dummy	5,355	227,043	2.4
250 km Dummy	22,360	227,043	9.8
500 km Dummy	57,044	227,043	25.1
1000 km Dummy	122,329	227,043	53.9
2500 km Dummy	192,364	227,043	84.7
5000 km Dummy	199,188	227,043	87.7

To avoid confusion when reading the presented models, the variable distance will always refer to the distance of the stage (in kilometers) while proximity will refer to the distance between rider birthplace and the stage finish line.

As mentioned earlier, a preferable way to capture the magnitude of home field advantage may be through classifying proximity in a binary way. In this respect, riders can be categorized as “locals” or “non-locals” based on whether they were born with a certain cutoff distance from the finish line of a given stage. In order to avoid potential bias in the estimated value of home field advantage due to discretization error, this paper initially focuses on a 10 kilometer cutoff point, and then summarizes regression results for larger cutoff values.

Figure 1 paints a clear picture of home field advantage in the Grand Tours of cycling. The model estimates that a rider born within 10 kilometers of the finish line of a Grand Tour stage can be expected to place nearly 11 positions better than a rider who was not. Additionally, those riders can be expected to shave their gap to the stage winner by over a minute in absolute terms and nearly one percentage point in relative terms, on average, compared to their peers. While the estimates of home field advantage in the first three models of Figure 1 are statistically significant, the sample size of flagged rider-stages is microscopic compared to the total number of rider-stages. As seen in Table 2, less than one fifth of one percent of rider-stages are flagged by the 10 kilometer dummy variable, thus it is beneficial to look at more encompassing thresholds for the proximity dummy variable to get a fuller picture of home field advantage.

For brevity’s sake, Figure 2 attempts to accomplish this task by compressing regression results into a

Figure 1

Racing Performance and 10 Kilometer Proximity - Grand Tours, 1998-2019

	(1)	(2)	(3)	(4)
Variables	Rank	Gap (min)	Gap (%)	Speed (km/h)
10 km Dummy	-10.82***	-1.225***	-0.811***	-0.0119
	(2.474)	(0.415)	(0.163)	(0.197)
Year	0.201***	0.0586***	0.0274***	-0.00770***
	(0.0364)	(0.00611)	(0.00240)	(0.00291)
Stage #	0.0417	0.281***	0.0873***	0.00415*
	(0.0285)	(0.00478)	(0.00188)	(0.00227)
Distance (km)	-0.00231	0.0451***	-0.0151***	0.0232***
	(0.00292)	(0.000490)	(0.000193)	(0.000233)
Rider Count	0.496***	-0.00199	0.00269***	0.0191***
	(0.00945)	(0.00159)	(0.000624)	(0.000754)
Constant	-411.0***	-122.9***	-51.38***	49.28***
	(72.77)	(12.20)	(4.802)	(5.804)
Observations	227,043	227,043	227,043	227,043
R-squared	0.259	0.504	0.572	0.523
Race Dummy	YES	YES	YES	YES
Rider Dummy	YES	YES	YES	YES
Stage Type Dummy	YES	YES	YES	YES

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

more compact form. Figure 2 reports only the coefficient for the respective proximity dummy variable, doing so across all four models for threshold values ranging from 10 to 5000 kilometers. Figure 2 provides clear evidence of home field advantage within the Grand Tours. Intuitively, the magnitude of the effect decays as the proximity cutoff point strays farther away from the stage destination. The effect becomes statistically insignificant / relatively negligible at various points between 500 and 2500 kilometers, depending on the model considered.

Overall, there appears to be substantial evidence that professional men’s road cyclists perform better when riding closer to their birthplace, at least in long stage races like the Grand Tours. Expanding the data set to shorter stage races or one-day classics, particularly those held outside of Europe, could help determine if this advantage is more generalizable. Data on national road race championships, where all riders are citizens of the same country, could specifically aid in the teasing out of purely proximity-driven home field advantage.

Additional data on the participating racers, like their age and team at the time of racing or their role on the team during race edition could be beneficial in improving the explanatory power of the models at the least. Race environmental variables, like the number of categorized climbs, total altitude gain, or crosswind speeds, could also be feed into the models presented in this article. Future research in this area must be careful to not obscure the total impact of home field advantage by including its constituent factors in the models themselves. Metrics like weather patterns and road quality might blur the line between explanatory as an independent variable and a fundamental component of “home road” advantage within cycling. Natural experiments in the sport, like adaptations in fan attendance and scheduling put in place during the 2020 season due to COVID-19,

Figure 2
Summary of Racing Performance and Proximity - Grand Tours, 1998-2019

	(1)	(2)	(3)	(4)
Variables	Rank	Gap (min)	Gap (%)	Speed (km/h)
10 km Dummy	-10.82***	-1.225***	-0.811***	-0.0119
	(2.474)	(0.415)	(0.163)	(0.197)
25 km Dummy	-15.02***	-2.092***	-1.039***	0.463***
	(1.660)	(0.278)	(0.110)	(0.132)
50 km Dummy	-11.17***	-2.028***	-0.892***	0.468***
	(1.035)	(0.174)	(0.0683)	(0.0826)
100 km Dummy	-6.174***	-1.363***	-0.506***	0.302***
	(0.622)	(0.104)	(0.0410)	(0.0496)
250 km Dummy	-4.195***	-0.599***	-0.246***	0.163***
	(0.336)	(0.0564)	(0.0222)	(0.0268)
500 km Dummy	-3.364***	-0.483***	-0.204***	0.219***
	(0.251)	(0.0421)	(0.0165)	(0.0200)
1000 km Dummy	-4.045***	-0.209***	-0.0877***	0.151***
	(0.255)	(0.0428)	(0.0168)	(0.0204)
2500 km Dummy	-0.746	-0.822***	-0.170***	1.003***
	(0.598)	(0.100)	(0.0394)	(0.0476)
5000 km Dummy	-0.356	-1.662***	-0.457***	1.285***
	(0.694)	(0.116)	(0.0458)	(0.0553)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

have potential to help distinguish the various components of home field advantage and their relative weights.

References

[1] Luca Filipas, Antonio La Torre, Paolo Menaspa, and Hedda Giorgi. Achieving grand tour success: a pilot study using cycling's world tour points. *The Journal of sports medicine and physical fitness*, 58(10): 1432–1438, 2017.

[2] Jeremy P Jamieson. The home field advantage in athletics: A meta-analysis. *Journal of Applied Social Psychology*, 40(7):1819–1848, 2010.

[3] Marshall B Jones. The home advantage in individual sports: An augmented review. *Psychology of Sport and Exercise*, 14(3):397–404, 2013.

[4] Alan M Nevill, Roger L Holder, Andrew Bardsley, Helen Calvert, and Stephen Jones. Identifying home advantage in international tennis and golf tournaments. *Journal of Sports Sciences*, 15(4):437–443, 1997.

[5] Richard Pollard and Gregory Pollard. Long-term trends in home advantage in professional team sports in north america and england (1876–2003). *Journal of sports sciences*, 23(4):337–350, 2005.

[6] Richard Pollard, Jaime Prieto, and Miguel-Ángel Gómez. Global differences in home advantage by country, sport and sex. *International Journal of Performance Analysis in Sport*, 17(4):586–599, 2017.

[7] Rishabh Saraf and Dries Goossens. Is there a home advantage in professional road cycling? In *First Conference of Mathsport Asia: Book of Abstracts*, pages 33–33, 2018.

[8] Daam VanR̃eeth. Globalization in professional road cycling. In *The Economics of Professional Road Cycling*, pages 165–205. Springer, 2016.

Broadcasting La Liga

Gustavo Bergantiños* and Juan D. Moreno-Ternero**

*ECOSOT, Universidade de Vigo.
**Department of Economics, Universidad Pablo de Olavide

1 Introduction

The Spanish government heavily regulated the business of broadcasting football on TV in 2015. The main aspect was to impose a collective sale of broadcasting rights. Thus, an ensuing key problem arose in which the revenues collected from the sale had to be shared among the clubs. To solve this problem, strict guidelines were also enforced by the Spanish government. As explicitly stated in the corresponding Royal Decree, the aim was to "limit differences among participating entities" by means of an "equitative distribution" according to sport outcomes, ticket sales and the capability to generate resources from selling broadcasting rights. To do so, the following hybrid proposal to divide the total endowment was implemented:

1. To divide half of it equally among all clubs.
2. To divide one quarter of it proportionally to sport performance in the last five seasons.[1]
3. To divide one twelfth of it proportionally to ticket sales in the last five seasons.
4. To divide one sixth of it according to social relevance.

2 The analysis

Using models of economics and management, as well as schemes inspired in other competitions, we scrutinize the hybrid proposal described above, considering alternative ways of dividing the amount assigned to each of the four dimensions:

For the first dimension, we note that dividing equally half of the endowment is a specific way of guaranteeing *lower bounds*, as formalized in the literature on claims problems, a canonical model of distributive justice with a long tradition in economic research (e.g., O'Neill, 1982; Thomson, 2019). We consider an alternative lower bound imported from that literature.

As for the second dimension, we consider two alternatives. In the first one (and in the same vein as the Premier League), we consider that, in each season, the score of the champion is 20, the score of the second is 19, and so on until the last one gets 1. The overall score is obtained after a weighted aggregation of the last

[1] More precisely, the champion gets 17% of the season's budget. The second one gets 15%. It continues down the line with 13%, 11%, 9%, 7%, 5%, 3.5%, 3%, 2.75%, 2.5%, 2.25%, 2%, 1.75%, 1.5%, 1.25%, 1%, 0.75%, 0.5% and 0.25%, respectively. Zero is given to those that played in the second division, or below, in one of those seasons. The allocation is proportional to the resulting weighted average of those 5-season standings, with a weight of 35% for the last season, 20% for the previous to last season and 15% for each of the other three.

5 seasons, with weights being as in La Liga. In the second alternative, each club would get, each season, a score equal to the points obtained. Again, the overall score is computed as in the case of La Liga (a weighted average).

As for the third dimension (economic performance), we interpret the amount of ticket sales as the claim of each club and consider the classical rules for claims problems to allocate the corresponding budget. One of them (the proportional rule) precisely gives rise to the allocation implemented by La Liga. We consider three other alternatives, given by the remaining classical rules (the so-called constrained equal awards, constrained equal losses, and Talmud rules).

Finally, for the fourth dimension (with we take as broadcasting performance), we consider the formal model we introduced at Bergantiños and Moreno-Ternero (2020a) to divide an endowment based on the data on audiences.[2] Two polar and somewhat focal rules are salient in such a model: the equal-split rule (which allocates the revenues from each game equally among the two clubs playing, and aggregates across games) and concede-and-divide (which concedes each club the audience from its fan base and divides equally the residual). We use both rules, as well as compromises among them to suggest possible allocations for this dimension.

The combination of each of the suggestions mentioned above gives rise to a large number of possible hybrid allocations of the whole endowment. We focus on three basic allocations:

- The one induced by enforcing in each dimension the smallest variance among clubs.
- The one induced by enforcing in each dimension the highest variance among clubs.
- The average (for each dimension) between both of the above.

Clubs with weaker (stronger) records in each dimension will benefit from the first (second) allocation. Now, we see that two clubs obtain more with the allocation implemented by La Liga than with all the allocations listed above. This casts doubts about the allocation implemented by La Liga.

No club prefers the third allocation listed above (each club obtains more with either the first or the second allocation). If we compare this (third) allocation with the one implemented by La Liga with the third one listed above, two clubs are clearly favored in the latter (obtaining almost 15 millions more). One club has the opposite treatment (obtaining around 31 millions less). Two other also receive considerably less (around 9 millions).

The previous analysis is assuming that the amount devoted to each dimension is the one implemented by La Liga. But other plausible ways can be endorsed too. For instance, considering only the broadcasting performance dimension (and, thus, dismissing the others). A rationale is that if revenues are raised from selling broadcasting rights, then broadcasting performance is the dimension that really matters. Alternatively, we could assign half of the endowment to the broadcasting performance, and the other half divided equally among the other three dimensions. The two powerhouses (namely, Real Madrid and Barcelona) would obtain more with the latter alternative than with the allocation implemented by La Liga.

3 Further insights

As mentioned above, the stylized formal model in Bergantiños and Moreno-Ternero (2020a) studies how to divide an endowment among clubs, based on the data on audiences. As the second and third dimensions listed

[2] See also Bergantiños and Moreno-Ternero (2020b, 2020c 2021).

above require enriching the informational basis of the model (bringing a new prior into the problem), one might consider rules in which only the first and fourth dimensions are combined. For instance, consider the rule in which one half of the overall amount is equally shared whereas the other half is shared according to the equal-split rule. It turns out this rule is precisely the intermediate member of the UE-family of rules we study elsewhere (e.g., Bergantiños and Moreno-Ternero, 2020c). If instead of considering equal weights for the first and the fourth dimension, we consider all possible convex combinations, we obtain the whole family. Likewise, we could consider concede-and-divide instead of the equal-split rule in the previous definition. Then, the rule is precisely the intermediate member of the UC-family of rules we also study elsewhere (e.g., Bergantiños and Moreno-Ternero, 2020c). If instead of considering equal weights for the first and fourth dimension, we consider all possible convex combinations, we obtain the whole family.

Additional sources of revenue, such as qualification to other tournaments and merchandising, transferring players, or ticket sales, are certainly relevant too. Nevertheless, the sale of broadcasting and media rights is now the biggest source of revenue for most sports clubs. Furthermore, we believe that these additional sources have a different nature to broadcasting revenues, which are collectively obtained. Merchandising is mostly individual. The same could be argued for performance bonuses (such as qualifying for other tournaments), transfers or ticket sales.[3]

Regarding ticketing, we believe it would be interesting to address the problem of setting the optimal pricing of season versus game tickets for each club. This is a similar problem to the so-called museum pass problem (e.g., Ginsburgh and Zang, 2003; Bergantiños and Moreno-Ternero, 2015), a specific problem of sharing the revenue from bundled pricing. A proper analysis of this problem would require to deal with the complex relationships that might exist between both prices.[4] This sort of considerations are beyond the scope of this paper.

4 Discussion

Our work allows us to obtain some interesting lessons for the schemes suggested by the Spanish Football League Association. On the positive side, they guarantee all participating clubs lower bounds, which have a long tradition in normative work (e.g., the conflicting claims literature, or the fair allocation literature). They also compromise between the "needs-blind" view carried by performance pay and the "incentives-blind" view carried by an equal sharing of the whole pie, which seems to be another reasonable desideratum. On the negative side, a key aspect of hybrid schemes is to decide how to share a portion of the pie based on audiences and that does not seem to be sufficiently justified in the allocation implemented by the Spanish Football League.

We conclude acknowledging that we have not treated another interesting (and somewhat related) issue: the (optimal) number of clubs participating in a league. As of today, only one of the five major football

[3] Although some competitions impose partial sharing on revenues from ticket sales, typically, these are entirely handled by the club owning the stadium.

[4] Betis, an important club from La Liga had a related controversial issue after the COVID-19 cancellation of the last games of the 2019/2020 season. Instead of returning the proportional amount of the season tickets for the cancelled games, it decided to return (with several alternatives to the direct cash rebate) a lower amount. The rationale was that none of the cancelled games were against the most attractive clubs in La Liga (namely, the two powerhouses Real Madrid, Barcelona, as well as Sevilla, the historic rival from the same town).

leagues in Europe does not have 20 clubs.[5] Nevertheless, the co-called Project Big Picture, recently unveiled, is suggesting the Premier League cut from 20 to 18 clubs (with the Championship, League One and League Two each retaining 24 clubs) and this trend might eventually be followed by La Liga, Serie A and Ligue 1. We believe there are several potential arguments playing a role in this decision. One could indeed be to maximize the joint revenues from broadcasting. More clubs imply more games to be broadcasted and, in principle, more revenues to be collected. On the other hand, one might argue that too many games might exhaust viewers and, thus, audiences might be hurt (which would eventually be translated into lower revenues from broadcasting). Another is a feasibility condition given by the calendar (a year simply cannot accommodate too many games, especially in sports like football in which it is compulsory to have at least 48 hours between two games played by a same club, and international competitions coexist with domestic ones). Entry costs can also impose a relevant feasibility condition. For instance, participating clubs might be required to own a stadium with a sufficiently large capacity. Political considerations might even play a role. Finally, strong clubs (which have stronger additional sources of revenue, mostly related to international competitions) normally favor smaller numbers, whereas weak clubs favor higher numbers, which requires a bargaining protocol.

References

[1] Bergantiños, G., Moreno-Ternero, J.D., 2015. The axiomatic approach to the problem of sharing the revenue from museum passes. Games and Economic Behavior 89, 78-92.

[2] Bergantiños, G., Moreno-Ternero, J.D., 2020a. Sharing the revenues from broadcasting sport events. Management Science 66 (6), 2417-2431.

[3] Bergantiños, G., Moreno-Ternero, J.D., 2020b. Allocating extra revenues from broadcasting sports leagues. Journal of Mathematical Economics 90, 65-73.

[4] Bergantiños, G., Moreno-Ternero, J.D., 2020c. On the axiomatic approach to sharing the revenues from broadcasting sports leagues. Mimeo Repec.

[5] Bergantiños, G., Moreno-Ternero, J.D., 2021. Compromising to share the revenues from broadcasting sports leagues. Journal of Economic Behavior and Organization 183, 57-74.

[6] Ginsburgh, V., Zang, I., 2003. The museum pass game and its value. Games and Economic Behavior 43 (2), 322-325.

[7] O'Neill, B., 1982. A problem of rights arbitration from the Talmud. Mathematical Social Sciences 2, 345-371.

[8] Thomson W., 2019. How to divide when there isn't enough: from Aristotle, the Talmud, and Maimonides to the axiomatics of resource allocation, Econometric Society Monograph. Cambridge University Press.

[5] Incidentally, La Liga had 22 clubs for a short period following a bizarre situation (with strong political ramifications) that occurred in 1995. During the summer of that year, the National Football League Association decided to relegate Sevilla FC and Celta de Vigo to the third division due to a lack of documents proving the economic viability of their budgets. Two clubs from the second division (Albacete and Real Valladolid) were promoted to get their seats in La Liga (and the same was done from the third to the second division). In the aftermath of that decision, massive demonstrations occurred in Seville and Vigo, which even prompted the Spanish government to request the National Football League admitting both clubs back into La Liga. To avoid counterpart demonstrations in Albacete and Valladolid, a *Solomonic* (and somewhat chaotic) decision was taken: it was sanctioned that La Liga would have 22 clubs during the upcoming two seasons. For the season 1997/1998, La Liga returned back to the 20-club format (whereas the second division endorsed then a 22-club format that lasts until today).

Home advantage of European major football leagues under COVID-19 pandemic

Eiji Konaka*

*Meijo University

Abstract

The main objective of this study is a quantitative evaluation of “crowd effects” on home advantage, using the results of these closed matches. The proposed analysis uses pairwise comparison method to reduce the effects caused by the unbalanced schedule. The following conclusions were drawn from the statistical hypothesis tests conducted in this study: In four major European leagues, the home advantage is reduced in closed matches compared to than in the normal situation, i.e., with spectators. The reduction amounts among leagues were different. For example, in Germany, the home advantage was negative during the closed-match period. On the other hand, in England, statistically significant differences in home advantage were not observed between closed matches and normal situation.

1 Introduction

Sice March 2020, the environment surrounding football has changed dramatically — because of the COVID-19 pandemic. Similar to many other crowd-pleasing events, most football leagues were suspended. Some of the leagues had even been calcelled, whereas others resumed after a few months’ break. With regard to the resumed leagues, however, no decision as before the suspension, has been announced. Re-scheduled matches have therefore been held behind closed doors without spectators.

The objective of this study was to analyze how this unfortunate closed-match situation affected the match outcomes of football. In particular, our main objective was a quantitative evaluation of “crowd effects” on home advantage.

Today, the existence of “home advantage” in sports, especially football, seems unquestionable [1]– [4]. However, definitive evidence on the factors that produce home advantage remains elusive. In a very brief and thorough review paper on home advantage [5], Pollard presented eight main factors that could generate home advantage and explained conventional studies for each.The first one is “’crowd effect,” i.e, effect caused by spectators.

This study utilizes the results of matches conducted behind closed doors during the COVID-19 pandemic to determine the relationship between the presence of spectators and home advantage. A similar study had already been performed by Reade et al., who used the results of matches conducted behind closed doors since 2003 and reported the results of their investigation on crowd effect [6]. The number of matches in empty studiums, however, were small (160, out of approximately 34 thousand matches). In addition, in these closed matches, teams were banned from admitting supporters into their stadiums typically as punishments for bad behavior off the football pitch (e.g., bacause of corruption, racist abuse, or violence). Therefore, in this analysis, the characteristics of the spectators could biased, e.g., they were excessively violent or exhibiting overly aggressive supporting behavior.

When the results of closed matches are used in such a study, it should also be noted that the schedule is unbalanced. In 2019-2020 season, many European league matches were already approximately two-thirds completeed by the mid-March suspension. Therefore, a possible bias for the strength of home teams in the resumed matches after the break should be considered. By contrast, most famous and extensive previous studies on home advantage [7] and the most recent report [8] used only a number of basic statistics, e.g., numbe of goals, fouls, and wins. These data could be biased if obtained under an unbalanced schedule.

For the problem on biased schedule, in this study, a statistical model that deremines match results based on the team strength parameters for each team and a home advantage parameter that is common for every team in the league was assumed. Through the fitting the parameters in this model to minimize explanation error, the home advantage was separated from team strength even when the schedule was unbalanced.

This paper is organized as follows: Section 2 describes the data and the detailed algorithm used in this study. An analysis on the five major top divisions of European football leagues, i.e., England, France, Germany, Italy, and Spain (in France, the top division has not resumed after suspension), is then presented. The match results were collected from 2010–2011 season. Section 3 then discusses statistical analysis. The following conclusions were able to be drawn from the statistical hypothesis tests that were performed in this study: In the four major European leagues that were examined, the home advantage was reduced when there were no spectators compared to that for a normal situation,i.e., with spectators. The reduction amounts among the leagues were different. For all four leagues, the home advantage remained even in closed matches.

Lastly, Section 4 summarizes and concludes this paper.

2 Methods

In this section, the leagues that were investigated and the content of the used data are described. A mathematical method for estimating home advantage is then explained.

Table 1 outlines the leagues and the numbers of matches examined in this study.

Table 1: Numbers of matches examined in this study

Country	League	Teams	Matches	Matches (2019/20)	
			(2010/11 — 2018/19)	Normal	Closed
England	Premier League	20	3420	290	90
France	Ligue 1	20	3420	279	0
Germany	Bundesliga	18	2754	216	90
Italy	Serie A	20	3420	240	140
Spain	LaLiga	20	3420	270	110

This study analyzed the home advantage among the top divisions in five European countries listed in Table 1, which are considered as the most major and highest-quality football leagues around the world. The match results from the 2010/11 season were collected from worldfootball.net (`https://www.worldfootball.net/`). The number of matches analyzed was 18159, including 430 closed matches.

All five leagues were suspended from mid-March because of the COVID-19 pandemic. Four of the leagues, i.e., excluding France, resumed by late June, and finished by early August. Ligue 1 in France, on the other hand, quickly decided and announced its cancellation at the end of April.

2.1 Mathematical model

We propose a unified and simple statistical estimation method for scoring ratios based on the scores in each match, which are always officially recorded and are subject to a scoring system common to all the games. This method extends [9] by incorporating home advantage. in the Rio Olympic Games compared to those of official world rankings.

The scoring ratio of a home team *i* in a match against an away team *j* (*i* and *j* are team indices), denoted as $p_{i,j}$, is estimated as follows:

$$p_{i,j} = \frac{1}{1+e^{-(r_i + r_{homeAdv} - r_j)}}, \tag{25}$$

where r_i is defined as the *rating* of team *i*, and $r_{homeAdv}$ is the quantitative value of home advantage. Given (s_i, s_j), the actual scores in a match between *i* and *j*,

$$s_{i,j} = \frac{s_i + 1}{s_i + 1 + s_j + 1} = p_{i,j} + \varepsilon_{i,j}, \tag{26}$$

where $s_{i,j}$ and $\varepsilon_{i,j}$ are the modified actual scoring ratio and the estimation error, respectively. This modification is known as Colley's method [10], and was originally used to rank college (American) football teams.

This mathematical structure is widely used in areas such as the winning probability assumption of Elo ratings in chess games [11].

The update method is designed to minimize the sum of the squared error E^2 between the result and the prediction , defined by the following equation:

$$E^2 = \sum_{(i,j)\in \text{all matches}} (s_{i,j} - p_{i,j})^2. \tag{27}$$

The rating r_i in (25) determines the scoring ratio. We convert the rating on the scoring ratio to that of a winning probability, as follows:

$$w_{i,j} = 1 \ (i \text{ wins}),\ \ 0.5 \ (\text{draw}),\ \text{ or } \ 0 \ (j \text{ wins}), \tag{28}$$

which denotes a win, draw, or loss, respectively, for team *i* against team *j*. Afterward, D_k^*, where *k* is an index of sports, that satisfies

$$\hat{w}_{i,j} = \frac{1}{1+\exp\left(-D_k\left(r_i + r_{homrAdv} - r_j\right)\right)}, \tag{29}$$

$$D_k^* = \arg\min_{D_k} \sum \left(w_{i,j} - \hat{w}_{i,j}\right)^2, \tag{30}$$

is obtained. r_i is then converted as follows:

$$\bar{r}_i = D_k^* r_i, \ \ i = 1, 2, \cdots, N_T, \text{and } homeAdv, \tag{31}$$

where N_T denotes the number of teams. Therefore, $\bar{r}_{homeAdv}$ is a quantitative home advantage estimation that explains the effect on the winning probability.

2.2 Short-term estimation of home advantage

The proposed method in the previous section was used to estimate the rating of each team and the home advantage in the league for every matchweek using the results of the last five matchweeks, including itself.

By using five matchweeks, we were able to estimate the average of each team's strength and league-wide home advantage over periods ranging from approximately three weeks to one month.

The calculated home advantage was classified into the following four classes based on spectator attendance: **Past**: Using the matches from the 2010/11 to the 2018/19 seasons. **Normal**: Using the matches before suspension in the 2019/20 season. Note that closed matches as punishments are also inclused here, if they exist. **Mixed**: Using both the matches that included spectators and those without spectators. **Closed**: Using the matches without spectators.

3 Results and discussions

This chapter describes the analysis results and discussions. Figure 1 depicts the basic statistics, e.g., goals difference per match and win ratio difference, in normal and closed-match periods. According to these data, home teams goaled approximately from 0.3 to 0.5 more per match than away teams on average under "past and normal" situations. In addition, the home advantage was apparently reduced under closed-match situations. In particular, in Bundesliga, the goals difference and win ratio difference were both negative in the closed matches. In this part of the study, however, possible schedule unbalance for the closed matches was not considered.

Figure 2 shows the results of the estimation of home advantage $\bar{r}_{homeAdv}$ for five leagues. The medians were all positive for every four classes. This result indicates that the home advantage remained even for matches that were closed and without spectators. The medians in the past and normal periods appeared similar. On the other hand, the median in the closed period was smaller than those of the past and normal periods.

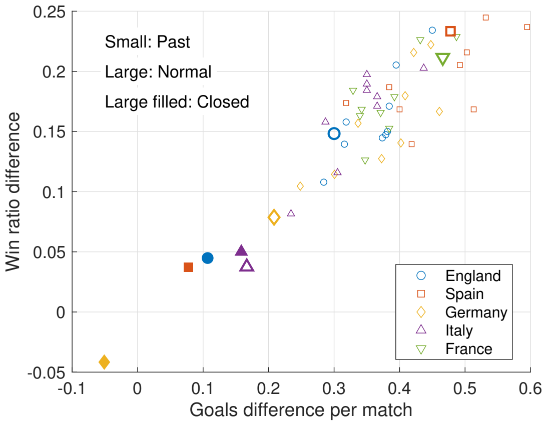


Figure 1: Win ratio difference with respect to goals difference

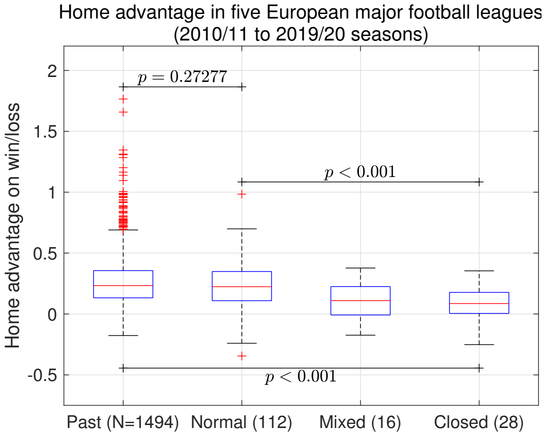


Figure 2: Distribution of $\bar{r}_{homeAdv}$ in five major European football leagues

We tested a null hypothesis that the home advantage $\bar{r}_{homeAdv}$ from two different categories were samples from continuous distributions with equal medians. Wilcoxon's rank sum test was used as a test method because any assumption on the shape of the distribution of $\bar{r}_{homeAdv}$ could not be posed. The p-values between classes are depicted in Figure 2.

From these test results, the following can be concluded: There was no significant difference in home advantage between the past's and 2019/20 season's normal matches ($p > 0.1$). There was significant difference in home advantage between the normal and closed matches $(p < 10^{-3})$. The significantnce of the difference was even more obvious between the past and closed matches ($p < 10^{-6}$). The median of the home advantage in the closed matches was clearly smaller.

Therefore, this study provides strong quantitative evidence of the impact of the crowd effect on home advantage among the top leagues in Europe. It should be noted, however, that all of the closed matches in this season have been with several months' suspension, and conducted in the summer, when matches are normally not played. Therefore, the possible effect of these factors on the home advantage should not be neglected.

In England, the home advantage value obtained via the proposed method for the closed-match period had no significant difference to those of the past and normal periods ($p = 2.13 \times 10^{-1}$ and 5.78×10^{-1}). This result demonstrates that the basic statistics were biased by the unbalanced schedule. In England, there were weak correlation between the number of home matches in the closed-match period and the final standings. The correlation value was 0.4052. By constrast, for the other leagues, i.e., Germany, Italy, and Spain, the correlation values were much smaller (0.2863, 0.0867, and -0.1561).

4 Conclusion

In this study, the results of matches conducted behind closed doors during the COVID-19 pandemic were used to determine the relationship between the presence of spectators and home advantage. To reduce the effect of schedule unbalance in the closed-match period, this paper proposed a short-term (e.g., five matchweeks) rating method that considers the home advantage. The proposed method was applied to the match results in five major European football leagues (England, France, Germany, Italy, and Spain) from the 2010/11 to the 2019/20 seasons.

The distributions of home advantage for both the past normal and closed-match periods were calculated. Their median values were compared using statistical hypothesis tests. A null hypothesis, "the home advantage $\bar{r}_{homeAdv}$ from two different periods are samples from continuous distributions with equal medians," were rejected because of sufficiently small p-value ($p < 10^{-3}$). More simply, the home advantage became smaller when the games were conducted behind closed doors

Our future work is to extend the proposed method to match results from all over the world. This future study could then clarify the crowd effect on home advantage.

References

[1] Richard Pollard. Home advantage in soccer: A retrospective analysis. *Journal of sports sciences*, 4:237–48, 02 1986.

[2] K. Courneya and A. Carron. The home advantage in sport competitions: A literature review. *Journal of Sport & Exercise Psychology*, 14:13–27, 1992.

[3] Alan Nevill, Sue Newell, and Sally Gale. Factors associated with home advantage in english and scottish soccer matches. *Journal of sports sciences*, 14:181–6, 04 1996.

[4] Alan M Nevill and Roger L Holder. Home advantage in sport. *Sports Medicine*, 28(4):221–236, 1999.

[5] Richard Pollard. Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1:12–14, June 2008.

[6] J. James Reade, Dominik Schreyer, and Carl Singleton. Echoes: what happens when football is played behind closed doors? `https://www.carlsingletoneconomics.com/uploads/4/2/3/0/42306545/closeddoors_reade_singleton.pdf`. accessed 2020/7/26.

[7] Richard Pollard and G Pollard. Long-term trends in home advantage in professional team sports in north america and england (1876-2003). *Journal of sports sciences*, 23:337–50, 05 2005.

[8] 21st Club. Empty stadiums have shrunk football teams' home advantage. `https://www.economist.com/graphic-detail/2020/07/25/empty-stadiums-have-shrunk-football-teams-home-advantage`, July 2020. accessed 2020/7/29.

[9] E. Konaka. A unified statistical rating method for team ball games and its application to predictions in the Olympic Games. *IEICE TRANSACTIONS on Information and Systems*, E102-D(6):1145–1153, June 2019.

[10] Wesley N Colley. Colley's bias free college football ranking method: The colley matrix explained. *Princeton University, Princeton*, 2002.

[11] Arpad E. Elo. *Ratings of Chess Players Past and Present*. Harper Collins Distribution Services, 1979.

Calibration and hyperparameter tuning in football forecasting with Machine Learning

E. Wheatcroft* and E. Sienkiewicz**

*London School of Economics and Political Science, Houghton Street, London, United Kingdom, WC2A 2AE. + email address: e.d.w

** London School of Economics and Political Science, Houghton Street, London, United Kingdom, WC2A 2AE.

Abstract

There is a great deal of interest in the use of Machine Learning and Neural Networks in sports forecasting. Sporting events, however, differ in nature to traditional applications of Machine Learning in which the aim is typically to provide deterministic classifications. For example, a spam filter classifies email as ‘spam’ or ‘non-spam’. For most sporting applications, probabilistic forecasts are more relevant. However, Machine Learning is not typically aimed at providing calibrated forecast probabilities. In this paper, we consider the calibration of probabilistic forecasts of football matches formed using Machine Learning algorithms and Neural Networks. We find that, in some cases, careful selection of the hyperparameters is enough to ensure that the forecasts are well calibrated, whilst, in others, an extra calibration step is required. We compare the performance of forecasts produced both with and without an additional calibration step. We then consider the use of the forecasts alongside two betting strategies and investigate whether the hyperparameters might be selected to optimise the utility of that particular decision, which, in this case, is whether to bet or not. We find some evidence that this approach may be fruitful in improving performance.

Keywords: Machine Learning, Neural Networks, Calibration, Football forecasting, Sports Betting

1 Title

Machine Learning has gained significant traction in sports forecasting in recent years, with many researchers investigating whether the flexible nature of Machine Learning algorithms is able to make better use of available information than traditional statistical forecasting methodologies [1, 3, 7]. Sports forecasting, however, is a somewhat different setting to more typical classification problems for which Machine Learning is often used. Classification aims to assign observations to groups. For example, a spam filter classifies an email as ‘spam’ or ‘not spam’ based on various ‘features’, the result being that emails are sent either to the inbox or spam folder [9]. On the other hand, there is often little value in ‘classifying’ a football match as a home win, draw or away win. Instead, *probabilistic* forecasts are usually more informative in terms of the decisions they are used to inform. In gambling, for example, bets are often placed when the forecast probability of an outcome exceeds the probability implied by the odds [5].

Machine learning algorithms are often not naturally aimed at providing calibrated forecast probabilities [8]. Algorithms such as Random Forests and Gradient boosting, for example, are ensemble learning methods which perform multiple applications of the same process to provide an aggregated estimate. In the case of a Random Forest, multiple decision trees are created on the same training set, each of which provides a single classification or 'vote' for a given label. The votes are then counted and the point is classified with the label that has the most votes. It is tempting to think of the proportion of votes for each possible label as a probabilistic forecast. However, this would be a misinterpretation of the algorithm's approach. In fact, the proportion represents a strength of evidence for a classification rather than a calibrated probability distribution [4]. To illustrate this, it is useful to take a football analogy. Since each tree has one 'vote', counting the proportion of votes and treating it like a probabilistic forecast is rather like asking one hundred football pundits who they think will win a match and treating the proportion of votes in the same way. If a strong team is playing a weak team, most, if not all, pundits will classify the match as a win for the former, yielding a proportion of one or very close to one on that team. However, we have no reason to believe that this proportion represents a probability in the frequentist sense. Simply put, the number of 'votes' do not easily translate into a probabilistic forecast distribution. Whilst not all Machine Learning techniques are ensemble methods, others also often fail to output calibrated probabilities. Gaussian Naive Bayes, for example, tends to output overconfident probabilities when the input features are correlated [2]. To attempt to produce calibrated probabilities, an extra 'calibration' step is often taken in which the proportions of votes assigned by the algorithm to each possible outcome is used as 'information' in producing the forecast [8].

In this paper, we are interested in the construction of probabilistic forecast distributions of football matches formed using Machine Learning and Neural Networks. We focus on two aspects: firstly, we consider the issue of calibration using Platt Scaling and demonstrate that it can sometimes be counterproductive and therefore should not be used as a matter of routine in football prediction. Instead, it is sometimes sufficient simply to optimise the hyperparameters (values that determine how the algorithm learns) with respect to a suitable probabilistic measure of performance such as log loss. We recommend that the performance of the forecasts formed both with and without calibration should be carefully checked when using Machine Learning methods for probabilistic forecasting. Staying with the issue of hyperparameter tuning, we then investigate whether improved decision making might be achieved by optimising hyperparameters with respect to the utility of that decision rather than an objective measure of performance. Taking gambling as an example, we find some evidence that optimising the hyperparameters with respect to mean profit can increase returns.

2 Calibration

Since Machine Learning algorithms do not necessarily provide calibrated forecast probabilities, techniques to attempt to improve the calibration are often used. The two most commonly used techniques in Machine Learning are Platt Scaling [8] and Isotonic Regression [4]. For simplicity, in this paper we consider only the use of Platt Scaling which takes the output of an ML algorithm as an input to a logistic regression model to attempt to produce calibrated probabilities. For conciseness, the details of Platt Scaling are omitted here but, for an overview, see [6].

3 Hyperparameter tuning

In Machine Learning, hyperparameters are values that control the learning process and they can have a considerable impact on the outputs of an algorithm. As such, the choice of hyperparameters is a crucial aspect of Machine Learning. A simple and common approach to hyperparameter tuning is to select a candidate set of hyperparameters, test each combination on a training set alongside cross-validation and select that which optimises some predefined measure of performance. In a spam filter, for example, the aim is usually to maximise the 'accuracy', that is the proportion of correctly classified observations. In probabilistic forecasting, one may aim, instead, to optimise with respect to a scoring rule such as the log loss, a function of a probabilistic forecast and its outcome [10].

4 All models are wrong

The focus of this paper is on probabilistic forecasting and it is worthwhile to note that no combination of hyperparameters or calibration techniques can ever lead to the output of 'true' probabilities, even in the case of an infinitely large training set. This is simply because we can never expect to gather all of the features that impact the outcome. Every combination of hyperparameters will therefore necessarily produce forecasts that differ from the true probabilities, each one 'wrong' in a slightly different way. This has an important impact on how we view hyperparameter tuning. We cannot expect to produce the true probabilities but, instead, can aim to find forecast probabilities that are as useful as possible for the context in which we plan to use them. This might be to produce forecasts for a generic audience without a specific purpose in mind, or alternatively, to inform a particular decision. In the latter case, we argue that it may be beneficial to select the hyperparameters with respect to the utility of the decision rather than some objective measure of forecast skill. In fact, assuming there are no fundamental changes over time in the relationship between the features, outcomes and the utility of the decisions to be made, for a large enough data set, one would expect this approach to yield optimal performance in terms of maximising utility. In reality, the training set is always of finite size so we are left with the question of whether the hyperparameters selected in this way are robust enough to increase performance.

5 Experimental design

We construct forecasts for football matches using data from 22 different European football leagues (data available at `www.football-data.co.uk`). For each match, we have match odds from multiple bookmakers (for which we take the maximum odds on each outcome) and we derive a set of features from existing academic papers as follows:

1. Predicted differences in the number of (i) shots on target, (ii) shots off target, and (iii) corners, for each team, formed using the Generalised Attacking Performance (GAP) ratings defined in [11].

2. Predicted differences in each team's probability of scoring from (i) any given shot, (ii) a shot on target, taken from [12].

We also include the odds-implied probability of a home win (the multiplicative inverse of the decimal odds) as an additional feature.

We divide the data set into a training set and a test set consisting of 26,484 and 35,733 matches, respectively, with the training set containing matches between the 2001/02 and 2009/10 seasons and the test set matches between the 2010/11 and 2018/19 seasons.

We build probabilistic forecasts of football matches using the following algorithms: Random Forest (RF), AdaBoost (AB), Support Vector Machine (SVM), K Nearest Neighbours (KNN), XGBoost (XGB), Gaussian Naive Bayes (GNB), Neural Network (NN).

To determine the hyperparameters for each algorithm, we use a gridsearch approach on the training set alongside two-fold cross-validation. The combination of hyperparameters that optimises the log loss is then used on the test set which is used for evaluation. Since it may be the case that the hyperparameters that optimise the performance of the forecasts without calibration are different to those that optimise the performance once the calibration step has been included, in the latter case we perform the grid search with the Platt Scaling included, i.e. we choose the combination of hyperparameters that optimises the forecasts once calibration has taken place.

We investigate the profitability of using the forecasts produced using each algorithm alongside two betting strategies. Under the *Value betting* strategy, a unit bet is placed whenever the forecast probability exceeds the probability implied by the odds. The *Kelly strategy* is similar to the Value betting strategy, but the amount staked is dependent on the size of the difference between the forecast probability and the odds-implied probability. The average stake for the Kelly strategy is set to one unit so that the results are comparable to those of the Value betting strategy. See [11] for details of the strategies. Only bets on a home or away win are considered since draws are difficult to predict.

6 Results

Table 1 shows, for the test set, the negative log loss (the negative is taken due to the convention in Machine Learning that higher scores indicate better performance) of the Platt scaled and non-calibrated forecasts for each algorithm. Here, in only two cases (Gaussian Naive Bayes and AdaBoost) is there a clear value in including the Platt scaling step. For each of the other algorithms, we find that Platt scaling is, in fact, counterproductive. This demonstrates that care should be taken when deciding whether to include the calibration step or not. In short, one should always check whether that step is, in fact, necessary. The reason for this difference is that Platt Scaling is unable to 'reduce down' to the original forecasts, that is there are no parameters in Platt Scaling that allow the original, possibly already calibrated, forecasts to be outputted. For the rest of this section, given the above results, we use Platt Scaled forecasts for Gaussian Naive Bayes and Adaboost and non-scaled forecasts for each of the other algorithms.

Algorithm	No calibration	Platt Scaling
Random Forest	-0.587	-0.588
AdaBoost	-0.602	-0.588
Support Vector Machine	-0.591	-0.599
K Nearest Neighbours	-0.591	-0.591
XGBoost	-0.588	-0.588
Gaussian Naive Bayes*	-0.642	-0.593
Neural Network	-0.587	-0.588

Table 1: Negative log loss on the test set for each algorithm with and without Platt Scaling. Note that higher scores indicate better forecast skill. Gaussian Naive Bayes is starred because it does not have any hyperparameters.

We now investigate the performance of the forecasts in terms of gambling profit. In general, one might expect that more skillful forecasts should be able to return a higher profit. In figure 1, the negative log loss is plotted against the mean percentage profit under the value betting (blue dots) and Kelly betting (red dots) strategies for each algorithm.

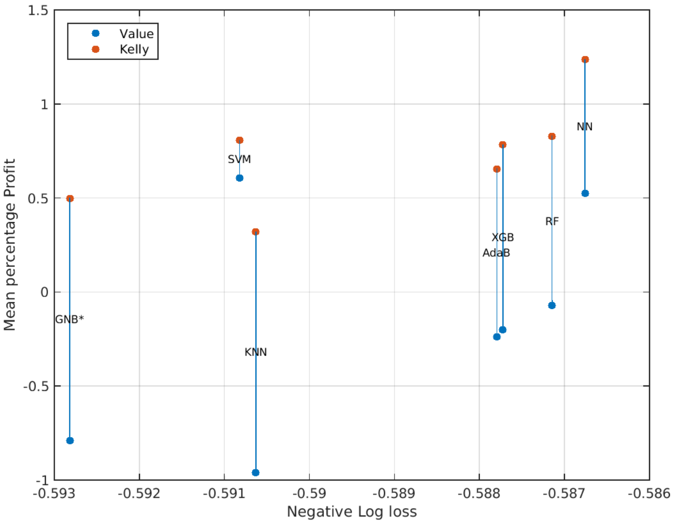


Figure 1: Log loss of each algorithm plotted against the mean percentage profit under Value betting (blue) and the Kelly Strategy (red). The star by the Gaussian Naive Bayes entry denotes that it has no hyperparameters.

Here, there appears to be some relationship between the skill of the forecasts and the mean profit. Forecasts produced using the Neural Network, for example, yield both the best negative log loss and the highest profit under the Kelly strategy. However, the relationship between the negative log loss and the mean profit is not a particularly strong one. This suggests that it might be possible to select hyperparameters that reduce the negative log loss but increase the mean profit. We therefore test the strategy of using Machine Learning algorithms with hyperparameters optimised with respect to mean profit under each of the two betting strategies. In figure 2, the mean profit from forecasts optimised with respect to the log loss is plotted against that obtained from forecasts optimised with respect to the mean profit. Here, points above the diagonal line are cases in which optimising with respect to profit yields better results than optimising with respect to log loss. Whilst the Support Vector Machine under Value betting is a notable exception, there seems to be some indication that this approach might be effective in increasing profits, though more work is needed to determine the robustness of this approach.

7 Conclusion

The results in this paper demonstrate that care should be taken before including a calibration step in probabilistic forecasting using Machine Learning algorithms and Neural Networks. It is important to test the performance of the forecasts formed both with and without calibration to ensure that the calibration step is indeed needed. This paper has also briefly considered the idea of hyperparameter tuning with respect to the

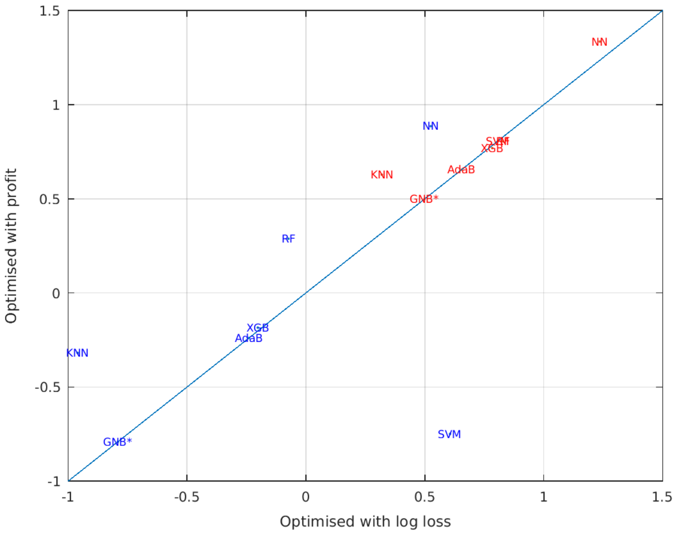


Figure 2: Mean percentage profit per bet under the Value betting (blue) and Kelly strategy (red) when the hyperparameters are optimised with respect to the log loss against when they are optimised with respect to the mean profit itself.

utility of the decision to be made. Whilst more work needs to be done to determine the robustness of the approach, these results are promising in that, for many of the algorithms considered, improved gambling performance would have been achieved.

References

[1] Baboota, R. and Kaur, H. (2019) , 'Predictive analysis and modelling football results using machine learning approach for English Premier League', *International Journal of Forecasting* **35**(2), 741–755.

[2] Bennett, P. N. (2000) , Assessing the calibration of naive bayes posterior estimates, Technical report, Carnegie-Mellon Univ. Pittsburgh PA School of Computer Science.

[3] Berrar, D., Lopes, P., Davis, J. and Dubitzky, W. (2019) , 'Guest editorial: special issue on machine learning for soccer', *Machine Learning* **108**(1), 1–7.

[4] Boström, H. (2008) , Calibrating random forests, *in* '2008 Seventh International Conference on Machine Learning and Applications', IEEE, pp. 121–126.

[5] Dixon, M. J. and Coles, S. G. (1997) , 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.

[6] Géron, A. (2019) , *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media.

[7] Hucaljuk, J. and Rakipović, A. (2011) , Predicting football scores using machine learning techniques, *in* '2011 Proceedings of the 34th International Convention MIPRO', IEEE, pp. 1623–1627.

[8] Platt, J. (1999) , 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Advances in large margin classifiers* **10**(3), 61–74.

[9] Tretyakov, K. (2004) , Machine learning techniques in spam filtering, *in* 'Data Mining Problem-oriented Seminar, MTAT', Vol. 3, Citeseer, pp. 60–79.

[10] Wheatcroft, E. (2019) , 'Interpreting the skill score form of forecast performance metrics', *International Journal of Forecasting* **35**(2), 573–579.

[11] Wheatcroft, E. (2021) , 'Forecasting football matches by predicting match statistics', *Journal of Sports Analytics (Ahead of Print)* .

[12] Wheatcroft, E. and Sienkiewicz, E. (2021) , 'A Probabilistic Model for Predicting Shot Success in Football', *arXiv preprint arXiv:2101.02104* .

A framework for monitoring all of club performance for amateur rugby union

P.J. Bracewell*, S.M. Bracewell**, J. K. Wilson* and D.K. Manchaiah*

* DOT loves data, Wellington, New Zealand: {paul | jordan | deepak}@dotlovesdata.com
** Wellington Football Club, Wellington, New Zealand: sarah@wfc.org.nz

Abstract

Falling amateur playing numbers in rugby union places increased pressure on the viability of clubs within New Zealand. Tackling this issue to ensure club sustainability requires consideration beyond just on-field performance.

Here, a framework for measuring all of club performance is outlined. In this instance, all of club encompasses junior to senior playing grades, as well as on-field and off-field activity. As the drivers at each playing level differ, especially within the amateur environment, the metrics embraced must reflect the relevant ages and stages.

The proposed framework utilises a balanced scorecard, a common strategic management performance metric used within business to improve internal process and resulting external outcomes. Balanced scorecards are used to measure and provide feedback. Data collection is critical to providing quantitative results to help make informed decisions in a timely manner.

Based on the results of a member survey, this framework is adapted to include player and participation, alongside rugby (coaching and performance) as the leading indicators, with member engagement and finance as the lagging indicators. Importantly, this aligns with the governance and operational structure of the club used in the case study which is presented here. The metrics and rationale for their creation of those metrics is outlined.

1 Introduction and Background

Founded in 1870, The Wellington Football Club (WFC) is Wellington's oldest rugby club, and one of the oldest clubs in New Zealand (`https://www.wellingtonfootballclub.org.nz`). Colloquially known as the Axemen, the club was established under the direction of Captain J. C. R. Isherwood, a Crimean war veteran and the adjutant of the 69 Foot Regiment, who bestowed the distinctive yellow and black colours upon the club. These colours have been adopted by the Wellington province and are widely worn by representative teams from myriad codes, such as the Wellington Firebirds and Blaze (Cricket), Wellington Lions and Pride (Rugby), Wellington Phoenix (Association Football) and Wellington Pulse (Netball).

WFC has fielded a team every winter Saturday since its formation. During this extensive period, the on-field fortunes of the club have ebbed and flowed. Arguably, the most successful period for the club was during the 1980s, when the highest honour in Wellington Club Rugby, the Jubilee Cup, was won three times. Overall, the Championship or Jubilee Cup has been won or shared 13 times (five Championships, contested from 1880 to 1928: 1883, 1885, 1890, 1901, 1914 and eight Jubilee Cups, contested from 1929 to date: 1939, 1957, 1972, 1978, 1982, 1983, 1985, 1987). In the 2020 season, disrupted due to Covid-19, the Axemen finished 13 of 14 teams competing in the Premier 1 competition.

The club has not been immune to a decline in rugby playing numbers (Cleaver, 2020; Tso, 2020; NZ Herald, 2019), fielding four senior teams in 2021 (Premier 1, Premier 2, Under 85kg and Presidents (aged over 30)). Junior players are registered in all age-based grades from Under 6 to Under 13. This season eight teams have taken the field, which is up from seven teams in the 2019 and 2020 seasons. This growth is positive, given that junior rugby teams across the Wellington region have fallen 15% in the last two seasons (dropping to 264 in 2021 from 311 in 2019). On-field performance, a focus on fielding a Colts team (players aged under 21) and the desire to increase the community engagement for the WFC highlight the key internal drivers for adopting a systematic approach for implementing and measuring change.

The external perspective continues to add pressure to ensure sound governance of "grass roots" club rugby. Rugby is clouded with uncertainty off the back of a tumultuous 2020 season where, according to the Wellington Rugby Football Union (WRFU), the impact of Covid-19 saw a 31% drop in revenue, a 49% decrease in Trust Funding and 41% drop in Sponsorship. The proposed investment by US private equity firm, Silver Lake (Knowler, 2021), further highlighted the financial predicament faced by New Zealand Rugby, with a $34 million dollar loss posted by the national governing body for 2020. In addition, WRFU have voiced the desire to have a strong premier competition and have this based around vibrant clubs and communities. This further reinforced the importance of change for the Axemen.

2 Response to Internal and External Pressures

Late in the 2020 rugby season WFC undertook a review of how the club is run. This work was commissioned by: Club Patron and Former All Black, Murray Mexted; then President and Former Sports Broadcaster, Keith Quinn; and then Club Chair, Scott Fuller. A transition committee of eight members was formed and they were given three areas of focus to investigate: 1) Brand and Culture, 2) Rugby 3) Finance.

The process ran for approximately three months. One of the compelling insights from the investigation was the response to a survey from approximately 27 members. Combining Likert responses and topic modelling of free text responses lead to an interpretation of the results as the desire from members for: "A competitive club that is welcoming and inclusive, built around family and friends." The methodology used for the natural language processing is similar to that outlined by Lisena et. al. (2020).

In addition, from a workshop conducted by Sam Gadd, founder of employee experience agency, Humankind (`https://www.humankind.nz`), the players indicated they wanted a voice as to how the club was run, input into all aspects of the club and greater visibility of decision being made about the club. This aligned with direct, verbal player feedback to members of the transition committee.

The transition committee found that as rugby becomes increasingly professional, the club must also adopt greater professionalism with respect to governance and strategic planning. The structure of the management committee gave insufficient voice to players and had limited emphasis regarding on-field performance.

Consequently, the implementation of a board was recommended to set and oversee the strategic direction of the club. This major change required a rewrite of the club's constitution which was tabled at the annual general meeting in November 2020 and ratified at a special general meeting held in December 2020 (`https://is-register.companiesoffice.govt.nz`).

3 Implementing a New Governance Structure

In creating a board to oversee the strategic direction of WFC, several organisational structures were considered. A key consideration was ensuring all members of the club had the opportunity to provide input and have visibility of decisions and outcomes. This review found that School Boards, Auckland Rugby and Wellington Cricket provided a suitable template that covered on and off field behaviour. The organisational structure was also considered from a Product Company perspective, with rugby being the product in question. A board consisting of seven people was adopted, with four of these directors being directly accountable for one of four key areas: **Participants** (People); **Rugby** (Product); **Marketing and Communication** (Customer); **Operations and Finance** (Finance).

The alignment for each of the four areas of focus and a balanced scorecard becomes immediately obvious when these elements are placed in the context of People, Product, Customer and Finance with direct reference to being either a lead or lag indicator. In this instance, the topic of People relates to our playing members and their on-field activities, whereas the customer topic covers all members and relates to off-field activities. A balanced scorecard is a common strategic management performance metric used within business to improve internal process and resulting external outcomes. Balanced scorecards are used to measure and provide feedback. Data collection is critical to providing quantitative results to help make informed decisions in a timely manner. As discussed in the next section, consideration must be given to the age and stage of each team, as well as ease data capture. The implication in this case study is that each indicator can be linked to the desired state of being a competitive club that is welcoming and inclusive, built around family and friends.

4 Quantifying Impact

The four areas are measurable. Several iterations were proposed for measuring the weekly impact of playing member behaviour. This needed to be tempered with an element of simplicity which respected the playing conditions for each team. To help encourage data capture and engagement with the process, the metrics have been used to define Team of the Year for the club and presents a unique opportunity to engage senior and junior teams. This does create challenges to ensure a level playing field that is also sympathetic of the different drivers influencing each team. The simplified outcome for each of the four indicators follows:

1. Measuring Rugby Performance (Leading Indicator): As mentioned previously, the top team in the club finished 13 out of 14 teams in 2020. Consequently, defining “competitive” from a binary perspective of win and loss is potentially problematic. Furthermore, emphasising winning for junior rugby teams can come at the expense of enjoyment which can be contrary to the NZR principles covered by Small Blacks (`https://www.smallblacks.com/how-to-play`).

To handle this scenario, the board determined that a junior team was competitive if they were within two converted tries of the opposition, or won. Essentially all junior teams are given a point start of +14.5 points. If they are competitive, the team earns one point, else 0 points. The President’s grade is social, and scores are not accurately kept. Thus, each game is assumed to be a hard fought, yet enjoyable draw, receiving 0.5 points. Premier 2 and Under 85 use a similar scoring system to the juniors, with a point start of 14.5 points being achieved yielding 0.7 points, with a win earning a full point. When the Colts re-enter the competition in 2022, they will adopt a scoring system like the Premier 1 team, outlined below.

Simply, if the Premier 1 team improves week on week, they will eventually be highly competitive.

Importantly, recent outcome data is readily available (`http://www.clubrugby.co.nz/wellington`), which creates an ideal opportunity for creating team ratings. Elo ratings provide a simple and explainable method for showing continuous, incremental improvement. In developing the Elo rating, three key features are adopted: 1) a high value of *k*, 2) adjusting for the margin of victory and 3) converting the ratings to a probability of beating an average team. The popular data website fivethirtyeight.com demonstrates the mainstream nature of Elo Ratings, highlighting the relevance of this type of statistic to show improvement (`https://fivethirtyeight.com/methodology/how-our-nfl-predictions-work`).

The update parameter, *k*, is set to 40 which is high and leads to volatile ratings. Given the nature of club rugby, substantial changes to playing XV can occur over the course of the season, due to injury, elevation to representative teams, work commitments and so forth. Table 1 compares the starting XV for the WFC Premier 1 team from round 1 of 2021 with round 8, revealing only three players started both these games.

Jersey	Round 1 vs Poneke	Round 8 vs J'ville
15	Zane Ainslie	William Cosgriff
14	Philip Elo	Beau Murphy
13	Jayden Kini	Ahtun Masun
12	Ranui Burchett	Jay Cameron
11	Aiden Fleming	Jayden Kini
10	William Cosgriff	Ben Stowe
9	Isaac Bracewell	Isaac Bracewell
8	Josh Michael	Vaea Fifita
7	Davis Eni	Manu Maihi-Ionae
6	Tom Robertson	Tom Robertson
5	Ben Strowger-Turnock	Regan Pope
4	Jack Laurenson	Josh Michael
3	Tuterangi Andersen	James Coburn
2	Liam Privett	Ben Hendry
1	Joe Reid	Tony Coburn

Table 1: Comparing Premier Starting XV for the Wellington Football Club in Round 1 (10th April) with Round 8 (29th May) of the Swindale Shield, 2021

Rather than using a binary win or loss metric to update the ratings, the margin of victory is used. Finally, to make the output easier to understand, the Elo Rating is converted to the probability of beating an average team (Rating = 1500), using the calculation outlined on the fivethirtyeight.com.

User testing of club members was effective in determining the relevance of these ratings. As the outcomes aligned with intuition, this mechanism for tracking performance has been adopted. The Axemen have been competitive four times in the first eight rounds of 2021, which includes one win over Johnsonville Hawks on the 29th of May 2021 to retain the Mick Kenny Memorial Cup. 2016 was the Axemen's strongest recent season, finishing third in the Swindale Shield (regular season), with nine wins from 13 matches. This is evident in Figure 1 which shows the chance of the WFC Premier team beating a team with a rating of 1500.

2. Measuring Participation (Leading Indicator): The simplest mechanism for measuring the engagement of playing members is attendance at training. For the senior grades, this is attendance at training. As an example of the value of tracking these numbers, in the build up to the round eight clash with Johnsonville, the Premier 1 team had more than 30 players at training. Consequently, a binary measure is adopted. If squad attendance exceeds 70% then a score of 1 is given, else the score is 0. For the President's grade, training is taken as being the post-match on Saturday.

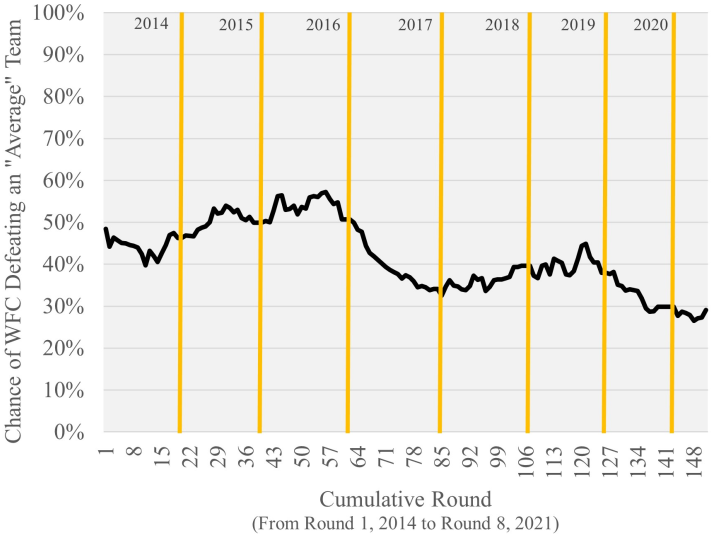


Figure 1: The chance of WFC beating an average team from Round 1 in 2014 to Round 8 in 2021.

This type of approach for measuring engagement is not entirely appropriate for junior teams due to the lack of independence for getting to training and the impact of other extracurricular activities. As such, the initial intended approach was to provide every player a quick three question survey after every game: 1) *did you have fun today?* 2) *did you get to run with the ball today?* and 3) *did you make any tackles today?* The intent of these questions was to encapsulate fun and skill acquisition. However, this was quickly found to be too onerous for parents, caregivers, managers, and coaches alike. Instead, using the match reports that every team provides, natural language processing algorithms are used to measure the sentiment and alignment with fun to create a single measure between 0 and 1. Sentiment and fun are given equal weighting in creation of this score using topic modelling (Lisena et. al., 2020) and sentiment analysis (Bracewell et. al., 2016).

3. Measuring Customer Engagement (Lagging Indicator): To encapsulate the community aspect of the rugby club, attendance at social events by some members from each team are noted. Of the three major events to date this season, up to round 8 of 2021: Season Launch, Celebrating Women in Rugby and Re:union Day (formerly known as Old Timers Day), there has been strong engagement from every team from Under 6 to Premier 1. This is simply a binary measure of more than one player or supporter (in the case of the juniors) attending major club events.

4. Measuring Financial Performance (Lagging Indicator): The financial impact for each team is the simplest metric of all four. It is the proportion of registered players in each grade who have paid their subscriptions. After eight rounds, approximately half the Premier squad have paid their subscription. Contrast that against the juniors where the tackle grade teams have 100% payment of subs (Under 8 to Under 13).

5 Overall Measure

Each of the four indicators, ranging from 0 to 1 are updated every week for each team and given an equal weight of 25% to create a composite score ranging between 0 and 1. To account for the different lengths of season across all grades, the average rating is taken. Essentially, this becomes a weighted average due to the impact of time. In essence, a team that pays their subscriptions in full earlier will be rewarded more.

The combination of all team results enables a whole of club score to also be generated and tracked over time, which is useful from a governance perspective to determine how well the club is tracking. Importantly, due to the transparent structure of the score, any substantial changes, for better or worse, can be identified.

The team with the highest score at the end of the season will be determined to be team of the year. A trophy first awarded at the club for Team of the Year in 1929 and in hiatus since 2007 is being re-purposed to become a whole of club award. This is an important step to have greater visibility of the desire for the club to be competitive, welcoming and inclusive, built around family and friends.

6 Conclusion

A framework for measuring the key drivers for an amateur rugby club was outlined. Four key drivers, that aligned with a balanced scorecard enable a simple, yet effective mechanism for tracking amateur team and club performance using leading indicators of member participation on-field and rugby performance, combined with lagging indicators relating to member engagement off-field and the financial contribution of subscriptions. Key considerations in developing this framework were outlined that paid respect to the different motivations for playing rugby at different ages and stages of members. This enables a comparable score between junior and senior teams to be allocated and tracked. The intent of the metric is to provide greater visibility to the newly formed board of directors about how the club is tracking and identifying areas of strength and weakness. Excitingly, the early results from this framework for the Wellington Football Club show that the club as a whole is tracking well towards the desired state of being competitive on-field with great community engagement off-field.

References

[1] Bracewell, P.J., McNamara, T.S, and Moore, W.E. (2016). How Rugby Moved the Mood of New Zealand. *Journal of Sport and Human Performance*. **4**(4). pp. 1-9.

[2] Cleaver, D. (2020, September 7) How to save NZ club rugby from dying - and what the numbers tell us about Super Rugby Aotearoa. NZ Herald. Retrieved from `https://www.nzherald.co.nz`.

[3] NZ Herald (2019, February 25). Rugby: Buck Shelford - New Zealand Rugby is in trouble over dwindling playing numbers. NZ Herald. Retrieved from `https://www.nzherald.co.nz`.

[4] Knowler, R. (2021, April 30). NZ Rugby and Silver Lake deal: What the fuss is all about. *Stuff*. Retrieved from: `https://www.stuff.co.nz`.

[5] Lisena, P., Harrando, I., Kandakji, O. and Troncy, R. (2020). ToModAPI: A Topic Modeling API to Train, Use and Compare Topic Models. *Proceedings of Second Workshop for NLP Open-Source Software (NLP-OSS)*, pp. 132–140.

[6] Tso, M. (2020, February 15). Rugby losing ground at high schools as students turn to other sports. *Stuff*. Retrieved from: `https://www.stuff.co.nz`.

Evaluation of soccer team defense based on ball recovery and being attacked

Kosuke Toda*, Masakiyo Teranishi**, Keisuke Kushiro*, Keisuke Fujii**

*Graduate School of Human and Environmental Studies, Kyoto University, Kyoto, Kyoto, Japan
** Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan.

Abstract

With the development of measurement technology, data on the movements of actual games in various sports are available and are expected to be used for planning and evaluating the tactics and strategy. In particular, defense in team sports is generally difficult to be evaluated because of the lack of statistical data. Conventional evaluation methods based on predictions of scores are considered unreliable and predict rare events throughout the entire game, and it is difficult to evaluate various plays leading up to a score. In this study, we propose a method to evaluate team defense from a comprehensive perspective related to team performance based on the prediction of ball recovery and being attacked, which occur more frequently than goals, using player behavior and positional data of all players and the ball. Using data from 45 soccer matches, we examined the relationship between the proposed index and team performance in actual matches and throughout a season. Results show that the proposed classifiers more accurately predicted the true events than the existing classifiers which were based on rare events (i.e., goals). Our results suggest that the proposed index might be a more reliable indicator rather than winning or losing with the inclusion of accidental factors.

1 Introduction

The development of measurement technology has allowed for the generation of data on the movements in various sports games for use in planning and evaluating the tactics and strategy. For example, tracking data during a game of soccer, including the positional data of the players and ball, is commonly used for individual players' conditioning. However, during a soccer match, all 22 players and the ball interact in complex ways for scoring goals or preventing being scored (it is sometimes referred to as conceding) for each team. Hence, it is then necessary to evaluate the performance of not only individuals but also the entire team [3]. Defensive tactics are particularly considered difficult to evaluate because of the limited amount of available statistics, such as goals scored in the case of attacks.

There are three main approaches to quantitatively evaluate teams and players in soccer, mainly from an attacking perspective. The first approach is based on scoring prediction, which evaluates plays based on changes in the expected values of goals scored and conceded based on a prediction of scoring using tracking data (e.g., [9]) and action data such as dribbling and passing [1], as well as other rule-based methods (e.g., [15]). The second approach is used to evaluate plays such as passes and effective attacks which lead to shots. For example, a previous study evaluated the value of passes based on relationships to the expected score and the difficulty in successfully completing a pass [11]. An effective attack can be defined as a play that will likely lead to a score [17]. Previous studies have analyzed pass networks [18] and three player interactions [19, 20], as well as pass reception [7] and of the related defensive weaknesses [8]. For defenses, researchers have evaluated interception [10] and the effectiveness of defensive play by the expected value of a goal-scoring opportunity conceded [12]. For the third approach, spatial positioning of the players is evaluated by calculating the dominant region with the use of a Voronoi diagram [14]. Recent research has also been conducted on the evaluation of movements that create space for teammates [2, 13].

However, these approaches have have several limitations. For evaluation based on the prediction of scoring (i.e., the first approach), the evaluation is not reliable because it predicts events that are rare throughout a game, and the process leading up to the goals is sometimes difficult to evaluate. Furthermore, the second approach to evaluate specific plays that lead to goals and the third approach regarding positioning have difficulties in relating the evaluation to overall performance (such as wins and losses). Also, since many studies on the first and second approaches have used only the actions and coordinates of players around the ball, it would be difficult to evaluate players at greater distances from the ball and the team as a whole.

To address these issues, we propose a method called *Valuating Defense by Estimating Probabilities* (VDEP), which utilizes the actions and positional data of all players and the ball. The main contributions of this work are as follows: (i) the proposed method is based on the prediction of ball possession and effective attacks, which occur more frequently than the rare goals; (ii) based on a comprehensive perspective related to team outcomes, we evaluated the team's defense. Methodologically, we modified the existing method called VAEP (Valuating Actions by Estimating Probabilities) [1], which is based on the classifiers to predict scoring and conceding, so that the defensive process can be evaluated by applying the approach to ball recovery and being attacked. We validated the classifiers of the proposed and existing methods and shows that the proposed classifiers predicted the true events more accurately than the existing classifiers. Moreover, we examined the relationship between VDEP and the team performance in actual matches and throughout the season, as compared with VAEP. Note that this paper is a part of the paper [16] in review.

Materials and methods

Dataset. In this study, we used event data (i.e., labels of actions, such as passing and shooting, recorded at 30 Hz and the simultaneous xy coordinates of the ball) and tracking data (i.e., xy coordinates of all players and the ball recorded at 25 Hz) of a total of 45 games from week 30 to week 34 of the Meiji Yasuda Seimei J1 League 2019 season provided by Research Center for Medical and Health Data Science in the Institute of Statistical Mathematics and Data Stadium Inc. Data acquisition was based on the contract between the soccer league (J League) and the company (Data Stadium, Inc), not between the players and us. The company was licensed to acquire this data and sell it to third parties, and it was guaranteed that the use of the data would not infringe on any rights of J.League players or teams. We obtained the data by participating in a competition hosted by the above organizations.

In all 45 games, there were 106 goals scored, 1,174 shots, 3,701 effective attacks, and 9,408 ball recoveries (all based on the provided event data). An effective attack is defined as an event that finally ends in a shot or penetrates the penalty area. Also, ball recovery is defined as a change in the attacking team before or after the play due to some factors other than an effective attack. In this study, an effective attack is defined as *being attacked* from the defender's perspectives. When calculating VDEP and VAEP values, we used a cross-validation procedure, which repeats the learning of classifiers using the data of four weeks (36 games) and a prediction using the data of one week (9 games) five times (i.e., data of all five weeks were finally predicted and evaluated) to analyze all games [4–6].

Proposed Method. The ultimate goal of defense in soccer is to prevent the opposing team from scoring a goal. However, since goal-scoring scenes are rare events, it may lead to ineffective training of a classifier and evaluating the events in a unreliable manner (the validation results of the VAEP method [1] will be presented later). Therefore, to reasonably evaluate the defense of a team, we propose the VDEP method to evaluate important factors for preventing goals from being scored. The VDEP method evaluates the potential increase in the number of ball recoveries and the potential decrease in the number of effective attacks. The number of effective attacks was chosen instead of the number of shots because of the following scenarios as defensive failures, in which an attacker selects to pass the ball rather than to shoot. Therefore, we evaluate the process of defense based on the expected value computed by the classifiers to predict ball recovery and being attacked in an analogous way of the VAEP method [1] based on the prediction of scoring and conceding.

Suppose that the state of the game is given by $S = [s_1, \ldots, s_N]$ in chronological order. We consider $s_i = [a_i, o_i]$, whereas [1, 12] used only a_i, which includes the ith action involving the ball and its coordinates. The proposed method utilizes classifiers trained with the state s_i, which includes the feature o_i far from the ball (off-ball) at the time of the action. Since all defensive and offensive actions in this study are evaluated from the defender's point of view, the following time index i is used as the ith *event*.

Given the game state S_i of a certain interval, we define the probability of future ball recovery $P_{recoveries}(S_i)$ and the probability of being attacked $P_{attacked}(S_i)$ in a state S_i at an event i based on the classifier trained from the data. Defensive players are considered to act so that $P_{recoveries}(S_i)$ becomes higher or $P_{attacked}(S_i)$ becomes lower. Therefore, the value of defense in the proposed method V_{vdep} is defined as follows:

$$V_{vdep}(S_i) = P_{recoveries}(S_i) - C * P_{attacked}(S_i), \tag{32}$$

where C is a parameter that adjusts the values of ball recoveries and effective attacks. In this study, we adjusted these values based on the frequency of each event in the training data. As described below, we determined $C \approx 3$ because the ratio of ball recoveries and effective attacks is approximately 3 : 1 (the value differs for each of 5-fold cross-validation). Since the main aim of this study is to evaluate the team, we define the evaluation value per game for team p as follows: $R_{vdep}(p) = \frac{1}{M}\Sigma_{S_i \in \boldsymbol{S}_M^p} V_{vdep}(S_i)$, where M is the number of events for team p in a match and $\boldsymbol{S}_M^p$ is the set of states S of team p up to the Mth event. Similarly, the sum of evaluation values using only $P_{recoveries}$ and $P_{attacked}$ are defined as $R_{recoveries}(p)$ and $R_{attacked}(p)$, respectively. For the VAEP [1] method, the value averaged by the playing time of each player was used. However, since the time each team played the game was almost the same, in this study, each team is evaluated by the sum of $S_{vaep}(p)$ as the VAEP value. Also, $S_{scores}(p)$ and $S_{concedes}(p)$ are used in the analysis as separate evaluation values, although the VAEP [1] value is calculated based on the prediction of goals scored and conceded.

Procedures. The feature a_i near the ball in this study was constructed using the action and tracking data with reference to [1]. Specifically, we used the types of events used in [1] (19 types including pass, shot, tackle, and so on), the start/end time of the event, the displacement of movement, and elapsed time from the start to the end of the event, the distance and angle between the ball and the goal, and whether there was a change in offense or defense from the previous event (73 dimensions in total). Moreover, in this study, the off-ball feature o_i at the time that the event occurred was included in the state s_i. Specifically, for each team, we used the x and y coordinates of positions of all players and the distance of each player from the ball, sorted in the order of closest to the ball (137 dimensions in total). We adopted XGBoost (eXtreme Gradient Boosting) used in [1], as the classifier to predict ball recoveries and being attacked. Gradient boosting methods are known to perform well on a variety of learning problems with heterogeneous features, noisy data, and complex dependencies. The time range of the input S_i to the classifier was ith, $i-1$th, and $i-2$th actions in [1]. In this study, since the effect of s_{i-2} on the prediction performance was small in the preliminary experiments, we used $s = [a, o]$ including the ith and $i-1$th actions.

In the first classification for estimating $P_{recoveries}(S_i)$, we assigned a positive label ($= 1$) to the game state S_i if the defending team in the state S_i recovered the ball in a subsequent k actions, and a negative label ($= 0$) if the ball was not recovered. Similarly, in the second classification for estimating $P_{attacked}(S_i)$, we assigned a positive label ($= 1$) to the game state S_i when an effective attack was made in a subsequent k actions. In both classifications, k is a parameter freely determined by the user. If k is small, the prediction is short-term and reliable, and if k is large, the prediction is long-term and includes many factors. In this study, we set $k = 5$ based on the results of preliminary validation.

In the data used in this study, defined by k above, the total number of events for all teams was 97,335, with 35,286 positive cases of ball recovery and 13,353 positive cases of being attacked. In terms of goals scored and conceded for the calculation of the VAEP value [1], there were 753 positive cases of goals scored and 227 positive cases of goals conceded (the total number of events was the same, but we set $k = 10$ in accordance with [1]). These indicate that goals scored and conceded are rare events compared to ball recoveries and being attacked. Therefore, the goals scored and conceded may not be correctly evaluated by the area under the receiver operating characteristic curve (AUC) and Brier scores used in [1].

Evaluation and Statistical Analysis. To validate the classifier, we used the F1 score in addition to the AUC and Brier scores used in [1].However, these evaluations may not be correct when there are extremely more negative than positive cases, as in this and previous studies (for example, AUC and Brier score are good even when all negative cases are predicted for the data with only 10% positive cases). In this study, we also used the F1 score to evaluate whether the true positives can be classified without considering the true negatives. In this index, only true positives are evaluated, not true negatives. To compare F1 scores among the various classifiers for testing our hypothesis (other AUC and Brier scores are shown only as references), a one-way analysis of variance was performed. As a post-hoc comparison, Tukey's test was used within the factor where a significant effect in one-way analysis of variance was found.

For the evaluation of defense using the VDEP and VAEP values [1], we present examples to quantitatively and qualitatively evaluate a game and a season of a specific team. Next, we examined the relationships with the outcomes of actual games (goals scored, conceded, and winning points, where win, draw, and lose were assigned as 3, 1, and 0 points, respectively) and the relationship with the team results throughout the season using the Pearson's correlation coefficient among all 18 teams. For all statistical analysis, $p < 0.05$ was considered significant. However, since the sample size was small ($N = 18$) in the correlation analysis, the r value indicating the magnitude of the correlation was also used as an effect size for evaluation.

Results

Validation of Classifiers. To validate the VDEP and VAEP [1] methods, we first investigated the prediction performances of their classifiers. In Table 1, the classifiers of VDEP shows more accurate predictions compared to those of VAEP [1] (note that the output and number of occurrences to be predicted are different). The AUCs of $R_{recoveries}$ and $R_{attacked}$ in VDEP were better than those of S_{scores} and $S_{concedes}$ in VAEP, and vice versa in regard to the Brier scores. However, again, these indices may not be validly evaluated because they include a large number of true negatives in the evaluation (thus, we did not perform statistical analysis in these variables). Instead, the F1 score was calculated, and the statistical analysis identified significant main effect among $R_{recoveries}$, $R_{attacked}$, and S_{scores} ($F = 144.40,\ p < 1.0 \times 10^{-6}$; $S_{concedes}$ was eliminated because of the average is near zero value). The post-hoc analysis shows that F1 scores of VDEP ($R_{recoveries}$, $R_{attacked}$) were significantly higher than that of S_{scores} ($ps < 0.002$). This indicates that the VDEP method predicted true positives correctly, while the VAEP did not.

Table 1: **Evaluation of classifiers for the proposed and conventional methods.**

	AUC	Brier score	F1 score
$R_{recoveries}$	0.770 ± 0.014	0.184 ± 0.009	0.522 ± 0.036
$R_{attacked}$	0.862 ± 0.003	0.079 ± 0.003	0.484 ± 0.038
S_{scores} [1]	0.698 ± 0.066	0.007 ± 0.002	0.201 ± 0.021
$S_{concedes}$ [1]	0.701 ± 0.040	0.003 ± 0.001	0.000 ± 0.000

Validation of Evaluation Methods. First, correlation analysis was performed between the outcome of the game and the proposed and existing indices. In the case of R_{vdep}, there were moderate positive correlations with winning points ($r_{16} = 0.464, p = 0.050$) and low positive correlation with goals scored ($r_{16} = 0.392, p = 0.106$). In the case of S_{vaep}, there were high positive correlation with winning points ($r_{16} = 0.830, p < 0.001$) and very high positive correlation with goals scored ($r_{16} = 0.953, p < 0.001$). It is obvious that S_{vaep} can accurately predict the number of goals scored in a match because it is based on the prediction of scores. Interestingly, even though S_{vaep} is also based on the prediction of conceded goals, it had slight almost negligible relationships with goals conceded ($r_{16} = -0.040, p > 0.05$). On the other hand , R_{vdep} had low correlation with the goals scored in the game ($r_{16} = -0.245, p > 0.05$).

Second, we performed the correlation analysis between the team's performance over the whole season and the evaluation indices. R_{vdep} had moderate positive correlations with winning points ($r_{16} = 0.397, p = 0.103$), and low correlation with goals scored ($r_{16} = 0.342, p = 0.162$) and goals conceded ($r_{16} = -0.291, p = 0.239$) . Meanwhile, S_{vaep} had moderate positive correlation with goals scored ($r_{16} = 0.497, p = 0.034$), but slight almost negligible relationships with winning points ($r_{16} = 0.177, p > 0.05$) and goals conceded ($r_{16} = -0.098, p > 0.05$). In the case of VDEP, the correlation coefficients with the game performances and those with the entire season were similar, whereas, in VAEP, the associations were very different.

Conclusion

In this study, we proposed a method to comprehensively evaluate a team's defense related to the team's performance, based on the prediction of ball recovery and being attacked, which occur more frequently than goals, using player actions and positional data of all players and the ball. We computed the F1 score and the results showed that the VDEP method predicted true positives correctly, while the VAEP did not. This suggests that the VDEP method was a reliable method that can evaluate defensive performances based on accurate predictions.

Regarding the team evaluations using the proposed and existing indices, the results of correlation analysis suggest that R_{vdep} could be a well-balanced indicator to evaluate both attacks (after the ball recovery) and defense itself (prevention of being attacked and the ball recovery). On the other hand, the VAEP method [1] is based on the prediction of offensive play and shows no correlation with the goals conceded. We expect that the use of VDEP in addition to the various indicators used so far will lead to the continuous strengthening of the team, regardless of immediate wins and losses which would be associated with contingent factors.

One possible future research direction is the determination of the weighting constant C in Equation 32 for ball recovery and being attacked. Although this study determined C based on the number of occurrences of both events, the constant should be determined in more suitable ways for the practical values in soccer.

Acknowledgments

The data used in this research was provided by the Research Center for Medical and Health Data Science in the Institute of Statistical Mathematics and Data Stadium Inc. We would like to thank Atom Scott and Masaki Onishi for their valuable comments on this work. This work was supported by JSPS KAKENHI (Grant Numbers 19H04941, 20H04075, 20H04087) and JST Presto (Grant Number JPMJPR20CA).

References

[1] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861, 2019.

[2] J. Fernandez and L. Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Proceedings of the 12th MIT Sloan Sports Analytics Conference*, 2018.

[3] K. Fujii. Data-driven analysis for understanding team sports behaviors. *arXiv preprint arXiv:2102.07545*, 2021.

[4] K. Fujii, Y. Inaba, and Y. Kawahara. Koopman spectral kernels for comparing complex dynamics: Application to multiagent sport plays. pages 127–139. Springer, 2017.

[5] K. Fujii, T. Kawasaki, Y. Inaba, and Y. Kawahara. Prediction and classification in equation-free collective motion dynamics. *PLoS Computational Biology*, 14(11):e1006545, 2018.

[6] K. Fujii, N. Takeishi, M. Hojo, Y. Inaba, and Y. Kawahara. Physically-interpretable classification of network dynamics for complex collective motions. *Scientific Reports*, 10(3005), 2020.

[7] K. Fujii, Y. Yoshihara, Y. Matsumoto, K. Tose, H. Takeuchi, M. Isobe, H. Mizuta, D. Maniwa, T. Okamura, T. Murai, et al. Cognition and interpersonal coordination of patients with schizophrenia who have sports habits. *PLoS One*, 15(11):e0241863, 2020.

[8] S. Llana, P. Madrero, J. Fernández, and F. Barcelona. The right place at the right time: Advanced off-ball metrics for exploiting an opponent's spatial weaknesses in soccer. In *Proceedings of the 14th MIT Sloan Sports Analytics Conference*, 2020.

[9] I. G. McHale, P. A. Scarf, and D. E. Folker. On the development of a soccer player performance rating system for the english premier league. *Interfaces*, 42(4):339–351, 2012.

[10] J. Piersma. Valuing defensive performances of football players. *Master Thesis in Erasmus School of Economics*, 2020.

[11] P. Power, H. Ruiz, X. Wei, and P. Lucey. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1605–1613, 2017.

[12] P. Robberechts. Valuing the art of pressing. In *StatsBomb Innovation in Football Conference*, 2019.

[13] W. Spearman. Beyond expected goals. In *Proceedings of the 12th MIT Sloan Sports Analytics Conference*, pages 1–17, 2018.

[14] T. Taki and J.-I. Hasegawa. Visualization of dominant region in team games and its application to teamwork analysis. In *Proceedings of Computer Graphics International*, page 227–235, 2000.

[15] M. Teranishi, K. Fujii, and K. Takeda. Trajectory prediction with imitation learning reflecting defensive evaluation in team sports. In *IEEE 9th Global Conference on Consumer Electronics (GCCE 2020)*, pages 124–125, 2020.

[16] K. Toda, M. Teranishi, K. Kushiro, and K. Fujii. Evaluation of soccer team defense based on prediction models of ball recovery and being attacked. *arXiv preprint arXiv:2103.09627*, 2021.

[17] F. Ueda, M. Honda, and H. Horino. The causal relationship between dominant region and offense-defense performance - focusing on the time of ball acquisition. In *Football Science*, volume 11, pages 1–17, 2014.

[18] Y. Yamamoto and K. Yokoyama. Common and unique network dynamics in football games. *PloS one*, 6(12):e29638, 2011.

[19] K. Yokoyama, H. Shima, K. Fujii, N. Tabuchi, and Y. Yamamoto. Social forces for team coordination in ball possession game. *Physical Review E*, 97(2):022410, 2018.

[20] K. Yokoyama and Y. Yamamoto. Three people can synchronize as coupled oscillators during sports activities. *PLoS Comput Biol*, 7(10):e1002181, 2011.

Classification of Japanese professional baseball players using a Gaussian mixture clustering model

Taishi Oda* and Nobuyoshi Hirotsu**

*Juntendo University, Inzai, Japan + email address: taishiflower1010@yahoo.co.jp
**Juntendo University, Inzai, Japan + email address: nhirotsu@juntendo.ac.jp

Abstract

The purpose of this study is to present a framework for grouping players in Nippon Professional Baseball (NPB). In order to group them, we could apply a method proposed by Soto-Valero (2017). Using data from FIFA's official website, he utilized principal component analysis (PCA) on 7,705 European footballers, and shrank their information such as Dribbling, Handling and Kicking into two variables. Then, he conducted cluster analysis using a Gaussian mixture model (GMM). After classifying them into four clusters consisting of players who have similar characteristics, excellent players were identified among them. In our study, we use a similar method. We apply PCA to 129 baseball players and shrink their performances into a small number of variables. Then, we conduct cluster analysis by GMM using 127 indexes such as Slugging percentage, On-base percentage and On-base plus slugging in the 2020 season taken from website "1.02 ESSENCE OF BASEBALL" provided by DELTA Co., Ltd. In each group, we evaluate the players according to their principal component scores corresponding to their characteristics and discuss the usefulness of this method.

1 Introduction

Baseball is one of the sports played worldwide. There are many professional baseball teams from Europe to Asian countries, including the country of origin, the United States. Of course, Japan is no exception. NPB (Nippon Professional Baseball), which manages professional baseball in Japan, has two leagues, the Central League and the Pacific League, and each league has 6 teams, for a total of 12 teams. They compete for the league title every year.

There are various types of players belonging to NPB. For fielders, there are batters like Kazuma Okamoto (currently playing in Yomiuri Giants) who perform a high hitting ability as a fourth batter. On the other hand, there are also players called “lead-off man” like Koji Chikamoto (currently playing in Hanshin Tigers) who have high batting averages and on-base percentage. However, the boundaries of these players' type are very vague. Therefore, the purpose of this study is to present a framework for grouping players and clarify the difference of the type of players in NPB. Soto-Valero (2017) provided a typical example of a player classification using a large dataset and presented a framework for grouping soccer players. He conducted principal component analysis (PCA) on a dataset of more than 10,000 European players available from the official FIFA website (http://sofifa.com/) and shrank information on 40 variables into a small number of variables. Then, he conducted cluster analysis by Gaussian mixture model (GMM). Furthermore, using the results of the cluster analysis as the objective variable and other variables as the explanatory variables, he also conducted gradient boosting decision tree analysis to investigate which index influences the results of the cluster analysis. In this study, we did a similar study to Soto-Valero by using baseball data. First, we apply PCA on a dataset available from the service “One Point Zero Two” (https://1point02.jp/op/index.aspx) provided by DELTA Co., Ltd. The goal of PCA is to replace a large number of possible correlated variables with a much smaller set of uncorrelated variables, while capturing as much information in the original variables as possible.

Second, we apply cluster analysis by Gaussian mixture model to the principal component score. The good point of this model is that the optimum number of clusters and the type of covariance matrix are selected based on the Bayesian information criterion called the BIC value. This eliminates the need to select the number of clusters as the k-means method. Finally, using the results of cluster analysis as the objective variable, we clarify the features of each cluster using a normal decision tree. In the previous research, Soto-Valero, the gradient boosting decision tree, which is said to be more accurate, is used, but we would like to make which model to use a future research subject, including the examination of the validity of the model. Therefore, this time, we use the regression tree, which is the simple model.

2 Dataset

The data used are 127 batting and defensive results recorded by fielders who come to bat 157 or more times among the players who participated in the official games in 2020. The number of players who come to 157 or more times is a maximum value that does not exceed the number of indicators. When performing PCA analysis, the number of observations (number of players) must be larger than the number of original variables. As described in the results of the decision tree analysis, Table 1 shows the 15 indexes with the highest features, along with the mean and standard deviation, as typical examples. For missing values, “0” is substituted for all values this time.

3 Procedures

Using this data set, we conduct PCA and shrink information on all the variables into two variables. Then we conduct cluster analysis by Gaussian mixture model. Furthermore, using the results of the cluster analysis as the objective variable and other variables as the explanatory variables, we also conduct decision tree analysis. All the experiments are developed using the R statistical computing software (version 3.5.1).

3.1 Principal components analysis

One of the most difficult problems in performing multivariate analysis is the reduction and selection of variables used in the analysis. This is a method of synthesizing a small number of uncorrelated variables called principal components that best represent the overall variation from a large number of correlated variables.

Assuming that $X_{n\times p}$ is a dataset consisting of n individuals and p variables, the composite variable is the following linear combination equation that contracts p-dimensional data to a lower k-dimensional ($k \leq p$).

$$z_j = a_{1,j}x_1 + a_{2,j}x_2 + a_{3,j}x_3 + \cdots + a_{p,j}x_p \quad (j = 1, \cdots, k)$$

The coefficient $a_{i,j}(i = 1, \cdots, p)$ at this time is called the main component. In PCA, this principal component is obtained under the constraint of $\sum_{i=1}^{p} a_{i,j} = 1$ so that the variance of z_j is maximized.

	Mean	SD
ISO	0.14	0.06
HR/FB	9.15	5.97
HR	8.92	7.77
SLG	0.4	0.08
SH	4.13	5.02
Zone%	44.63	2.95
GB%	46.63	7.12
GB/FB	1.13	0.39
FB%	43.87	7.65
Mid%	41.96	6.96
Spd	3.32	1.91
wSB	0.48	0.64
GDP	6.23	4.18
wCH	1.68	1.67
wCH/C	2.35	2.12

Table 1: Examples of typical indicators

3.2 Gaussian finite mixture model-based clustering

Mixture clustering is a method of finding the parameters of the original probability distribution, assuming that the continuous variables at hand are generated from several different probability distributions.

Now, let G probability density functions be $f_1(x;\theta_1),\cdots,f_G(x;\theta_G)$, and their mixed ratios be $\pi_1,\cdots,\pi_G$. However, $\theta_g(g=1,\cdots,G)$ is a vector consisting of parameters included in the probability (density) function $f_g(x;\theta_g)$. For the mixing ratios $\pi_1,\cdots,\pi_G, 0 \leq \pi_g \leq 1(g=1,\cdots,G), \sum_{g=1}^{G}\pi_g = 1$ shall be satisfied. At this time, the probability (density) function of the mixture distribution model is given as follows.

$$f(x;\theta) = \sum_{g=1}^{G} \pi_g f_g(x;\theta_g)$$

The EM algorithm is used to estimate the parameters $\theta = (\theta_1^T,\cdots,\theta_G^T,\pi_1,$ $\cdots,\pi_{G-1})^T$ included in this model. Cluster analysis can also be performed using a mixture distribution model. Conditional expectation used in the E step of the EM algorithm (see reference [5] P178,179) for which data each observation belongs to,

$$\gamma_{ig} = E(Z_{ig}|x_i) = Pr(Z_{ig}=1|x_i)$$

$$Pr(Z_{ig}=1|x_i) = \frac{\pi_g f_g(x_i;\theta_g)}{\sum_{h=1}^{G}\pi_h f_h(x_i;\theta_h)}$$

The i-th observed value is classified into the component that maximizes the estimated value of these formulas.

3.3 Decision tree

Decision tree analysis is a method of dividing data in stages and outputting tree-like analysis results. It is an algorithm that sets a branch according to the condition, traces from the root, and divides into the one that best meets the condition. When the result of cluster analysis is used as the objective variable, it is possible to know how much the explanatory variables such as slugging percentage and on-base percentage affect the classification result.

4 Results

We show the results of PCA, cluster analysis and decision tree analysis following the above procedure.

4.1 Principal components analysis

Table 2 below shows the results of PCA. It represents up to the fifth principal component.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	5.178	3.791	2.780	2.434	2.360
Proportion of Variance	0.211	0.113	0.061	0.047	0.044
Cumulative Proportion	0.211	0.324	0.385	0.432	0.476

Table 2: Results of PCA

Looking at Table 2, the cumulative proportion rate up to the second principal component is only 30%, but since the previous study Soto-Valero uses up to the second principal component, we use up to the second principal component in this study as well.

4.2 Gaussian finite mixture model-based clustering

The results of applying a mixture model using these principal component scores are as follows. The horizontal axis is the first principal component score calculated with the obtained principal component, and the vertical axis is the second principal component score.

4.3 Decision tree

Then, we apply to the classification results as the objective variable and the other variables as the explanatory variables. Figure 2 shows the attributes' importance.

5 Further study

In this study, we were able to present a framework for grouping players using PCA, Gaussian mixture model, and decision tree analysis, and to clarify the differences in the types of athletes in NPB. In the future, we will conduct three main analyses to develop this study.

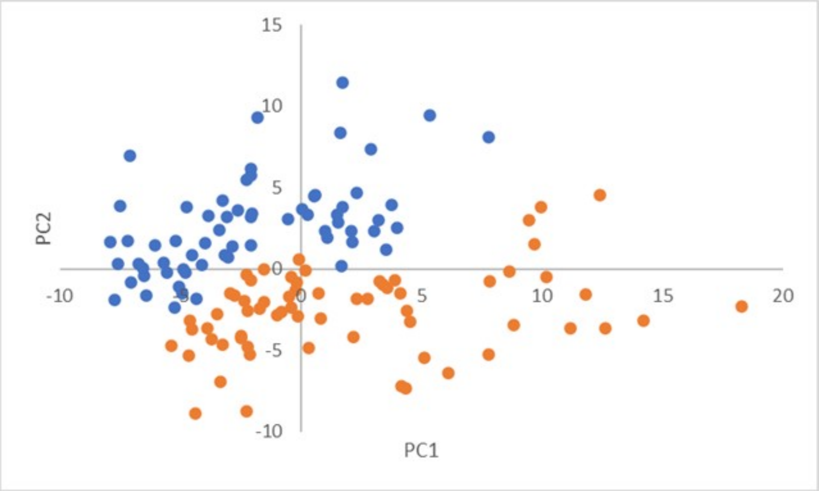


Figure 1: Scatter plot of principal component scores showing classification.

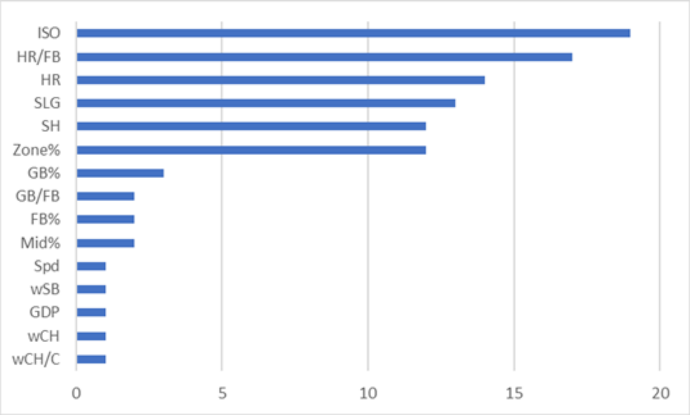


Figure 2: Importance of attributes

First, this time we used the indexes recorded by the fielders, but we would like to analyse the indexes recorded by the pitchers as well. Second, “0” is substituted to make up for the missing values in the dataset this time, but we would like to try other methods such as the average substitution method and the listwise method. Finally, in calculating the features, we used decision tree, the simple model, but next time we would like to use the gradient boosting decision tree as Soto-Valero.

Acknowledgements

This study was supported by Grants-in-Aid for Scientific Research (C) of Japan (No.19K04911). The used data was provided by Delta Inc.

References

[1] BA Bushong, "Fuzzy clustering of batting averages", NAFIPS 2006-2006 Annual Meeting of the North American Fuzzy Information Processing Society (2006).

[2] BA Bushong, "Fuzzy clustering of baseball statics", NAFIPS 2007-2007 Annual Meeting of the North American Fuzzy Information Processing Society (2007).

[3] DELTA Co., Ltd., "Online Baseball Analysis Course "Baseball Analytics Now"", https://peatix.com/event/1561186/view (viewed on August 5, 2020).

[4] K Imaizumi, Y Nagao, G Yoshida & M Taguchi, "Analysis of the strength of teams participating in Koshien by cluster analysis - using the results of regional qualifiers", the 47th Annual Conference of Japan Society of Physical Education (1996). (in Japanese)

[5] H Matsui & K Koizumi, "Statistical model and guess", Koudansha, Tokyo (2020). (in Japanese)

[6] S Miyamoto, "Introduction to Cluster Analysis-Theory and Application of Fuzzy Clustering", Morikita Publishing, Tokyo (2006). (in Japanese)

[7] K Nishiuchi, "With Yasuhito Endo, the team will have 117% of points", SoftBank Shinsho, Tokyo (2012). (in Japanese)

[8] C Soto-Valero, "A Gaussian mixture clustering model for characterizing football players using the EA Sports’ FIFA video game system", RICYDE. Revista International de Ciencias del Deporte (2017).

Analysing football tactics using high-frequency tracking data

Marius Ötting* and Dimitris Karlis**

*Bielefeld University + email address: marius.oetting@uni-bielefeld.de
** Athens University of Economics and Business + email address: karlis@aueb.gr

1 Introduction

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.

Whereas tracking data sets have previously been investigated in other (mostly US) sports such as basketball and American football (see, e.g., [2, 7]), corresponding analyses of association football are fairly rare. Such analyses in football cover passing performance (see, e.g., [5, 6]), team formation (see, e.g., [3, 8]), and space creation [1]. [4] provide an overview over the existing literature on football tracking data.

In this paper, 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 excluding the goalkeeper. This metric 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. The EPS relates to several game characteristics. First, the team formation results in different convex hull sizes as a result of the positions of the players. Second, it also relates to the way the team attacks. Third, when relating the EPS to specific events of the game (such as goals), we can infer how goal scoring chances are created. Fourth, we observe the convex hull as a fine-grained time series which helps to investigate the speed of transition from defence to offence.

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 allow for within-state dependence of the two teams' EPS, we formulate multivariate state-dependent distributions using copulas.

2 Data

The high-resolution data set considered here covers the (x, y) positions of all players and the ball with a resolution of 25 Hz. Since the ball is out of play several times in a football match, we consider only those observations where the ball was in play. An example situation of the match analysed is given in Figure 1, which shows all players, the corresponding convex hulls, and the ball. In the full sample, the means of the EPS are 1352 for the blue team (min: 98, max: 2512) and 1244 for the black team (min: 114, max: 3156). We calculate the size of the convex hull (i.e. the EPS) for both teams at each time point where the ball was in play, which is our main quantity of interest here. The corresponding bivariate time series is shown in Figure 2.

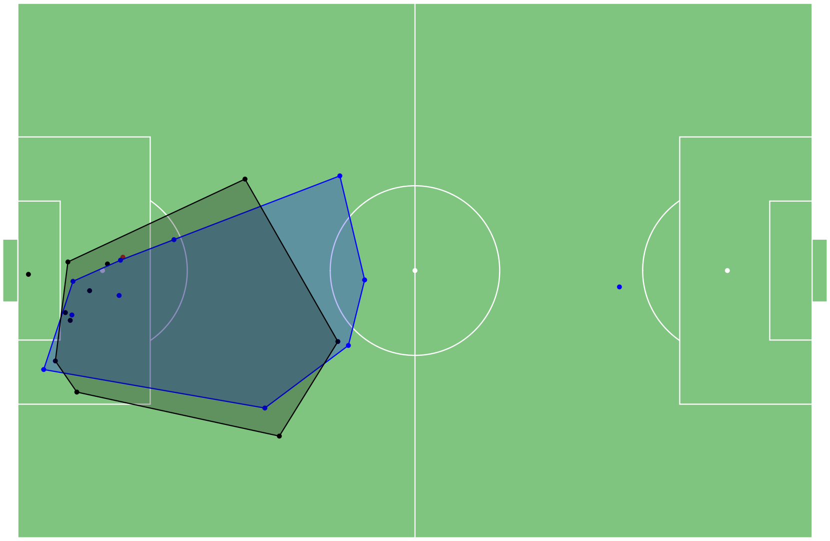


Figure 1: Example situation found in our data. Team blue (playing from right to left) is attacking. The sizes of the convex hulls are 785 square meters for the blue team and 800 square meters for the black team.

3 Modelling EPS using hidden Markov models

Figure 2 further underlines that there are periods in the match where both teams' EPS are fairly high (e.g. after the halftime break), as well as periods where one team has a fairly low EPS, while we observe a high EPS for the other team (e.g. at the end of the match). HMMs thus constitute a natural modelling approach for our time series data, as they accommodate the idea of a match progressing through different phases, with potentially changing tactics of the two teams.

In the basic model formulation, HMMs involve two components: an unobserved Markov chain with N possible states, and an observed state-dependent process, whose observations are assumed to be generated by one of N distributions as selected by the Markov chain. As the EPS is positive and continuous-valued, we consider the Gamma distribution here. Moreover, to account for potential within-state dependence between the two teams' EPS, we model the teams' EPS jointly by assuming a bivariate Gamma distribution for each state. This bivariate Gamma distribution is created using a copula. We thus aim at using a distribution with joint density

$$f(x,y) = f_1(x) f_2(y) c_\theta(F_1(x), F_2(y))$$

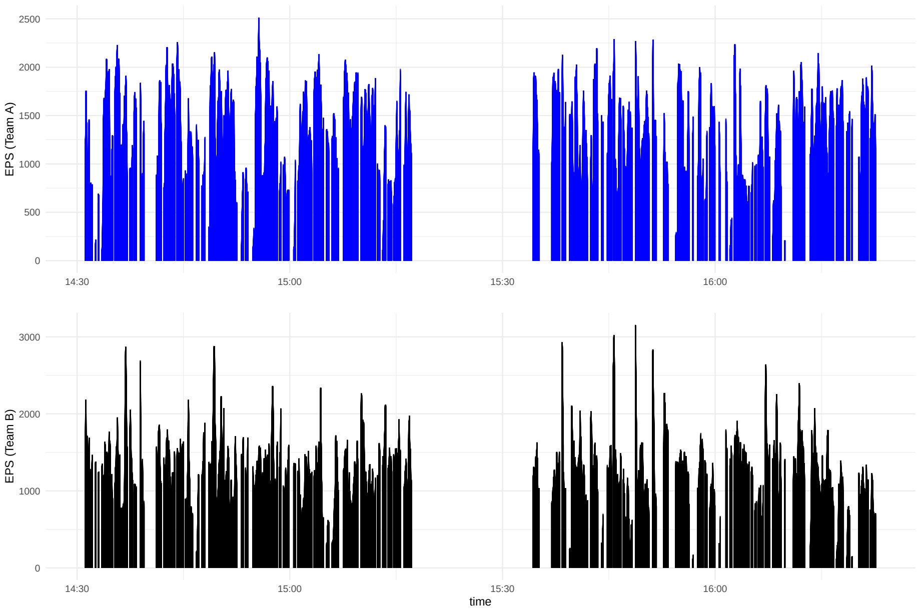


Figure 2: Bivariate time series of the two teams' EPS (top: Team A, bottom: Team B).

where

$$c_\theta(u,v) = \frac{\theta(e^\theta - 1)e^{\theta(1+u+v)}}{\left[e^\theta - e^{(\theta+\theta u)} + e^{\theta(u+v)} - e^{(\theta+\theta v)}\right]^2}$$

i.e. the density of a Frank copula with dependence parameter $\theta \in \mathscr{R} \setminus \{0\}$. In our case we assume that the marginals are Gamma distributions and hence

$$f_j(x) = \frac{\beta_j^{\alpha_j} x^{\alpha_j - 1}}{\Gamma(\alpha_j)} \exp(-\beta_j x)$$

for $j = 1,2$. The HMM likelihood is then maximised numerically in R (for further details see [10]).

4 Results

We fit a three-state copula-based HMM to the tracking data. Since the parameters of the Gamma distribution are somewhat cumbersome to interpret, Table 1 displays the corresponding means and standard deviations of the estimated state-dependent distributions. We observe a higher mean of the EPS for Team A in states 1 and 2, while in state 3 Team B's mean EPS is larger. In addition, we observe a negative dependence in state 1, and a positive dependence in states 2 and 3. Based on further investigations of the fitted model (not shown here), we found that state 1 refers to situations where Team A is in forward play and Team B is defending. In state 2, Team A is in forward play and Team B is pressing, whereas in state 3 Team B is in forward play.

We further investigate typical tactical patterns according to the state process. For that purpose, we use the Viterbi algorithm to investigate the most likely trajectory of the states. This allows to analyse the (median)

Table 1: Parameter estimates for the state-dependent distributions of the Frank-copula HMM with three states.

Variable	State 1	State 2	State 3
EPS (Team A)	$\hat{\mu} = 1233, \hat{\sigma} = 203$	$\hat{\mu} = 1781, \hat{\sigma} = 175$	$\hat{\mu} = 1071, \hat{\sigma} = 489$
EPS (Team B)	$\hat{\mu} = 882, \hat{\sigma} = 340$	$\hat{\mu} = 1271, \hat{\sigma} = 199$	$\hat{\mu} = 1631, \hat{\sigma} = 406$
Dependence	$\hat{\theta} = -0.569$	$\hat{\theta} = 3.310$	$\hat{\theta} = 8.288$

Table 2: Median number of time points spent in each state.

median number of consecutive time points spent in state 1:	157
median number of consecutive time points spent in state 2:	265
median number of consecutive time points spent in state 3:	308

number of consecutive time points spent in each state (shown in Table 2), which reveal some further tactical insights. Based on these summary statistics, we can, for example, compare how the two teams behave in their forward play. For states 1 and 2, which refer to Team A being in forward play, we see that the median number of consecutive time points spent in this state is lower compared to state 3, which refers to Team B's forward play. In other words, if Team B is in forward play, their main aim is potentially to control the match by keeping possession of the ball, whereas Team A's forward play is associated with more pressure on goal.

5 Outlook

In this paper, we provide a case study on modelling high-resolution football tracking data. The variables to be modelled are the two teams' EPS. As a team's tactics — and therefore the corresponding EPS — changes during a match, we consider HMMs to analyse our data. While the results indicate clearly distinguishable states (and hence tactics), the modelling approach considered here could be further extended. For example, in future research other response variables such as indicators for pressing or space control could be considered (see [9]). In addition, covariates such as players' position on the field could be included. Still, using a simple bivariate HMM as considered here, we are able to provide useful insights into teams' tactics, which may be of great interest to managers and sports fans.

References

[1] Fernandez, J. and Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. *Sloan Sports Analytics Conference*.

[2] Franks, A., Miller, A., Bornn, L., Goldsberry, K., et al. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *Annals of Applied Statistics*, 9(1):94–121.

[3] Frencken, W., Lemmink, K., Delleman, N., and Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. *European Journal of Sport Science*, 11(4):215–223.

[4] Goes, F., Meerhoff, L., Bueno, M., Rodrigues, D., Moura, F., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., et al. (2020). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*.

[5] Goes, F. R., Kempe, M., Meerhoff, L. A., and Lemmink, K. A. (2019). Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big Data*, 7(1):57–70.

[6] Kempe, M., Goes, F. R., and Lemmink, K. A. (2018). Smart data scouting in professional soccer: Evaluating passing performance based on position tracking data. In *2018 IEEE 14th International Conference on e-Science*, pages 409–410. IEEE.

[7] Lopez, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: an introduction to a special issue on tracking datain the National Football League. *Journal of Quantitative Analysis in Sports*, 16(2):73–79.

[8] Memmert, D., Raabe, D., Schwab, S., and Rein, R. (2019). A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs. 11 game set-up. *PLoS ONE*, 14(1):e0210191.

[9] Memmert, D. and Rein, R. (2018). Match analysis, big data and tactics: current trends in elite soccer. *German Journal of Sports Medicine/Deutsche Zeitschrift fur Sportmedizin*, 69(3):65–72.

[10] Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman & Hall/CRC.

Investigating drivers of stakes in a football live betting market

Marius Ötting* and Rouven Michels** and Roland Langrock*** and Christian Deutsch

*Bielefeld University + email address: marius.oetting@uni-bielefeld.de
Bielefeld University + email address: r.michels@uni-bielefeld.de ***Bielefeld University**Bielefeld University + email address:

1 Introduction

Gambling and sports betting markets have grown very rapidly in recent years, as illustrated by an increased gross gaming revenue. This value, which measures the amount of money bookmakers win from their clients, increased to almost 100 billion euro in 2019 in Europe. As one key part of betting markets, the live betting market has seen a rise in popularity as this market nowadays accounts for about 55% of the overall volume [3].

However, in contrast to the pre-game betting market, for which betting behaviour and market inefficiencies have been intensively studied, the live betting market to date has not nearly as thoroughly been studied, likely caused by the fact that appropriate data is not publicly available. Previous work in the literature on live betting markets investigated stakes placed in betting exchanges or live betting odds. However, to the best of our knowledge, we are the first to investigate stakes placed in live betting markets at a regular bookmaker.

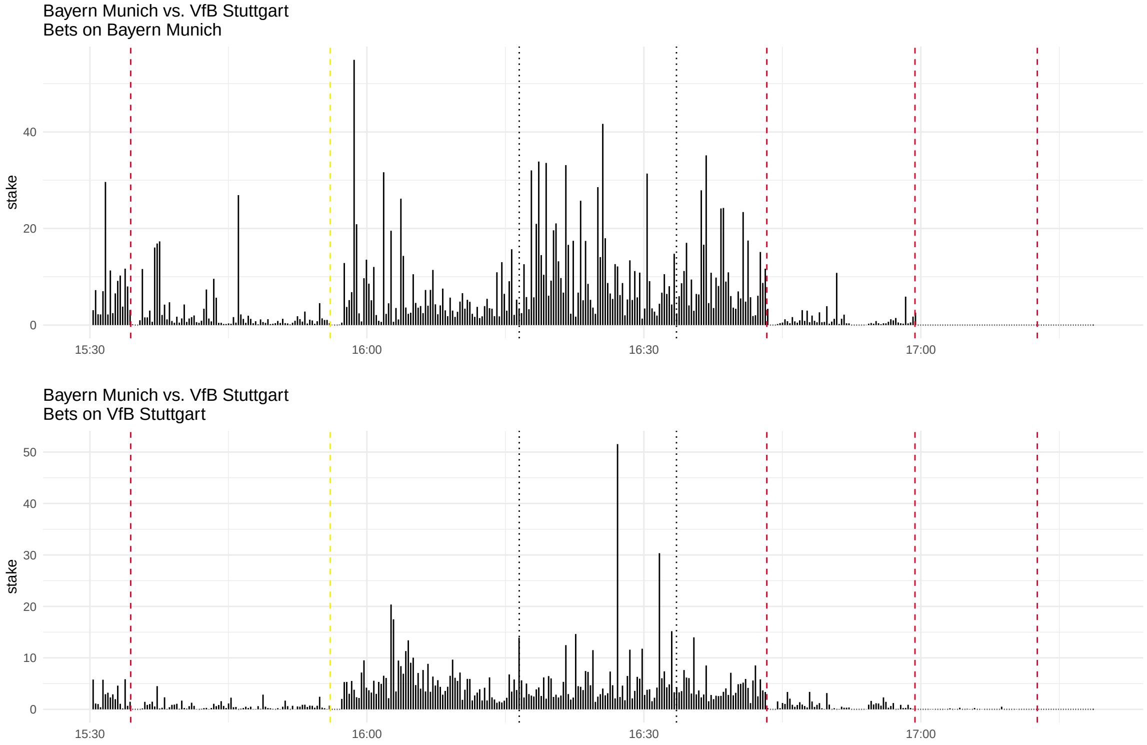
We consider a unique high-resolution data set on bets placed during football matches. Further information on matches, such as goals scored, are also included in the data.

With these detailed information at hand, our high-resolution data set enables a fine-grained analysis of live betting markets with the aim to investigate patterns in betting behaviour. Specifically, accounting for the general market activity level within a state-space modelling framework, we focus on the market's response to events such as goals, but also to the general situation within a match, e.g. the uncertainty about the game outcome.

2 Data

We consider high-resolution live betting data provided by a large European bookmaker. The data comprises second-by-second stakes for all 306 matches of the 2018/19 season in the German Bundesliga, i.e. the highest German football league. We split each match into two time series, such that all bets placed on a win of the respective team per match, serving as our response variable, are collected in one time series. Further, we aggregate the stakes into 15-second intervals to obtain regular time series. Hence, each match consists of a minimum of 420 intervals per match. As we also include the halftime as well as injury times, the exact number of intervals differs across matches. Moreover, the number of stakes have been multiplied with a constant beforehand as we are not allowed to provide any kind of information regarding the actual number of stakes.

Figure 1 shows two example time series for one match — Bayern Munich playing at home against VfB Stuttgart on January 27, 2019. These time series suggest various patterns with respect to how stakes are being


Bayern Munich vs. VfB Stuttgart
Bets on Bayern Munich

Bayern Munich vs. VfB Stuttgart
Bets on VfB Stuttgart

Figure 1: Time series of the stakes placed for one example match (Bayern Munich vs. VfB Stuttgart). Top: stakes placed on Bayern Munich. Bottom: stakes placed on VfB Stuttgart.

placed depending on the development of the match. In particular, the market strongly reacts to goals being scored — indicated by the red (for Bayern Munich) and yellow (for VfB Stuttgart) dashed lines.

The covariates we consider can be split into two groups: static and dynamic covariates. Regarding the first group, we consider the strength of teams [4], which are proxied by the market values of the teams (taken from www.transfermarkt.com). Further, we assume betting volumes to differ across weekdays [5] and hence account for kick-off times.

Dynamic covariates can be split into two groups as ’dynamic’ can relate to the market status as well as the game status. As stakes cannot be placed if the bookmaker closes the market (e.g. after goals or red cards), we include a dummy variable indicating whether the market is open or not in interval t. We further consider the number of remaining intervals in the match as well as the live betting odds in interval t. Taking into account the bookmaker’s margin, these odds can be transformed into winning probabilities. The inequality between these probabilities — represented by the Gini coefficient — is then taken into account for the imbalance of the match. Here, a Gini coefficient of zero refers to a fully balanced match, whereas a Gini coefficient of one refers to an decided match. Next to the Gini coefficient, we also generate a covariate where we subtract the Gini coefficient in interval t from the Gini coefficient at the beginning of the match. This covariate serves to measure rather surprising match courses.

We further include a covariate for the goals scored of the team analysed as well as for the opponent, and a dummy variable for the halftime break. The existing literature [1] indicates that the amount of stakes placed tends to increase during halftime. All covariates are briefly summarised in Table 1.

Table 1: Descriptive statistics of the variable analysed (stake), as well as the *static* and *dynamic* covariates.

	mean	st. dev.	min.	max.
stake	3.755	8.347	0	1027
market values (*mvteam* and *mvopp*)	203.3	180.8	42.88	835.6
matches played on Friday (*friday*)	0.095	–	0	1
matches played on Saturday afternoon (*saturdayaft*)	0.513	–	0	1
matches played on Saturday evening (*saturdayeve*)	0.101	–	0	1
matches played on Sunday (*sunday*)	0.216	–	0	1
matches played on weekday (reference category)	0.075	–	0	1
status (*open*)	0.908	–	0	1
halftime (*halftime*)	0.149	–	0	1
remaining intervals (*int*)	210.1	121.7	0	472
intervals since last goal (*goalteam* and *goalopp*)	47.85	79.22	0	422
Gini coefficient (*gini*)	0.544	0.297	0	0.993
difference between the Gini at interval t and interval 1 (*ginidiff*)	0.205	0.292	-0.794	0.916

3 Model

The bottom panel in Figure 1 indicates clear phases of rather low (e.g. before 16:00) and high betting activity (e.g. around 16:30). To capture such serial correlation we assume the general betting behaviour to be driven by an unobservable (latent) market activity level. Specifically, we use (covariate-driven) state-space models (SSM) which comprise a continuous-valued state process corresponding to the latent market activity level. The SSM considered here further consists of an observed state-dependent process — in our case the stakes placed — which are assumed to be driven by the unobserved state process.

We use an SSM where the observed stakes follow a zero-adjusted gamma distribution (ZAGA) [7]:

$$y_t \sim \text{ZAGA}(\mu_t, \sigma, \pi_t), \;\; \text{with } f(y_t) = \begin{cases} \pi_t, \text{ if } y_t = 0; \\ (1-\pi_t)h(y_t), \text{ if } y_t > 0, \end{cases} ,$$

with $h(t)$ the density of the Gamma distribution. We use the parametrisation of the gamma distribution with mean μ_t and standard deviation σ instead of scale and shape [7]. The parameter π_t gives the probability of no bets being placed in interval t.

Intervals without any bets placed mostly occur while the market is closed, e.g. after goals, when odds are updated. However, intervals without stakes placed are also observed although it is possible to place a bet. We see in the data that this almost exclusively happens when matches seem to be clearly decided. Hence, we incorporate the Gini coefficient and the remaining number of intervals, as well as an interaction term between both as covariates such that the predictor for π_t is given by

$$\pi_t = \text{logit}^{-1}(\gamma_0 + \gamma_1 gini_t + \gamma_2 gini_t^2 + \gamma_3 int_t + \gamma_4 gini_t \cdot int_t + \gamma_5 gini_t^2 \cdot int_t)$$

if $\text{open}_t = 1$, and $\pi_t = 1$ otherwise, i.e. if $\text{open}_t = 0$. Further, we incorporate exogenous static covariates for the state-dependent process and dynamic covariates for the state process. The static covariates are included into the predictor for μ_t:

$$\begin{aligned}\mu_t = \exp\big(&\alpha_0 + \alpha_1 \log(mvteam) + \alpha_2 \log(mvopp) + \alpha_3 \log(mvteam) \cdot \log(mvopp) \\ &+ \omega_1 friday + \omega_2 saturdayaft + \omega_3 saturdayeve + \omega_4 sunday + g_t\big),\end{aligned}$$

whereas the dynamic covariates are associated with in-game dynamics. We mainly focus on goals scored as well as on the difference between the winning probabilities in interval t and at kick-off. As we assume in-game dynamics to affect the latent market activity level, we include the corresponding covariates into the latent process g_t:

$$g_t = \phi g_{t-1} + \beta_1 halftime_t + \beta_2 ginidiff_t + \beta_3 \tfrac{1}{goalteam_t} + \beta_4 \tfrac{1}{goalopp_t} + \omega \eta_t.$$

One drawback of this approach is the non-existence of a standard software package to fit such models. Here, we make use of a combination of numerical integration and recursive computing (first suggested by [6]). Furthermore, following [8], we approximated the likelihood function and used a recursive algorithm to evaluate the approximate likelihood. Finally, we obtained parameter estimates by numerically maximising the likelihood in R using the function `nlm()`.

4 Results

The resulting parameter estimates are shown in Table 2. As the persistence parameter $\hat{\phi}$ is estimated to be close to one, the results indicate a strong serial correlation in the state process, i.e. the persistence in the underlying market activity level is fairly high. Further, our results are in general in line with the existing literature as the betting volume tends to increase during halftime ($\hat{\beta}_1 > 0$) [1] while less stakes are placed if further matches are played in parallel ($\hat{\omega}_2 < 0$), such as at Saturday afternoon, which is in-line with [2].

Moreover, goals scored by the team analysed tend to increase the market's activity level ($\hat{\beta}_3 > 0$), whereas the market's activity level tends to decrease when the team analysed concedes a goal ($\hat{\beta}_4 < 0$). While these estimated effects seem intuitively plausible, this is a remarkable result as bookmakers decrease the odds after goals have been scored. Hence, the results indicate the effect of goals to be greater than the effect of lowered odds.

Furthermore, unexpected match dynamics and surprising match courses tend to increase betting volumes ($\hat{\beta}_2 < 0$). This is intuitively plausible as bettors may prefer situations in matches in which, for example, weaker expected teams score an equalizing goal when playing against a clear favourite, rather than matches where the clear favourite scores a goal which puts them two or more goals in front.

5 Discussion

Despite the fact that our approach is explanatory in nature, it builds the basis for further investigations in this research area. Among other things, our model can be modified from a methodological point of view such that the underlying state process will be allowed to affect not only the stakes placed on the team analysed but also the stakes of the opposing team. This extension would lead to a framework with a two-dimensional

Table 2: Parameter estimates with 95% confidence intervals for the final model.

parameter	estimate	95% CI
ϕ	0.970	$[0.969; 0.972]$
ω	0.203	$[0.198; 0.208]$
σ	0.870	$[0.869; 0.872]$
α_0	6.355	$[4.759; 7.95]$
$\alpha_1 \; \big(\log(mvteam)\big)$	-0.737	$[-1.119; -0.491]$
$\alpha_2 \; \big(\log(mvopp)\big)$	-1.482	$[-1.861; -1.237]$
$\alpha_3 \; \big(\log(mvteam) \cdot \log(mvopp)\big)$	0.228	$[0.179; 0.303]$
$\omega_1 \; \big(friday\big)$	0.04	$[-0.095; 0.180]$
$\omega_2 \; \big(saturdayaft\big)$	-0.246	$[-0.355; -0.136]$
$\omega_3 \; \big(saturdayeve\big)$	0.198	$[0.058; 0.339]$
$\omega_4 \; \big(sunday\big)$	0.047	$[-0.073; 0.166]$
γ_0	-4.438	$[-4.65; -4.225]$
$\gamma_1 \; \big(gini\big)$	-10.05	$[-10.72; -9.392]$
$\gamma_2 \; \big(gini^2\big)$	14.605	$[14.08; 15.13]$
$\gamma_3 \; \big(int\big)$	-0.542	$[-0.779; -0.305]$
$\gamma_4 \; \big(gini \cdot int\big)$	3.638	$[2.938; 4.338]$
$\gamma_5 \; \big(gini^2 \cdot int\big)$	-4.29	$[-4.814; -3.767]$
$\beta_1 \; \big(halftime\big)$	0.002	$[0; 0.005]$
$\beta_2 \; \big(ginidiff\big)$	-0.093	$[-0.098; -0.089]$
$\beta_3 \; \big(goalteam\big)$	0.246	$[0.224; 0.269]$
$\beta_4 \; \big(goalopp\big)$	-0.054	$[-0.075; -0.033]$

time series rather than two distinct time series. In particular, such a modification would result in replacing the autoregressive process by a vector-autoregressive process. One major advantage of this extension would be the possibility to capture dependencies between the unobserved market activity levels which might play a major role in the general betting behaviour.

Furthermore, our models could play a role in the area of betting fraud, which has grown rapidly in recent years reflected by the rather large number of match-fixing scandals. Our model might contribute to detecting fixed matches via the investigation of extreme outliers. Specifically, more focus would need to be placed on large residuals. If we observe large deviations between the actual stakes and the stakes expected by our model, this could be a first indicator for match-fixing.

References

[1] Croxson, K. and Reade, J. (2014). Information and efficiency: Goal arrival in soccer betting. *The Economic Journal*, 124(575):62–91.

[2] Deutscher, C., Ötting, M., Schneemann, S., and Scholten, H. (2019). The demand for English Premier League soccer betting. *Journal of Sports Economics*, 20(4):556–579.

[3] European Gaming & Betting Association (2020). European Online Gambling 2020 Edition.

[4] Feddersen, A., Humphreys, B. R., and Soebbing, B. P. (2017). Sentiment bias and asset prices: Evidence from sports betting markets and social media. *Economic Inquiry*, 55(2):1119–1129.

[5] Humphreys, B. R., Paul, R. J., and Weinbach, A. P. (2013). Consumption benefits and gambling: Evidence from the ncaa basketball betting market. *Journal of Economic Psychology*, 39:376 – 386.

[6] Kitagawa, G. (1987). Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041.

[7] Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. CRC press.

[8] Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman & Hall/CRC.

Predicting Tennis Outcomes Using Random Walks

Tomáš Kouřim*

*FNSPE CTU Prague + email address: tom@skourim.com

Abstract

Recently introduced random walk model showed its possible use in predicting tennis men Grand Slam tournaments outcomes. The model was then further examined and its theoretical background was described in more detail. In this paper the model is trained using the most recent data and then applied for in-play betting against a bookmaker. Different betting strategies are tested and the overall performance of the model is evaluated.

1 Introduction

Random walk was first introduced by K. Pearson over 100 years ago [4]. Since then the concept was further elaborated, the random walk was described in much detail and many alternatives of the original model were introduced. One of the most recent variations of a random walk is a random walk with varying probabilities, a concept introduced recently by the authors [2]. It is based on a similar concept of a random walk with varying step size introduced by L. Turban [5]. This model seems particularly well suited to model different sporting events, especially tennis matches.

A proper model of a tennis match, and more generally any sporting event, is of a great value. It can serve the athletes and their coaches to improve their strengths and eliminate weaknesses, it can bring new insides into the game and it is especially useful in the sports betting industry. Such model can be used by bookmakers to provide more accurate odds, it can used by professional bettors to gain an edge against the bookmakers and a robust model can also help to detect fraudsters selling their matches or betting on games where the result was agreed long before the game even started.

The paper is organized as follows. Next chapter briefly introduces the model used for tennis modeling. Section 3 provides general description of the data used, Section 4 describes the model fitting procedure. In Chapter 5 the model is tested against a database of real life bookmaker's odds and its performance is evaluated. Section 6 concludes this paper.

2 Random walk with varying probability

The idea behind the model of a random walk with varying probability is that the probability distribution of the next step depends directly on the result of the previous step in the following manner. Let us have a basic random walk with steps $X \in \{-1, 1\}$ with Bernoulli distribution depending on a parameter p so that $P(x=1) = p$ and $P(x=-1) = 1-p$. After each step this probability is altered using a memory coefficient λ and the result of previous step. We can define two major types of random walk with varying probability -

success punishing and *success rewarding*. The first denotes that every time an event occurs, the probability of its repetition decreases, i.e. if step $X_t = 1$ then the probability the next step $X_{t+1} = 1$ decreases by a parameter λ so that $p_{t+1} = \lambda p_t$. For the *success rewarding* model, the opposite holds and every time an event occurs, the probability of its repetition increases.

The model can be further refined by using two separate parameters, one for each direction, i.e. $\lambda = [\lambda_1, \lambda_2]$. Again, there can be *success punishing* and *success rewarding* model. Thus, together there are four basic model variants. Formal definitions of all four model variants and their detailed description together with the derivation of useful formulas can be found in [3].

2.1 Application for tennis modeling

The random walk model can be applied to model any discrete time random process. Such processes often occur in sports, especially in sports played not for certain amount of time, but for certain amount of points. These sports include for example volleyball, badminton or tennis. Tennis seems as particularly interesting sport in terms of discrete random processes. There are several such processes. Series of matches can be considered a discrete random process, the sets played within a match, games within a set, points within a game or even strokes within a point, they all can be considered discrete random processes. In this paper, the model of a random walk with varying probability was used to model the sets played within a match.

3 Data description

For the purpose of this paper an automatic data gathering tool was created. This tool was continuously scraping data from *www.tipsport.cz*, the website of the biggest Czech bookmaker, gathering provided odds for the winner of the first set of each match and, if available, also the *in-play* odds for each set played. The tool also stored the results of each particular set played. The tool was developed using the Python programming language with the help of Selenium framework. PostgreSQL database deployed on Digital Ocean infrastructure was used for data storage. For the purpose of this paper, data from February until May 2021 was available.

The data was split into training and testing datasets. Training dataset contains odds and results of matches played from February till the end of April, May matches represent the testing dataset. In the training dataset, there were 12 372 matches. Complete information, i.e. both odds and result, were available for 3 365 matches (mostly because the bookmaker did not provide *in-play* odds for the matches). As the matches are played as *best-of-three* (or *best-of-five* in case of 2021 ATP Australian Open) there were 8 310 sets played where all information required for the presented model was available. The testing dataset contains 3 076 matches, 1 150 of which with complete information, and there are 2 796 relevant sets.

4 Model training

As mentioned in Section 2, there are four basic variants of the model altogether with two (or three in case of $\lambda = [\lambda_1, \lambda_2]$) unknown parameters. Model training thus consists of two steps, best model variant selection and finding the optimal parameter.

4.1 Starting probability selection

The first unknown parameter of the model is the starting probability p_0. In tennis modeling, coefficient $p_0 = p_A$ means the probability that Player A will win the first set. This probability (or its estimate) is implicitly given by the odds $o = [o_A, o_B]$ provided by bookmaker. Roughly, $p_A \cong \frac{1}{o_A}$. This estimate can be further refined. First, for typical odds there holds that $\frac{1}{o_A} + \frac{1}{o_B} > 1$ (i.e., typically the bookmaker provides *subfair* odds with bookmaker's winning margin). First step is thus the normalization of odds. It turns out that the simple normalization, i.e. $p_A = n_1(o) = \frac{\frac{1}{o_A}}{\frac{1}{o_A}+\frac{1}{o_B}}$ does not provide reasonable results. In fact, bookmaker's margin is usually unevenly distributed towards the outsider. Therefore, an alternative normalization function can be used

$$p_A = n_2(o) = \frac{1}{o_A} + \frac{o_A o_B - o_A - o_B}{o_B(o_A + o_B)}. \tag{33}$$

This can be further improved by introducing a third normalization function $\mu(o,t)$, with a real parameter t linearly extrapolating n_1 and n_2 such that $\mu(o,0) = n_2(o)$ and $\mu(o,2) = n_1(o)$ [1].

To obtain reasonable estimation of the model parameter p_0 from provided odds, the first step is to estimate the parameter t. This was done using the maximal likelihood estimate method and the training set. Bookmaker's odds for the first set and the actual result of the first set were used. The likelihood function is defined as

$$L = \prod_{i=1}^{N} (x_i \mu(o_i, t) + (1 - x_i)(1 - \mu(o_i, t)),$$

where N is the number of matches in the training dataset, o_i is the bookmaker's odds for the first set in the $i-th$ match and x_i is the result of the first set in the $i-th$ match, $x_i = 1$ if Player A won the set, $x_i = 0$ otherwise. For computational reasons the *log-likelihood* $L_l = ln(L)$ was used, i.e. the function

$$L_l = \sum_{i=1}^{N} ln(x_i \mu(o_i, t) + (1 - x_i)(1 - \mu(o_i, t)).$$

Numerical methods, namely the function *minimize_scalar* from the Python SciPy package, were used to obtain optimal value of parameter $t = 4.03e - 06$. We can thus consider $t = 0$ and use function n_2 for normalization.

4.2 Model and memory parameter selection

For each match with data for at least 2 sets available (i.e. with at least 1 prediction opportunity), the first set winning probability p_0 was derived using the procedure and parameter from previous paragraph. Then, for each of the four model variants the optimal value of coefficient λ was found again using the training dataset and the maximum log-likelihood estimate

$$L_l = \sum_{i=1}^{N} \sum_{j=1}^{N_i} ln(x_{ij} p_{ij} + (1 - x_{ij})(1 - p_{ij})),$$

where N denotes the number of matches in testing dataset, N_i the number of predictable sets in $i-th$ match, x_{ij} is the result of the $j-th$ set in the $i-th$ match, $x_{ij} = 1$ if Player A won the set, $x_{ij} = 0$ otherwise and p_{ij} is the probability of Player A winning the set computed using the currently considered model variant.

Strategy	Bet	$E(w)$	$E(w\|p=\frac{1}{o})$	$Var(w)$	$Var(w\|p=\frac{1}{o})$
Naive	1	$po-1$	0	$po^2(1-p)$	$o(1-\frac{1}{o})$
Odds	$\frac{1}{o}$	$p-\frac{1}{o}$	0	$p(1-p)$	$\frac{1-\frac{1}{o}}{o}$
Prob.	p	$p(po-1)$	0	$p^3o^2(1-p)$	$\frac{1-\frac{1}{o}}{o}$
General	u	$u(po-1)$	0	$u^2po^2(1-p)$	$u^2o(1-\frac{1}{o})$

Table 1: Theoretical values of wins and variances for different betting strategies. p is the probability of winning, o is the odds provided by bookmaker.

Finally, the Akaike Information Criterion $AIC = 2k - 2L_l$, which considers the number of parameters to find the optimal model, was used. Here k is the number of model parameters and L_l is the maximal log-likelihood.

Based on the available training data the *single lambda success rewarding* model variant was selected, with $\lambda = 0.826$ and p_0 computed for each match using the n_2 normalization function defined in (33).

5 Model application for in-play betting

To test the quality of the model, following experiment was performed. For each match in the testing dataset, first the starting probability was derived using the procedure described in Section 4.1. Then for each set played, the set winning probability was computed using the *success rewarding* model and memory parameter $\lambda = 0.826$ as specified in Section 4.2. Finally, this probability was compared to odds provided by the bookmaker and a virtual bet was made.

Choosing the correct betting strategy is one of the key elements of successful betting. It depends on the underlying model, available bookmaker's odds, bankroll, internal bookmaker's policies and many other parameters. There exists a large number of possible approaches and the detailed description of them is beyond the scope of this paper. For testing purposes, there were three basic strategies tested. First, the naive betting strategy, where simply 1 unit was bet every time. Then, the probability based strategy, where p (probability of winning) units was bet. Finally the odds based strategy, where $\frac{1}{odds}$ units was bet. The strategies differ in the expected wins and their variance. The theoretical properties of the betting strategies can be observed in Table 1 with the special case where $p = \frac{1}{a}$, i.e. in case of *fair* odds [1]. Besides the different betting amount strategies, it is also important to choose when to bet. The basic strategy is to bet always when $p > \frac{1}{odds}$. It turns out that it is more favorable to bet only when there is a margin present, i.e. when $p > \frac{1}{odds} \cdot m$, where m is some margin parameter, $m \geq 1$.

For $m = 1.2$, i.e. 20% margin, there are 65 virtual bets made and a significant profit is achieved. One of the parameters to evaluate a quality of a betting strategy is the return of investment (ROI), i.e. $\frac{profit}{bankroll\,needed}$. The naive betting yields $ROI = 148\%$, probability based betting $ROI = 98\%$ and odds based betting $ROI = 111\%$. The development of the portfolio balance can be seen in Figure 1.

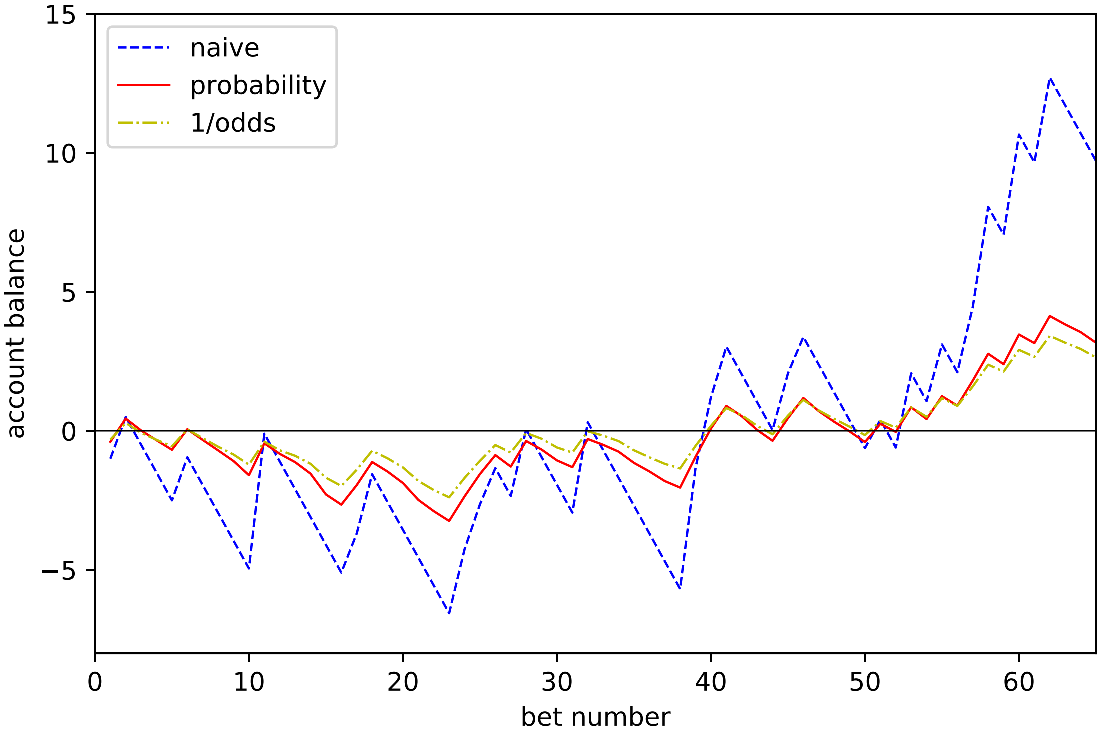


Figure 1: A graph showing the account balance development for different betting strategies.

6 Conclusion and future work

In this paper, a recently introduced model of a random walk with varying probabilities was used to predict the development of a tennis match. The quality of the predictions was tested against a real life odds provided by a bookmaker and the results confirm big potential of the model for tennis modeling. To better validate the model quality a larger dataset has to be acquired. The model can be further improved, for example by introducing a variable memory coefficient λ or by combining the model with some of the more classical, regression based approaches. This will be subject of a further research.

7 Remarks

The source code containing all functionality mentioned in this article is freely available as open source at GitHub[1]. More results can be also obtained from the same repository. The data used in this paper are available from the author upon request.

References

[1] Tomáš Kouřim. Mathematical models of tennis matches applied on real life odds. In *Proceedings of Doktorandské dny FJFI*, pages 83–91. Czech Technical University in Prague, 2015. Available at https://km.fjfi.cvut.cz/ddny/historie/historie/15-sbornik.pdf.

[2] Tomáš Kouřim. Random walks with varying transition probabilities. In *Proceedings of Doktorandské dny FJFI*, pages 141–149. Czech Technical University in Prague, 2017. Available at https://km.fjfi.cvut.cz/ddny/historie/historie/17-sbornik.pdf.

[3] Tomáš Kouřim and Petr Volf. Discrete random processes with memory: Models and applications. *Applications of Mathematics*, 65:271–286, 2020.

[4] Karl Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[5] Loïc Turban. On a random walk with memory and its relation with markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 43(28):285006, 2010. MR2658904.

[1] https://github.com/tomaskourim/mathsport2021

Funding: This work was financially supported by the PhD grant no. 2020.05672.BD, funded by national funds through FCT - Portuguese Foundation for Science and Technology. This work was financially supported by the research unit on Governance, Competitiveness and Public Policy (UIDB/04058/2020), funded by national funds through FCT. This paper is financed by national funds of the FCT within the project "UIDB/03182/2020".

A bibliometric study about the European Super League of football – through the scientific paths of a utopia?

Anthony Macedo* and Marta Ferreira Dias** and Paulo Reis Mourã

*GOVCOPP, DEGEIT, University of Aveiro, Aveiro, Portugal + email address: anthony
**GOVCOPP, DEGEIT, University of Aveiro, Aveiro, Portugal + email address: mfdias@ua.pt ** NIPE, EEG, University of Minho, E

1 Introduction

The European Super League (ESL) of football is the idea of grouping the top European clubs in a breakaway league. Over the years, this idea has repeatedly threatened the established structure regulated by UEFA, motivating several references in scientific literature. From the perspective of sports economics, this study aims to discuss the existing scientific knowledge on the subject and to create a well-structured groundwork to explore the topic of the ESL.

The literature about the ESL is particularly limited and most studies only refer the ESL in a few sentences or dedicate a small part of the study. Although a few studies with higher focus reached important findings, there is frequently a missing link between them. When these gaps had been observed in previous scientific fields, researchers have found a comfortable reply by developing bibliometric studies able to identify the well-developed areas, the networks of research and of citations on the field and the followed directions by the identified works. Therefore, we intend to pursue such goals with a bibliometric analysis.

2 Methodological procedure

In order to consider more studies in this review, we followed Boanares and de Azevedo (2014) and considered the Google Scholar database, as it does not limit the text search only to title, abstract, and keywords as do Web of Science (WoS) and Scopus, two of the most prominent databases for bibliometric analysis (Mourao & Martinho, 2020). This limitation would be a drawback for a bibliometric review seeking to be exhaustive on a very specific topic, such as the ESL. A large proportion of the studies that address the topic do not focus on it primarily, but more commonly debate it in other sections (Franck, 2018), sometimes suggesting that the ESL may become a consequence of their findings (Késenne, 2007) or only referring the historical importance of the idea of ESL for the main focus of the study (Geeraert & Drieskens, 2015). Consequently, its reference is often omitted from the sections title, abstract, and keywords.

Based on previous literature, a search string was derived[1], and the search done on 3rd December 2020 resulted in a total of 627 documents. After removing duplicates and a first screening round aiming to eliminate

[1] ("football" OR "soccer") AND ("european super league" OR "european superleague" OR "european major league" OR "pan-european league")

studies clearly not about the ESL of football[2], studies published in magazines or Bachelor/Master theses, and studies missing/not available to the authors, 190 studies remained.[3]

The main drawback of using the Google Scholar database is that the search results include articles from journals ranked with usually low scores if considering several indicators of impact factor and the so-called grey literature. Therefore the eligibility criterium of the studies for the bibliometric analysis is the presence of the study in Scopus or WoS databases. Only 81 studies meet this criterium (64 articles, 5 books, and 12 book chapters). Using the text data and bibliographic data of the documents collected and the software VOSviewer, several bibliometric indicators are created and presented in the next section.

3 Bibliometric analysis

3.1 The sources

Of a total of 55 sources in the sample, there are 40 academic journals and 15 books. The Journal of Economic Perspectives (JEP) and the Journal of Sports Economics (JSE) are clearly the most cited ones with over 220 citations, although JEP achieves this through only one study (Kahn, 2000). In terms of concentration, Soccer and Society (S&S) ranked first with 8 publications related to the ESL, which is 2 more than the JSE. It should be noted that 13 journals are specialised in sports and they represent 53% of the articles in analysis.

The most cited book is "Sport beyond television: The internet, digital media and the rise of networked media sport" (Hutchins & Rowe, 2012) with 152 citations, followed by "The economics of sports broadcasting" (Gratton & Solberg, 2007), "Globalization and football" (Giulianotti & Robertson, 2009), and "The sport business future" (Smith & Westerbeek, 2004).

Unsurprisingly, a tendency is observed for studies published earlier having more citations, which may, for example, explain why S&S does not have more citations and why some journals present zero citations at the moment.

If, in the other way around, we analyse the sources cited by the 81 studies considered in the analysis, the JSE is by far the most cited one (252 citations from 121 articles). Other commonly cited journals are Scottish Journal of Political Economy (91 from 34), Applied Economics (58 from 36), Journal of Political Economy (55 from 21), and S&S (51 from 40).

3.2 The authors

Stefan Szymanski was the author with the most studies discussing the ESL (7) and also the author with the most citations (270). In terms of number of contributions he is followed by Nicolas Scelles, Girish Ramchandani, Daniel Plumley, and Rob Wilson with 4 articles published recently (on average, they have not been published more than 3 years ago), so the number of citations is expected to increase substantially. Note that these latter three are co-authors. Regarding the number of citations, Anthony King and Wladimir Andreff follow Stefan Szymanski with, respectively, 118 and 102 citations.

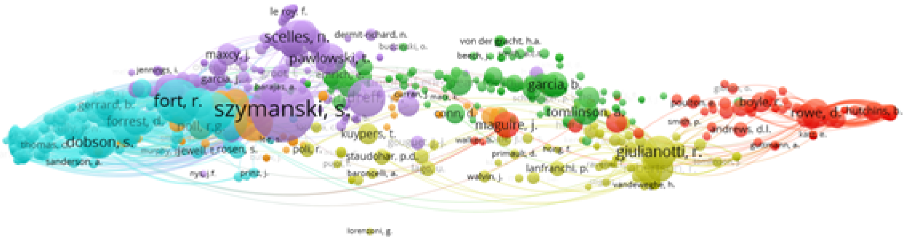
The co-citation map in Figure 1 shows the relatedness of authors. This map was created with VOSviewer, which uses the VOS mapping technique (Van Eck & Waltman, 2007) to draw the network of authors cited

[2] It is very often confused with rugby Super League.

[3] With WoS and Scopus only 8 relevant results would have been collected.

by the sample of studies in analysis (van Eck & Waltman, 2010).[4] The network is constituted of a set of authors and the connections between them. In the map, each node is an author, and each link is a connection between two authors based on the number of times they are cited together. When this number increases, the link between the two authors become stronger. So, the total number of links of an author differs from the total strength of his links.

Figure 1: Co-citation map of authors with at least two citations



The visualization of Figure 1 allows to group the authors in six clusters. The blue cluster is the biggest one, with 229 authors and about 5.75 citations per author, containing several studies about optimal contest design and sports demand. In terms of thematic, it is difficult to distinguish the blue cluster from the purple one, however, this latter gives more focus to European football and the former to professional sports in general. The purple cluster is considerably smaller (127 authors) but has the highest number of citations per author of Figure 1 (8.07 citations), largely driven by the 169 citations of Stefan Szymanski.

The orange cluster is small (43 authors) but has a high number of citations per author (6.02). It aggregates some of the pioneering names in sports economics (Andreff & Szymanski, 2006), such as Simon Rottenberg, Peter Sloane, James Quirk, and Mohamed El-Hodiri.

The green cluster is rich in authors who studied how the evolution of football business influenced European integration and how football governing bodies relate to European institutions and EU's law. It is the second cluster with more authors (213), but it is the cluster with less citations per author (3.83).

There are the yellow and red clusters left. The yellow cluster include studies discussing the interrelation between sports and culture, pointing to the influence of sports in globalisation, and the red cluster adds the media to this interrelation, also emphasizing the power of some sport governing bodies. These clusters present similar size (145 authors in the red cluster and 132 in the yellow cluster) and similar number of citations per author (3.97 and 4.10, respectively).

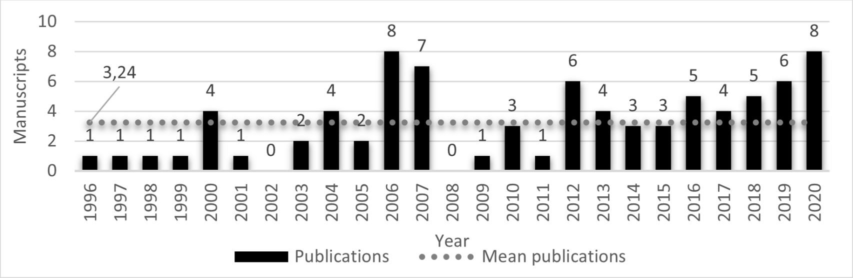
3.3 The studies

Figure 2 shows that, on average, from 1996 to 2020, 3.24 studies mentioning the ESL were published each year. The number of publications is influenced by institutional/environmental reasons that create stimulus for authors and interest for the journals. Some peaks can be observed in Figure 2 and the first one is in 2000, possibly due to the proposal of MPI to create an ESL and the creation of the G14. Another peak is observed in 2006-2007, which can be related to the renegotiation of the Champions League (CL) format in 2007, which some believed could have been a trigger for a ESL (Szymanski, 2006). In following years, different reasons could have motivated the emergence of a ESL. For example, changes in regulations from UEFA, in particular

[4] To turn the map understandable, only authors with at least two citations were considered.

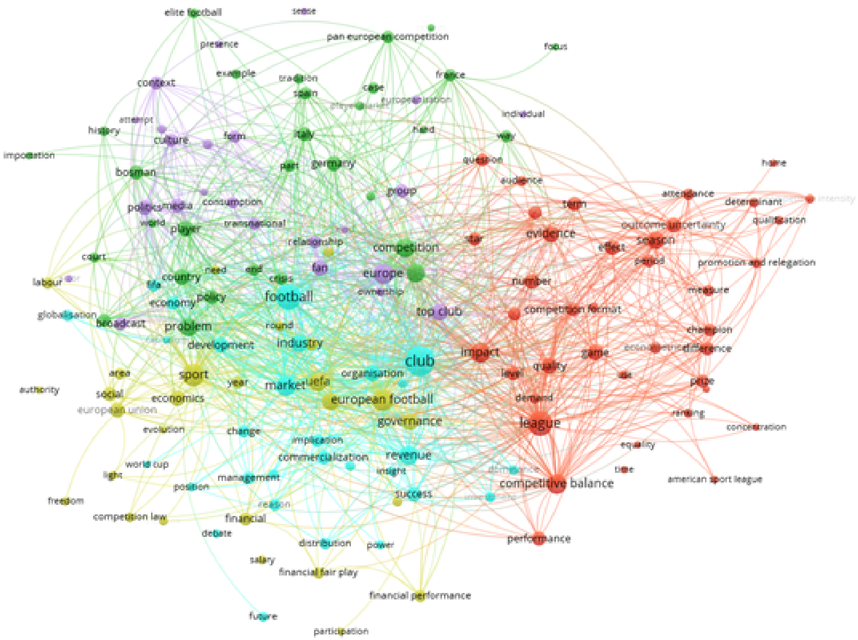
the financial fair play (FFP) regulations (Drut & Raballand, 2012), the pressure from media and sponsors (Maguire, 2011), the lack of competitive balance (CB) in current domestic leagues (Kringstad, 2020), and the desire from top clubs to increase their profits even more (Wills et al., 2020).

Figure 2: Evolution over time of publications related to the ESL



The network of Figure 3 was created in a similar way to Figure 1, but here each node is a term, and each link between two terms is based on the number of times they are used together in the sample of studies considered.[5]

Figure 3: Co-occurrence map of terms with at least three occurrences



The red cluster has 40 terms, being the biggest one. The terms in this cluster mainly concern studies estimating the determinant of demand for sporting events, being highlighted terms such as "league", "competitive

[5] Figure 3 considers the terms in title, abstract, and keywords sections. To make the map understandable, only terms with at least 3 occurrences were considered. Additionally, based on the analysis of the original studies, some terms were edited to create an informative map.

balance", "outcome uncertainty", "attendance", and "audience". These terms tend to be used in more recent publications and with fewer citations.

The green cluster has 30 terms, such as "elite football", "competition", the name of the big five football nations (England, Spain, France, Italy, and Germany), "policy", "Bosman", and "player market". The studies using these terms commonly refer to how the Bosman case influenced European football[6], but a notable aspect of these terms is their high number of links, which point to their versatility. An example of this is that the typical studies of the red cluster can be applied to any of the big five nations.

Next, the blue cluster has 29 terms, including the two with most occurrences in the sample of studies, which are "football" and "club". Considering the methodological approach followed in this study, it is not surprising that these terms are so common. Although these two terms are highly versatile, being the ones with more links, the blue cluster is more focused on the business side of football, highlighting terms such as "industry", "revenue", "economy", and "commercialisation". The yellow cluster presents some similarities with terms such as "economics" and "financial performance", but among its 27 terms there are also a few that approach institutional aspects, such as "governance", "regulation", and "UEFA".

Finally, the purple cluster is the one with fewer terms. "Europe" is the term with more occurrences and links in this cluster, while terms such as "fan", "consumption", "media", "culture", and "politics" show high relevance too. The terms in this cluster tend to be present in older studies and with more citations, however, it is interesting to observe that, on average, they have fewer links to other terms.

4 Conclusion and future research agenda

Although the shadow of European Super League has never entirely disappeared in recent decades, its discussion in scientific literature has generally been brief and infrequent (particularly in journals with a high impact factor). Among the existing studies, the discussion ranges from the ESL reducing consumer welfare and being incompatible with EU competition law (van der Burg, 2019) to arising as the most natural solution for a polarised European football (Hoehn & Szymanski, 1999). Pros and cons of different formats for the ESL are analysed, although without arriving at a definitive model. Given, on the one hand, this diversity of positions on the ESL and, on the other hand, the current stage of the debate constructed, we felt that a bibliometric study was necessary to mark the moment of scientific involvement.

One of the potential explanations for the lack of attention given to the ESL is its lack of tangibility. The ESL could adopt so much different features that several assumptions need to be made to focus on only a few formats. Besides, there is obviously no historical ESL data, so empirical works are very limited. This requires using data from other competitions and assuming that supply and demand would have a similar behaviour with the emergence of a ESL.

The low credibility of the ESL threats in the past may also be an explanation to the low interest of researchers and journals. There is no real certainty about club owners' objectives and some authors point (Franck, 2018; Millward, 2006a) are doubtful about the real will of top clubs to join a ESL, suggesting that the current system is the one that best accentuate polarization in football. The creation of a ESL would only be a threat to control UEFA.

[6] The Bosman ruling of the European Court of Justice liberalised labour market in European football by allowing free transfers of players at the end of their contract between clubs from the EU and banning the restrictions on the number of foreign EU players (Késenne, 2007).

In future research, it could be interesting to compare the evolution of studies on the ESL (Figure 2) with other topics, such as the competitive balance in domestic leagues and the revenues (in particular coming from broadcasting and UEFA) of top clubs. To identify an inverse relationship with the former would suggest that the discussion is aroused by concerns with fans, while to identify an inverse relationship with the latter would reinforce the idea that top clubs use the threat of ESL only to strengthen their position.

Finally, the recent ESL project in 2021, led by Florentino Pérez, may cause publications on the subject to increase in the near future and it will be stimulating to analyse whether it will have any effect on the industry.

References

Andreff, W., & Szymanski, S. (2006). Introduction: Sport and economics. In W. Andreff & S. Szymanski (Eds.), Handbook on the Economics of Sport (pp. 1-11). Edward Elgar Publishing

Boanares, D., & de Azevedo, C. S. (2014). The use of nucleation techniques to restore the environment: A bibliometric analysis. In Natureza e Conservacao (Vol. 12, Issue 2, pp. 93-98). Elsevier B.V.

Drut, B., & Raballand, G. (2012). Why does financial regulation matter for European professional football clubs? International Journal of Sport Management and Marketing, 11(1-2), 73-88.

Franck, E. (2018). European club football after "five treatments" with financial fair play - Time for an assessment. International Journal of Financial Studies, 6(4), 97.

Geeraert, A., & Drieskens, E. (2015). The EU controls FIFA and UEFA: A principal-agent perspective. Journal of European Public Policy, 22(10), 1448–1466.

Giulianotti, R., & Robertson, R. (2009). Globalization &amp; Football. In Globalization and Football. SAGE Publications.

Gratton, C., & Solberg, H. A. (2007). The economics of sports broadcasting. Routledge.

Hoehn, T., & Szymanski, S. (1999). The Americanization of European football. Economic Policy, 14, 203-240.

Hutchins, B., & Rowe, D. (2012). Sport beyond television: The internet, digital media and the rise of networked media sport. In Sport Beyond Television: The Internet, Digital Media and the Rise of Networked Media Sport. Taylor & Francis.

Kahn, L. M. (2000). The sports business as a labor market laboratory. Journal of Economic Perspectives, 14, 75-94.

Késenne, S. (2007). The peculiar international economics of professional football in Europe. Scottish Journal of Political Economy, 54, 388-399.

Kringstad, M. (2020). Comparing competitive balance between genders in team sports. European Sport Management Quarterly, 1-18.

Maguire, J. A. (2011). The global media sports complex: Key issues and concerns. Sport in Society, 14(7-8), 965–977.

Millward, P. (2006). Networks, power and revenue in contemporary football: An analysis of the G14. International Review of Modern Sociology, 32(2), 199–216.

Mourao, P. R., & Martinho, V. D. (2020). Forest entrepreneurship: A bibliometric analysis and a discussion about the co-authorship networks of an emerging scientific field. Journal of Cleaner Production, 256, 120413.

Neale, W. C. (1964). The peculiar economics of professional sports: A contribution to the theory of the firm in sporting competition and in market competition. The Quarterly Journal of Economics, 78.

Rottenberg, S. (1956). The baseball players' labor market. Journal of Political Economy, 64, 242-258.

Smith, A., & Westerbeek, H. (2004). The sport business future. In The Sport Business Future. Palgrave Macmillan.Szymanski, S. (2003). The economic design of sporting contests. Journal of Economic Literature, 41, 1137-1187.

Szymanski, S. (2003). The economic design of sporting contests. Journal of Economic Literature, 41(4), 1137–1187.

Szymanski, S. (2006). Football in England. In S. Szymanski & W. Andreff (Eds.), Handbook on the Economics of Sport (pp. 459-462). Edward Elgar Publishing.

van der Burg, T. (2019). EU Competition Law With Respect to Football Clubs and National Markets. SSRN Electronic Journal.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523-538.

Van Eck, N. J., & Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. Studies in Classification, Data Analysis, and Knowledge Organization, 299-306.

Wills, G., Tacon, R., & Addesa, F. (2020). Uncertainty of outcome, team quality or star players? What drives TV audience demand for UEFA Champions League football? European Sport Management Quarterly.

The Fifth International Timetabling Competition (ITC 2021): Sports Timetabling

D. Van Bulck*, D. Goossens**, J. Beliën***, and M. Davari****

*Ghent University, david.vanbulck@ugent.be
**Ghent University, dries.goossens@ugent.be
***KU Leuven, jeroen.belien@kuleuven.be
****SKEMA Business School, morteza.davari@skema.edu

Abstract

This paper discusses the organization of the most recent International Timetabling Competition (ITC 2021). This competition focused on sports timetabling, where the problem is to decide on a suitable date for each of the matches to be played in the tournament. This is a complex and challenging problem, even for tournaments with few contestants. As a consequence, state-of-the-art typically focuses on a particular season of a sports competition for which a tailored algorithm is developed which is then compared to a manual solution. The aim of this competition was therefore to promote and provide insights in the development of more generally applicable sports timetabling solvers. To this purpose, participants required to solve a rich and diverse set of 45 sports timetabling instances involving up to 9 different constraints that are common in real life. We discuss the format of these instances, how the instances were released during the competition, and conclude with an overview of the finalists.

1 Introduction

Creating timetables for sports competitions has been a topic of research since the 1970s (e.g., [1]). Ever since, academic papers about sports timetabling have increased considerably in numbers and sports timetabling has become a specialized field [10]. Sports timetabling is often complex and challenging, even for a small number of teams. While generating a timetable where each team plays against each other team once and no team is involved in simultaneous matches is easy (e.g. [6]), some rather basic sports timetabling problems are already NP-hard. For instance, [2] show that there is no constant-factor approximation (unless $P = NP$) for a sports timetabling problem where certain matches cannot be played on a set of predefined rounds. Furthermore, real-life sports timetabling problems are characterized by a wide diversity of constraints, and conflicting interests of many stakeholders. At the same time, in professional sports, the timetable has an impact on commercial interests and revenues of the clubs, broadcasters, sponsors, as well as an impact on society through resulting traffic and policing costs.

Since 2002, there have been frequent timetabling competitions, which have been beneficial for the research community. The first international timetabling competition was organized in 2002 and focused on (a simplified version of) the university course timetabling problem (see [16]). The next ITC competition (2007) aimed to further develop interest in the general area of educational timetabling and involved three problems: curriculum-based timetabling, examination timetabling, and post-enrolment timetabling (see [12, 13]). With high-school timetabling, the ITC placed yet another educational timetabling problem in the spotlights in 2011 (see [17, 18]). The fourth ITC is again devoted to university course timetabling: it introduces the combination of student sectioning together with time and room assignment of events in courses (see [14, 15]). In between, there have been two international nurse rostering competitions in 2010 (see [9]) and 2014 (see [4]), as well as a cross-domain heuristic search challenge (CHeSC 2011), where the challenge was to design a high-level search strategy that controls a set of problem-specific low-level heuristics, which would be applicable to different problem domains (see [3]).

Many of the sports timetabling contributions in the literature read as a case study, describing a single instance for which a tailored algorithm is developed (which is then typically compared to a manual solution). Moreover, the state-of-the-art does not offer a general solution method, or even much insight in which type of algorithm would work well for which type of problem (see [19]). One notable exception is the travelling tournament problem [7], an artificial and somewhat simplified sports timetabling problem where the objective is to minimize the total team travel in a timetable. For this problem, substantial algorithmic progress has been reported after [7] made a set of artificial benchmark instances publicly available, and for which best results can be submitted to a website maintained by professor Michael Trick (see `http://mat.tepper.cmu.edu/TOURN/`). A long standing obstacle to benchmark algorithms for sports timetabling problems that are real-world-like was the absence of a file format to express the wide amount and variety of constraints that are typically present in real-life problem instances. Given the recent efforts by [19] to overcome this obstacle, we believed the time was right to organize an international timetabling competition on sports.

The remainder of this paper is as follows. 2 provides a general description of the type of problems offered in the competition, and 3 outlines the competition rules. We conclude in 4 with a short discussion of the competition timeline, the prizes that could be earned, and the announcement of the finalists.

2 Problem description and file format

In essence, sports timetabling is deciding on a suitable round for each of the matches to be played in the tournament. In practice, rounds typically correspond to weekends, and consist of several time slots (e.g., Saturday evening, or Sunday afternoon), however, each team plays at most once per round. The competition focuses on the construction of round-robin timetables, meaning that each team plays against every other team a fixed number of times. Although many other tournament formats are conceivable (e.g. the knock-out tournament), round-robin tournaments are probably the most researched format (see [11]) and are very common in practice (see e.g. [8]). Most sports competitions organize a double round-robin tournament (2RR) where teams meet twice but single, triple, and even quadruple round-robin tournaments also occur. Existing literature distinguishes two types of round-robin tournaments: time-constrained timetables and time-relaxed timetables. A timetable is time-constrained (also called compact) if it uses the minimal number of rounds needed, and is time-relaxed otherwise. In this competition, we only consider time-constrained double round-robin tournaments with an even number of teams. Under this setting, the total number of rounds is exactly equal to the total number of games per team, and hence each team plays exactly one game per round. Although there is a line of research that focuses on the simultaneous scheduling of multiple leagues with dependencies [5], we focus on the construction of a 2RR for a single league. For an example of a time-constrained 2RR timetable, we refer to 1.

Table 1: A time-constrained double round-robin timetable for a single league with 6 teams. Each game is represented by an ordered pair in which the first element is the home team, and the second element is the away team.

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
(1,2)	(2,5)	(2,4)	(2,3)	(6,2)	(4,2)	(5,2)	(2,1)	(3,2)	(2,6)
(3,4)	(4,1)	(1,6)	(5,1)	(4,5)	(6,1)	(1,4)	(4,3)	(1,5)	(5,4)
(5,6)	(6,3)	(5,3)	(6,4)	(1,3)	(3,5)	(3,6)	(6,5)	(4,6)	(3,1)

The constraints that appear in real-life problem instances are extremely diverse: apart from some basic constraints, each competition has its own requirements. In this competition, we assume that there are two types of constraints: hard constraints that represent fundamental properties of the timetable that can never be violated, and soft constraints that represent preferences that should be satisfied whenever possible. While many possible optimization objectives appear in the literature (e.g. the minimization of travel), this competition considers problem instances only where the objective is to minimize the penalties from violated soft constraints. This assumption makes the problem formulation more attractive for a wider timetabling community, while retaining the empirical complexity of the problems. In total, nine types of constraints were considered that can be categorized into the following five constraint classes as introduced by [19]. Capacity constraints force a team to play home or away and regulate the total number of games played by a team or group of teams. Game constraints enforce or forbid specific assignments of a game to rounds. Constraints to increase the fairness or attractiveness involve balancedness of, e.g., home advantage, travel distances, etc. Break constraints regulate the frequency and timing of breaks in a competition; we say that a team has a break if it has two consecutive home games, or two consecutive away games. Finally, separation constraints regulate the number of rounds between consecutive games involving the same teams.

The problem instances are expressed using the standardized XML data format developed by [19]. The main intention of this data format is to promote problem instance data sharing and reuse among different users and software applications, and this is exactly what the timetabling competition envisioned. The XML data format is open, human readable (i.e., no binary format), software and platform independent, and flexible enough to store the problem instances. Most of the sports timetabling constraints are easy to express in words but are hard to enforce within specific algorithms such as mathematical programming or metaheuristics. We believe this format minimized the specification burden and maximized the accessibility. The main advantage of XML over plain text-only file formats lies in the structured way of data storage which separates data representation from data content.

3 Competition rules

Prior to the competition, all rules and a number of sample instances were made available at the competition website (`itc2021.ugent.be`). The website provides more details on the rules of the competition, the problem instances and their XML format, the awards for the winners, and intermediate results. The website also provides access to a validator, allowing participants to verify whether their solution satisfies all hard constraints and to determine its score on the objective function.

We are much indebted to the various organizers of the previous international timetabling competitions.

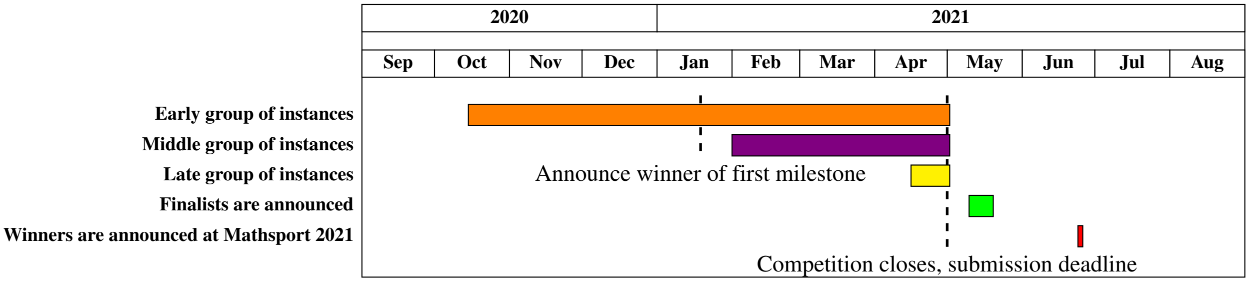


Figure 1: Timeline for the International Timetabling Competition 2021

Their experience has crystallized into the rules that were used for the ITC 2019 competition [15], and to which we largely adhered for this competition. In particular, we enforced no bound on the computation time. In fact, the objective function value of the best submitted solution was the only criterion that mattered. While computation time is obviously not unimportant, a fair comparison in terms of computation time is quite challenging, and it could easily lead to disputes that we as organizers prefer to avoid. Moreover, from a practical point of view, sports timetabling problems are often not so time-critical, as there are often several days or even weeks available to obtain a good solution.

We also allowed to make use of any commercial solver. In this way, we tried to lower the threshold to participate, and reach out to the largest possible research community. Obviously, to keep it interesting, the instances for the competition were designed such that a straightforward implementation on e.g., state-of-the-art integer or constraint programming solvers, could not solve the problem instances to optimality. In fact, for most problem instances, a straightforward integer programming formulation could not even generate a feasible solutions within a reasonable amount of time.

Although we allowed parameter tuning, we required that the same version of the algorithm was used for all instances. In other words, the algorithm should not 'know' which instance it is solving. While the algorithm may analyze the problem instance and set parameters accordingly, it should apply this same procedure for all instances. The programmer should not set different parameters for different instances, however, if the program is doing this automatically, then this is acceptable.

We believe these rules are efficient (in the sense that they do not require the organizer to run the participant's code) and fair/simple (in the sense that the only thing that matters is the obtained objective value; it avoids all discussion about measuring, e.g., computation time, the impact of random seeds, etc.).

4 Competition timeline and results

An overview of the competition timeline is given in 1. In total, we released three groups of 15 artificially generated problem instances each: early, middle, and late instances. While all instances contributed to the final ranking of participants, instances that were released later in the competition had a higher weight. For instance, the overall best found solutions was respectively awarded 10, 15, and 25 instances for an early, middle, and late problem instance. The early group of instances were already available from our website at the time the competition was officially announced (mid October 2020), while the middle group of instances were only released in February 2021. The late instances followed half April 2021, which gave the participants

Team name	Research institute	Participants
TU/e	Eindhoven University of Technology	F. Spieksma, H. Christopher, R. Lambers, and J. van Doornmalen
Saturn	HSE University	S. Daniil and R. Ivan
MODAL	Zuse Institute Berlin	T. Koch, T. Berthold, and Y. Shinano
GOAL	Federal University of Ouro Preto	G. H. G. Fonseca and T. A. M. Toffolo
UoS	University of Southampton	T. Martínez-Sykora, C. Potts, C. Lamas-Fernández
Udine	University of Udine	R. M. Rosati, M. Petris, L. Di Gaspero, and A. Schaerf

Table 2: Overview of the 6 finalists (ordered randomly)

two weeks to come up with solutions.

Around half January 2021, we organized a first milestone event where participants had the possibility to submit their best solutions found at that time. Although optional, participation in the first milestone was strongly encouraged as it provided participants with the feedback on where their algorithms ranked among their peers as well as a chance to win a small prize (free registration for Mathsport 2022). The first milestone was won by team UoS, followed by team Udine and TU/e (see 2).

At the time of the final submission deadline, 13 research teams from over 10 different countries successfully submitted solutions. As a comparison, the cross-domain heuristic search challenge attracted 17 teams, the two international nurse rostering competitions each attracted 15 teams, and the third and fourth international timetabling competition each attracted 5 teams that submitted one or more solutions by the final submission deadline. Out of all 13 participating teams, the 6 finalists given in 2 were selected. Note that the order of this list was generated at random and hence is unlikely to represent the final ordering. The prize fund is 1,750EUR to be split between the first, second, and third place competitors. Moreover, a discount on registration for the upcoming PATAT conference is awarded to the top three overall. The final ordering of the finalists together with an overview of the best found solutions will be announced at the Mathsport International 2021 conference.

Given the large number of teams that participated in the competition and the fact that feasible solutions were found for all problem instances, we conclude that the ITC 2021 competition was a huge success.

References

[1] B. C. Ball and D. B. Webster. Optimal scheduling for even-numbered team athletic conferences. *Aiie T.*, 9:161–169, 1977.

[2] D. Briskorn, A. Drexl, and F. C. R. Spieksma. Round robin tournaments and three index assignments. *4OR-Q J Oper Res*, 8:365–374, 2010.

[3] Edmund K. Burke, Michel Gendreau, Matthew Hyde, Graham Kendall, Barry McCollum, Gabriela Ochoa, Andrew J. Parkes, and Sanja Petrovic. The cross-domain heuristic search challenge – an international research competition. In Carlos A. C. Coello, editor, *Learning and Intelligent Optimization*, pages 631–634, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[4] S. Ceschia, N. Dang, P. De Causmaecker, S. Haspeslagh, and A. Schaerf. The second international nurse rostering competition. *Ann. Oper. Res.*, 274:171–186, 2019.

[5] M. Davari, D. Goossens, J. Beliën, R. Lambers, and F. Spieksma. The multi-league sports scheduling problem, or how to schedule thousands of matches. *Oper. Res. Lett.*, 48:180 – 187, 2020.

[6] D. de Werra. Geography, games and graphs. *Discrete Appl. Math.*, 2:327–337, 1980.

[7] K. Easton, G. Nemhauser, and M. Trick. The traveling tournament problem description and benchmarks. In T. Walsh, editor, *Principles and Practice of Constraint Programming — CP 2001*, pages 580–584, Berlin, Heidelberg, 2001. Springer.

[8] D. R. Goossens and F. C.R. Spieksma. Soccer schedules in Europe: an overview. *J. Sched.*, 15:641–651, 2011.

[9] S. Haspeslagh, P. De Causmaecker, A. Schaerf, and M. Stølevik. The first international nurse rostering competition 2010. *Ann. Oper. Res.*, 218:221–236, 2014.

[10] G. Kendall, S. Knust, C. C. Ribeiro, and S. Urrutia. Scheduling in sports: An annotated bibliography. *Comput. Oper. Res.*, 37:1–19, 2010.

[11] S. Knust. Classification of literature on sports scheduling, 2021. URL `http://www2.inf.uos.de/knust/sportssched/sportlit_class/`.

[12] B. McCollum, A. Schaerf, B. Paechter, P. McMullan, R. Lewis, A.J. Parkes, L. Di Gaspero, R. Qu, and E.K. Burke. The second international timetabling competition (ITC2007), 2007. URL `http://www.cs.qub.ac.uk/itc2007/index.htm`.

[13] B. McCollum, A. Schaerf, B. Paechter, P. McMullan, R. Lewis, A. J. Parkes, L. Di Gaspero, R. Qu, and E. K. Burke. Setting the research agenda in automated timetabling: The second international timetabling competition. *Informs J. Comput.*, 22:120–130, 2010.

[14] T. Müller, H. Rudová, and S. Müllerová. University course timetabling and international timetabling competition 2019. In E. K. Burke, L. Di Gaspero, B. McCollum, N. Musliu, and E. Özcan, editors, *Proceedings of the 12th International Conference on the Practice and Theory of Automated Timetabling*, Vienna, 2018. PATAT.

[15] T. Müller, H. Rudová, and S. Müllerová. International timetabling competition (ITC2019), 2019. URL `http://www.itc2019.org/home`.

[16] B. Paechter, L.M. Gambardella, and O. Rossi-Doria. International timetabling competition (ITC2002), 2003. URL `http://sferics.idsia.ch/Files/ttcomp2002/`.

[17] G. Post, L. Di Gaspero, J.H. Kingston, B. McCollum, and A. Schaerf. International timetabling competition (ITC2011), 2011. URL `http://www.utwente.nl/en/eemcs/dmmp/hstt/itc2011/`.

[18] G. Post, L. Di Gaspero, J. H. Kingston, B. McCollum, and A. Schaerf. The third international timetabling competition. *Ann. Oper. Res.*, 239:69–75, 2016.

[19] D. Van Bulck, D. Goossens, J. Schönberger, and M. Guajardo. RobinX: A three-field classification and unified data format for round-robin sports timetabling. *Eur. J. Oper. Res.*, 280:568 – 580, 2020.

Betting Market Inefficiencies in European Football

David Winkelmann* and Marius Ötting** and Christian Deutscher***

*Bielefeld University, Universitätsstrasse 25, Bielefeld, Germany + email address: david.winkelmann@uni-bielefeld.de
Bielefeld University, Universitätsstrasse 25, Bielefeld, Germany*Bielefeld University, Universitätsstrasse 25, Bielefeld, Germany

Abstract

Sports betting markets are continuously growing in popularity and economical relevance. With increasing competition among bookmakers, betting markets have to be excellent predictors of game outcomes. Efficiency of sports betting markets implies that betting odds reflect all available information while bettors cannot use simple systematic strategies to generate profits. Yet, previous research has identified various strategies that would have led to positive returns to bettors. While such research typically used relatively short observation periods, we analyse whether short-term inefficiencies last long-term. For data from the English Premier League our results show that inefficiencies appear and disappear over a total of 14 seasons, while we still uncover systematic strategies generating positive returns in the long-run.

Keywords: Betting Markets, Biases, Market Efficiency

1 Introduction

Market efficiency in sports betting implies that betting odds reflect all information publicly available, similar to general financial markets [4]. However, millions of bettors worldwide try to find strategies to beat the market as they presume to have bettor knowledge about the game outcome than the bookmaker. In the short-run, bettors could generate positive returns on investment (ROI) even in efficient markets due to randomness of game outcomes. In the long-run, statistical noise levels out and bettors cannot make profits in efficient markets [11].

Previous literature focuses on certain characteristics of a team or a match, such as the venue or the competitive balance of a match. Various studies report the occurrence of systematic betting strategies associated with positive returns for different leagues and periods. Still, these studies typically cover few seasons only and analysed strategies do not translate into bets on every game. Hence, our approach covers a relatively long period of time with 14 seasons of the English Premier League, the economically most relevant football league with a comprehensive betting volume [3]. Here, we pick up the profitable short-term strategies found in previous work and determine if they are also profitable long-term. We first present previous findings on betting markets, followed by our own long-term analysis.

2 Literature Review

Research on betting markets evaluates systematic betting strategies which enable bettors to net positive returns over the span of few seasons. Respective studies can be found in general economic outlets as well as in the fields of operational research and forecasting. Previous literature has uncovered multiple match and team characteristics which led to the possibility of generating profits for bettors as the result of inefficient betting markets.

While currently most matches are played behind closed doors due to the COVID-19 pandemic, in general, the venue of the match has impact on the winning probabilities of the two teams. Hence bookmakers offer, on average, lower odds for bets on the home team. [1], [7], and [12] provide evidence for the so-called *home bias*, where bookmakers underestimate the home advantage and enable bettors to generate positive ROIs when consistently betting on the home team. As the home advantage disappeared for matches which has been played behind closed doors during the COVID-19 pandemic, several studies analyse the (non-)reaction by bookmakers to this change in match characteristics (see, e.g., [6, 13]).

Furthermore, literature discusses the *sentiment bias*. This phenomenon is referred to biased betting odds for more popular teams. Promising strategies uncovered by [7] and [8] include consistently betting on teams with higher sentiment.

While named studies cover full seasons, single seasons can be split into different periods. As squads change during the off-season period, bookmakers have shown difficulties to asses team strength at the very beginning of seasons. [10] provide evidence of temporal inefficiencies for the first rounds as well as the end of seasons. For the German Bundesliga, [2] show that betting on recently promoted teams is also a promising strategy.

3 Data

While most of the studies mentioned above cover only a limited number of seasons (typically two or three) and hence rather short-term strategies, we consider 5,320 matches within a period of 14 seasons between 2005/06 and 2018/19 and analyse possible long-term inefficiencies of the betting market in the English Premier League. Our analysis relies on data from `www.football- data.co.uk`. In addition to match results this data set comprises pre-game betting odds for home and away wins as well as draws for each match. We use the average betting odds over all up to 42 bookmakers. As bets on wins are way more popular than bets on draws, our data set covers two rows (one for each team) for each match [7]. We find a considerably higher proportion of home wins (46.6%) than away wins (28.7%) over all seasons. To control if the bookmakers anticipate such home advantage in a correct way, we introduce the covariate *Home* to capture a potential home bias. This dummy variable takes value one if we bet on the home team and zero otherwise.

The stadium capacity in the Premier League varies between about 20,000 (for some seasons even lower) and about 75,000. Therefore, we consider the difference in the average attendance between both teams in the corresponding season *DiffAttend* as a proxy for the team's sentiment. In about 75% of the matches, the difference in the attendance does not exceed 25,000.

As previous studies revealed promising strategies related to betting on matches containing promoted teams (see [2]), we introduce binary variables *OnPromotedHome*, *OnPromotedAway*, *AgainstPromotedHome*, and *AgainstPromotedAway*. In our sample, 26.8% of the matches include one promoted team. For matches between two promoted teams we set these binary variables to value zero for both teams.

4 Analysis of the English Premier League

Under the assumption of a fully efficient market, the bookmakers' odds perfectly correlate with the expected winning probability for a given match. We use a logistic regression model to explain whether the binary variable *Won* equals zero or one using several covariates. Beyond the implied winning probability for a bet, derived from the bookmakers' odds, we control for the round (i.e. the current matchday) as well as the difference in the average attendance between both teams. We further include dummy variables for betting on the home team and on/against promoted teams, leading to the following model:

$$\begin{aligned}\eta_i = {} & \beta_0 + \beta_1 \mathit{ImpliedProbability}_i + \beta_2 \mathit{Home}_i + \beta_3 \mathit{DiffAttend}_i \\ & + \beta_4 \mathit{AgainstPromotedHome}_i + \beta_5 \mathit{AgainstPromotedAway}_i \\ & + \beta_6 \mathit{OnPromotedHome}_i + \beta_7 \mathit{OnPromotedAway}_i \\ & + \beta_8 \mathit{Round}_i .\end{aligned}$$

The binary response variable is linked to the linear predictor by the logit function. This approach has been considered in previous literature (see, e.g., [5, 7, 9]) and allows to test whether any covariate beyond the implied winning probability has explanatory power on the outcome. If there is any such covariate, the efficient market hypothesis would be challenged, and there is evidence for a potentially promising betting strategy.

Table 1: Estimation results for the regression model fitted to all seasons of the English Premier League

	Response variable:
	Won
ImpliedProbability	4.963^{***}
	(0.188)
Home	0.111^{*}
	(0.058)
DiffAttend	0.002
	(0.001)
AgaintPromotedHome	0.160^{*}
	(0.091)
AgainstPromotedAway	0.005
	(0.091)
OnPromotedHome	0.045
	(0.092)
OnPromotedAway	−0.022
	(0.110)
Round	0.002
	(0.002)
Constant	-2.544^{***}
	(0.081)
Observations	10,640
Note:	*p<0.1; **p<0.05; ***p<0.01

We fit our model to the whole data set, i.e. all season from 2005/06 until 2018/19. The results, obtained from Table 1, suggest that a comprehensive amount of variation in the probability to win a bet can be explained by the bookmakers' implied probability of that very outcome. However, our model provides evidence for a home bias as the variable *Home* has a statistically significant impact on the chances to win a bet. These chances are increased for bets on the home team compared to bets on the away team. Furthermore, betting against promoted teams in their home games increases the chance of winning the bet. As we neither find significant effects for *DiffAttend*, *Round*, and the remaining promoted team dummies, our model does not provide evidence for other systematic betting strategies in the long-run, considering the full data set.

However, results may differ in the short-run. Therefore, we conduct a season-by-season analysis in the following. For each season, Table 2 displays the ROI when consistently following a systematic strategy such as betting on the home team in each of the 380 matches during a single season. Results are given for seasons 2005/06 (first column) to season 2018/19 as well as over the whole period of 14 seasons (last column).

Table 2: Returns on presented strategies for all seasons

bet	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19	all
Home	**0.036**	**0.008**	-0.094	-0.062	**0.084**	**0.002**	-0.043	-0.124	-0.007	-0.035	-0.104	**0.054**	**0.010**	**0.031**	-0.017
DiffAttend	**0.012**	**0.117**	-0.011	**0.249**	**0.012**	-0.161	**0.080**	**0.143**	**0.017**	-0.11	**0.072**	-0.024	-0.098	**0.129**	**0.026**
AgPromHo.	-0.091	**0.045**	**0.280**	**0.019**	**0.093**	-0.132	**0.005**	**0.014**	**0.062**	**0.116**	-0.020	**0.142**	-0.055	**0.131**	**0.044**
AgPromAw.	**0.193**	-0.357	-0.101	-0.125	-0.326	-0.099	-0.318	-0.205	-0.112	-0.088	**0.223**	-0.020	-0.142	**0.204**	-0.091
OnPromHo.	-0.289	-0.040	-0.293	**0.108**	**0.015**	-0.174	**0.255**	-0.126	**0.151**	-0.188	-0.362	**0.314**	**0.196**	-0.075	-0.036
OnPromAw.	**0.108**	-0.458	-0.674	-0.307	-0.239	**0.557**	-0.036	-0.459	-0.304	-0.350	**0.099**	-0.749	-0.279	**0.053**	-0.217

Even if the regression model provides evidence for a significant home bias, the average return over the full period is negative. This is mainly driven by the bookmakers' margin. The average ROI of about -1.7% over all seasons shows that the betting strategy cannot make up for the margin taken by the bookmaker.

Considering the difference in the average attendance, we bet on a team when *Diff- Attend* exceeds the 95%-quantil of the corresponding season. The regression model provides a positive but insignificant effect. However, betting only on these extreme cases, we are able to gain returns of up to 25% in a single season. While we can realise positive returns in nine out of 14 seasons, the overall return is positive but only at 2.6%.

When betting on/against promoted teams, returns vary considerably over seasons. This is driven by the small number of matches containing promoted teams in a single season. However, a return of 4.4% can be generated when consistently betting against promoted teams in their home matches over the 14 season contained in the data set. Considering single seasons, a positive return can be generated in ten out of these 14 seasons. This confirms the positive significant effect of *AgainstPromotedHome* in the regression model.

5 Discussion

We analyse 14 seasons of the English Premier League to test whether betting markets are efficient in the long-run. For single seasons there are different strategies to obtain positive returns while the same strategies lead to comprehensive losses in a following season. However, our results show that betting on popular teams against unpopular teams and betting against promoted teams in their home games enabled bettors to generate positive returns in the long-run. Therefore, our analysis provides evidence that betting markets are not fully efficient even in the long-run.

In the light of a comprehensive market volume, these findings are somewhat surprisingly. However, further research is necessary to detect whether these systematic strategies can be also found in other leagues. Furthermore, in the long-run, single seasons with positive returns can occur by chance. Therefore, it should be analysed whether the findings in this study exceed what would be expected by chance.

References

[1] Angelini, G. and De Angelis, L. (2017). PARX model for football match predictions. *Journal of Forecasting*, 36(7): 795–807.

[2] Deutscher, C., Frick, B., and Ötting, M. (2018). Betting market inefficiencies are short-lived in German professional football. *Applied Economics*, 50(30): 3240–3246.

[3] Deutscher, C., Ötting, M., Schneemann, S., and Scholten, H. (2019). The demand for English Premier League soccer betting. *Journal of Sports Economics*, 20(4):556–579.

[4] Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25(2): 383–417.

[5] Feddersen, A., Humphreys, B. R., and Soebbing, B. P. (2017). Sentiment bias and asset prices: evidence from sports betting markets and social media. *Economic Inquiry*, 55(2): 1119–1129.

[6] Fischer, K. and Haucap, J. (2020). Betting market efficiency in the presence of unfamiliar shocks: The case of ghost games during the COVID-19 pandemic. *CESifo Working Paper*.

[7] Forrest, D. and Simmons, R. (2008). Sentiment in the betting market on Spanish football. *Applied Economics*, 40(1): 119–126.

[8] Franck, E., Verbeek, E., and Nüesch, S. (2011). Sentimental preferences and the organizational regime of betting markets. *Southern Economic Journal*, 78(2): 502–518.

[9] Franck, E., Verbeek, E., and Nüesch, S. (2013). Inter-market arbitrage in betting. *Economica*, 80(318):300–325.

[10] Goddard, J. and Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1): 51–66.

[11] Thaler, R. H. and Ziemba, W. T. (1988). Anomalies: parimutuel betting markets: racetracks and lotteries. *Journal of Economic Perspectives*, 2(2): 161–174.

[12] Vlastakis, N., Dotsis, G., and Markellos, R. N. (2009). How efficient is the European football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28(5): 426–444.

[13] Winkelmann, D., Deutscher, C., and Ötting, M. (2021). Bookmakers' mispricing of the disappeared home advantage in the German Bundesliga after the COVID-19 break. *Applied Economics*, 0(0):1–11.

Ranking rankings as predictor the final ranking

Roel Lambers* and Frits C.R. Spieksma**

*Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Nederlands + email add
** Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, the Nederlands + email addr

Abstract

As the covid pandemic spread over Europe, almost all of the top football competitions were suspended. Some temporary, some were eventually canceled. In the latter cases, for the subsequential season international tickets such as spots for the Champions League and Europa League, were given based on the current standings in the respective competitions. This led to controversy, as some teams missing out on these tickets had played less matches than others, and had met different, presumably stronger, opponents thus far. We examine a different method to determine a ranking, called True Ranking, and compare the predictive value of this alternative with that of the regular ranking throughout the season.

1 Introduction

During the 2019/2020 season, many professional sports leagues have been halted before the competition - usually a Double Round Robin - was completed. Some of these eventually continued later in the year, others were stopped indefinitely. As examples for the first case, we name mens football competitions such as the English Premier League and the Italian Seria A, for the latter case we name the French Ligue Un and the Dutch Eredivisie [BBC(June 2020)].

A relevant question whenever a competition is stopped, is how to determine the *final ranking* out of the matches played so far. The naive idea to look at the ranking based on the number of points collected so far, appears to have several shortcomings, such as:

- Teams have played a different number of matches.
- Teams played against different sets of opponents, influencing their results thus far.

Neither of these issues can be overcome while scheduling the league, as an abrupt ending of the competition can happen anytime - hence *abrupt*. To highlight the relevance of both points, the Eredivisie was stopped after Round 26, while 2 matches still had to be played - table leader *Ajax* versus sixth ranked *FC Utrecht*, and no. 2 versus no.3, *AZ-Feyenoord*. The standings in the Top-6 are shown in Table 1 [Wikipedia()].

As we can see, both Ajax and AZ have an equal amount of points, with Ajax leading on goal difference. The champion usually gets the automatic spot for the Champions League next season, where the runner-up only gets a ticket for a qualification round of the Champions League. Ranking fifth instead of sixth also makes a big difference, as the fifth spot guarantees participation in next years Europa League, where the

Standing	Team	Games Played	Points	Goal Difference
1	Ajax	25	56	45
2	AZ	25	56	37
3	Feyenoord	25	50	15
4	PSV	26	49	26
5	Willem II	26	44	3
6	FC Utrecht	25	41	16

Table 1: Top-6 when the Eredivisie was stopped

sixth ranked team gets nothing. *Willem II* is ahead of *FC Utrecht* by 3 points, exactly the amount the latter could gain by winning their match against *Ajax*, which would also lead them past *Willem II* based on their superior goal difference. The decision to take the rankings in this state as leading for the distribution of the european tickets was thus very controversial, with law suits from *AZ* and *FC Utrecht* directed at the football association KNVB. It is important to realise that the impact of the final ranking is not just a matter of honour and opportunity, but also of money - both *AZ* and *FC Utrecht* missed out on revenues that would come from playing (higher) in Europa. [RTL Nieuws(April 2020)]

Amidst the controversies that arose around several competition, such as the aforementioned Eredivisie and also the Ligue Un in France, alternatives were proposed to get to a fairer determination of the final rankings based on incomplete competitions. Guyon [Guyon(March 2020)] came with an ELO-like approach while sports analysts from Hypercube [Hypercube(March 2020)] simulated millions of matches to get a stochastic ending of the Eredivisie, the authors came with the *direct ranking* or *true ranking* [Spieksma and Lambers(April 2020)], an eigenvector method described before by Keener [Keener(1993)] and dating back to the beginning of the 20th century, while Csato [Csato(2020)] came with a procedure based on a generalized sum. In the aftermath of the first covid wave, sports associations implemented specific regulations for the case of canceled competitions such as the Dutch Fieldhockey Association KNHB, who based the final standings in the season 2020/2021 on points per game rather than total amount of points when the season was canceled after only a few rounds [KNHB(August 2020)].

Almost none of the involved parties lobbied for using the standings after round k as final rankings in the case of a cancellation, an as mentioned, some associations even specifically implemented alternatives for these situations. However, when looking at alternatives, almost none used statistical arguments to propagate any alternative compared to the regular standings. In this work we perform a small statistical analysis comparing the predictive value of True Ranking with that of the regular standings after every round, using data of the highest male football leagues in Spain, Italy and the UK of the past 10 years.

2 True Ranking

The true ranking or direct ranking as it is called in Keener, assumes that all N teams in the competition have a fixed strength and that their performance is scaled with that strength. Their performance is evaluated in such a way that a win or draw is valued based on the strength of the opponent.

Let N be the number of teams and let r_i be the strength of team $i \in [N]$, with $a_{i,j}$ the number of regular points scored by team i versus team j - 3 for a win, 1 for a draw, and let n_i be the number of matches played by team i. The number of points s_i scored by a team is then given by:

$$s_i = \frac{1}{n_i} \sum_{i \neq j} a_{i,j} r_j \tag{34}$$

Of course, a priori it is not clear what the strength of r_j is. In fact, the strength of r_j is what we set out to determine in the first place. However, when we assume that the strength is proportional to the score of a team, we expect there is a $\lambda \in \mathbb{R}$ such that:

$$s = Ar = \lambda r_i \tag{35}$$

Where $A = (\frac{a_{i,j}}{n_i})_{i,j \in [N]}$.

And indeed,

Theorem 2.1 (Perron-Frobenius)**.** *Under some minor constraints, there is a unique r such that $\lambda_{\max} \in \mathbb{R}$ and $Ar = \lambda r$.*

The constraints for the existance of r in practice means that there should be no set of teams $U \subset [N]$ such that the matches $(u, \overline{u})$ with $u \in U$ and $\overline{u} \in \overline{U}$, were either all won or all lost. There should be a certain entanglement in schedule and in the results, thus in practice the True Ranking is applicable after a few rounds.

3 Statistical analysis

To see how the True Ranking holds up to the regular standing in terms of predictive value, we've compared its performance for seasons starting in 2010 to 2019 for three of the biggest mens football competitions in Europe, the Premier League (England), Primera Division (Spain) and Serie A (Italy) [Football Data()]. To compare two rankings R^*, R_k - the final standings R^* with the prediction R_k after round k - we use *Kendall's tau* [Kendall(1938)].

Definition 3.1 (Kendall's tau)**.** *For teams t, t' we have $r_k(\cdot)$ their ranking after round k and $r^*(\cdot)$ as final ranking. We define sets:*

$$C = \{(t,t') : sign(r_k(t) - r_k(t')) = sign(r^*(t) - r^*k(t')), t, t' \in [N]\}$$
$$D = \{(t,t') : sign(r_k(t) - r_k(t')) \neq sign(r^*(t) - r^*k(t')), t, t' \in [N]\}$$

With these sets, where C represents all couples of which the partial order is similar, and D where the order is opposite, we can define τ to be the fraction of similar ordered couples:

$$\tau(R_k, R^*) = \frac{C - D}{C + D} \tag{36}$$

As the total number of ordered couples equals $\frac{N(N-1)}{2}$ and a couple is either ordered equal or opposite, we can rewrite τ to $\tau(R_k, R^*) = 1 - \frac{4D}{N(N-1)}$.

When two rankings are equal, all teams are ordered alike and τ will be 1. If two rankings are complete opposites, their τ will be -1. Intuitively, when a ranking is made after k rounds, τ measures how many times a team t can find another team t' that is ranked higher at that time while eventually finishing lower.

As we want to compare the predictive value of the True Ranking compared to the regular standings after k rounds, we are interested which of the two has a better average τ per round. All three analysed competitions have $N = 20$ teams, play a double round robin, and after half of the games (19) most teams have played all other teams once - bar a few exceptions. For the True Ranking to be sure to exist, we only look at the standings after round 8 and on wards.

The results of the τ-scores are shown in Figure 1 and 2. We see that throughout the season, both the regular current ranking as well as the True Ranking consistently improve on their predictive power, something one might expect. However, the regular standings on average consistently outperform the True Ranking as a predictor for the final rankings, albeit the difference in their τ-scre is very close to 0. What is noticable is that in the early days of the competition, when not that many games have been played, the True Ranking scores poorly - this might be explained as upsets - i.e., a team that win a lot of matches losing to a team that loses most of its matches - early in the season are given too much value in the eigenvector, while in the normal standings, 3 points are only 3 points, no matter who your opponent is.

Whenever the season is halfway, the predictive power of the True Ranking is less susceptible to these effects, but is still less than just applying the standings as they are. In the final rounds of the season, there is again a growing discrepancy, as the rankings will ultimately converge to the final ranking after 38 rounds, while for the True Ranking this need not be the case.

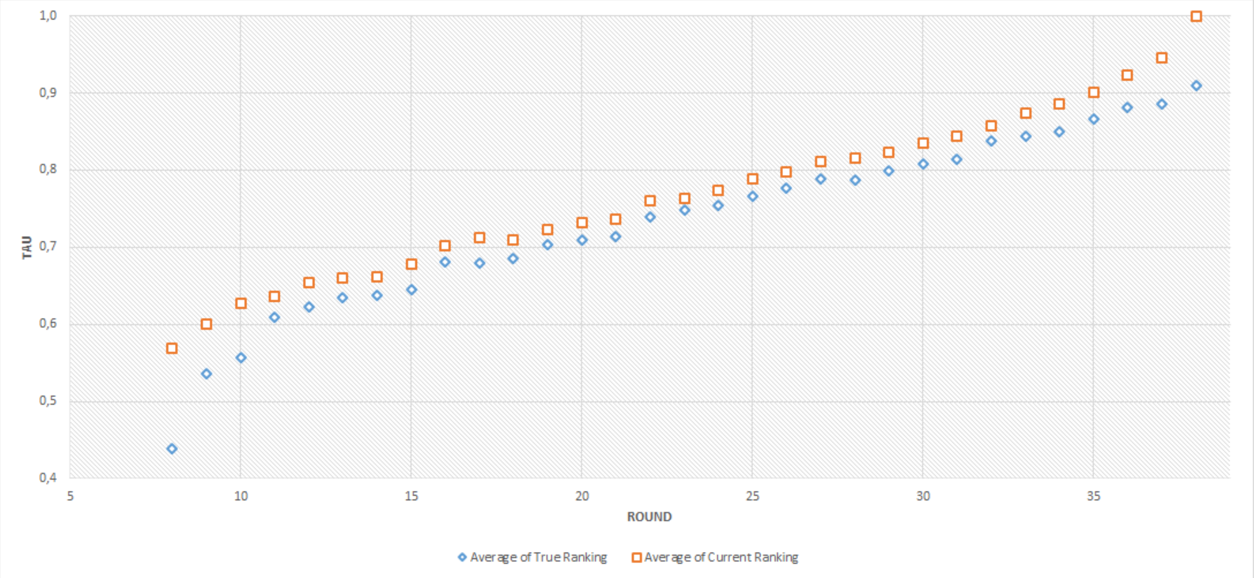


Figure 1: Average τ per round per ranking type

4 Discussion

As shown in Figures 1 and 2 and by the data behind these figures, the True Ranking is not a better predictor for the final outcome of a competition than just the regular ranking after k rounds. Even more, it seems to perform slightly worse throughout the season. A somewhat surprising result, as commentators usually refer

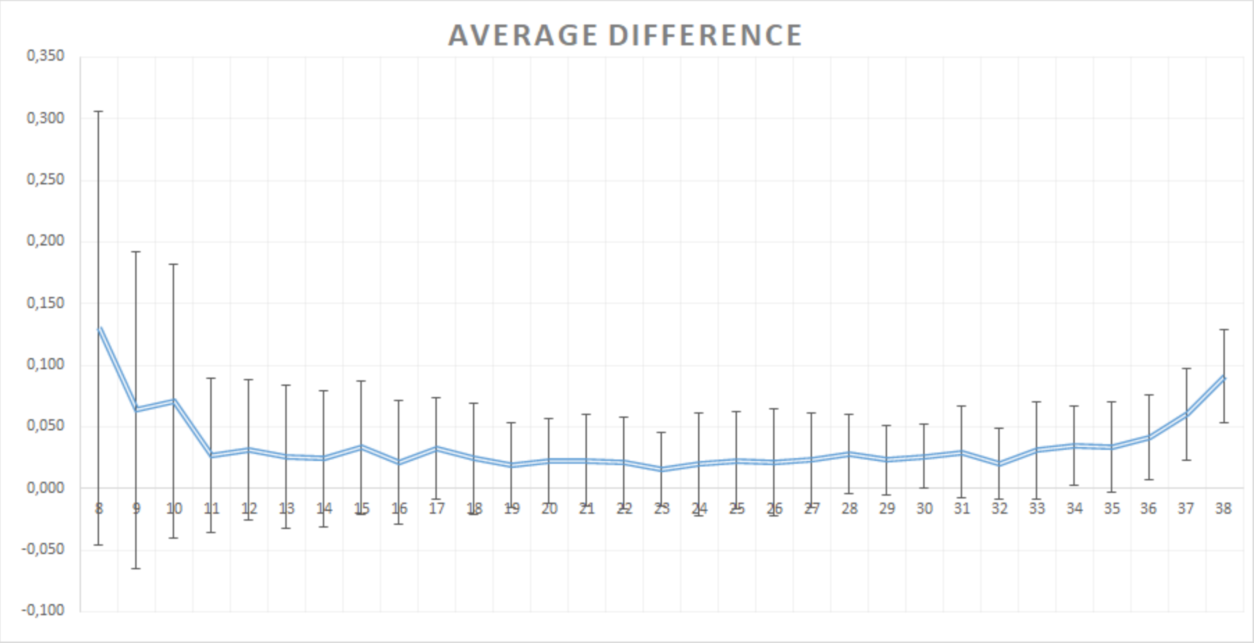


Figure 2: Average difference in τ with standard deviation

to teams having had hard/easy matches in periods of the competition - a correction for these difference in strength of competition faced so far does not perform better as predictor than ignoring these differences all together. However, there are some perks that might influence the final ranking in a way negative to the True Ranking.

When the season is ending, some teams might have achieved there goals or have nothing left to play for. For instance, if they secured a safe spot and are not in the running for European football, or have already got crowned champions. These teams might not give their all in a match, while other teams who are still fighting against relegation or for European tickets, can be extra eager to score against these opponents.

It is also unclear from these figures where the difference in predictive power lies exactly. Correctly ranking the teams placed around the 10-th position at the end, is not of much interest to the league organizers, as the most crucial spots are usually at the top and bottom of the competition.

As a third point, it would be interesting to check if there are elements where the True Ranking outperforms the regular standing and if it is possible to use this to create a hybrid method that is superior to both individual ways of ranking teams during the season.

References

[BBC(June 2020)] BBC. Coronavirus: How the virus has impacted sporting events around the world, June 2020. URL `https://www.bbc.com/sport/51605235`.

[Csato(2020)] L. Csato. Coronavirus and sports leagues: obtaining a fair ranking when the season cannot resume. 2020. URL `https://arxiv.org/abs/2005.02280`.

[Football Data()] Football Data. URL `football-data.org`. Last visited May 2021.

[Guyon(March 2020)] J. Guyon. The model to determine premier league standings, March 2020. URL `https://www.thetimes.co.uk/article/the-model-to-determine-premier-league-standings-ttt8tnldd`.

[Hypercube(March 2020)] Hypercube. Bureau hypercube: 'computermodel voor uitspelen eredivisie', March 2020. URL `https://www.vi.nl/nieuws/bureau-hypercube-computermodel-voor-uitspelen-eredivisie`.

[Keener(1993)] J. Keener. The perron-frobenius theorem and the ranking of football teams. *SIAM Review*, 35:80–93, 1993.

[Kendall(1938)] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[KNHB(August 2020)] KNHB. Scenario's in case of suspension due to corona, August 2020. URL `https://hockey.nl/nieuws/de-sport/coronascenarios-bij-afbreken-competitie-eindstand-na-helft-alle-duels/`.

[RTL Nieuws(April 2020)] RTL Nieuws. Knvb wijst geen kampioen aan: Ajax naar champions league, geen promotie/degradatie, April 2020. URL `https://www.rtlnieuws.nl/sport/voetbal/artikel/5102066/eredivisie-eindstand-knvb-ajax-rkc-ado-coronavirus`.

[Spieksma and Lambers(April 2020)] F. Spieksma and R. Lambers. True rankings, April 2020. URL `https://www.win.tue.nl/~fspieksma/papers/TrueRanking.pdf`.

[Wikipedia()] Wikipedia. Eredivisie 2019/20 (mannenvoetbal). URL `https://nl.wikipedia.org/wiki/Eredivisie_2019/20_(mannenvoetbal)`. Last visited May 2021.

Perspectives on the Impact of Covid-19 on Football (Soccer) and the 1x2 Betting Markets

J. Pym*, N. Patel*, T. Balamuralee* and A. Owen**

* BSc (Hons) Mathematics and Statistics, Coventry University, UK.
** sigma Mathematics and Statistics Support Centre, Coventry University, UK. email address: aa5845@coventry.ac.uk

Abstract

The impact Covid has had on football in terms of goals, shots and disciplinary measures is assessed along with how this has affected the home, draw and away win betting markets. The work is the result of original research undertaken by Final Year BSc (Hons) Mathematics and Statistics students at Coventry University.

1 Introduction

Since the COVID-19 pandemic first struck in early 2020, football (soccer) has been played in largely empty stadia with no fans present, and with other restrictions and changes imposed on the way that matches are managed. Previous evidence has suggested that the presence of fans has been one of the main reasons for the so-called home advantage in football, largely due to pressure being applied to referees by the home supporters, [1], [2] and [3]. Since football has returned to empty (or near empty) stadia, interest has naturally returned to this question of the extent to which fans contribute to the home effect, [4] and [5]. Here we not only explore how the absence of fans has affected football more generally, including home advantage, but also how the betting markets have struggled to deal with the correct pricing (i.e. the setting of fair odds) for match outcomes. Section 2 looks at the performance of the betting markets pre-Covid, whilst the impact of Covid is then considered in Section 3. Following this, Section 4 examines the impact Covid has had on football more generally in order to explain the impact on the betting markets .The work presented is the result of original research undertaken by Final Year BSc (Hons) Mathematics and Statistics students at Coventry University.

Data on 54,782 matches from 11 of the top European leagues from the start of the 2005/06 season to 12th May 2021 were gathered from www.football-data.co.uk, of which 4,102 have taken place in empty (or near empty) stadia since the resumption of top-flight football following the Covid pandemic. The data includes goals, shots, fouls, yellow cards and red cards for each team in each match, as well as the average betting odds recorded for the home win, draw and away win outcomes. The leagues included are the top league within Belgium, England, Germany, Greece, France, Italy, Netherlands, Portugal, Scotland, Spain and Turkey.

2 Football Betting Market Efficiency pre-Covid

Attention is focused here on the most popular betting market in football, the 1x2 market, where bets are placed on whether a match will end in a home win, a draw or an away win. The market is typically made up of odds, which are often referred to as prices, being offered by bookmakers who we refer to as the market, but the popularity of peer to peer betting exchanges (such as Betfair) now means that these market makers also include individuals offering odds (laying bets) to those willing to accept the bet. To assess the accuracy or efficiency of the 1x2 market, we consider the implied market probabilities derived from the reciprocal of the average odds recorded across a range of bookmakers (included in the data). For example, decimal odds of 4.0 commonly used in Europe (equivalent to 3/1 in fractional odds more common in the UK), equates to an implied probability of 1/4 or 0.25. Typical odds for a match might for example be 1.8, 3.5 and 5.0, which imply probabilities of 1/1.8=0.56, 1/3.5=0.29 and 1/5=0.20, for the home win, draw and away win respectively. Note that these market implied probabilities sum to 1.05. This additional amount by which the total of the (implied) probabilities exceeds 1 is referred to as the over-round, which in this example would provide the market makers with a profit of around 5%. That is assuming the odds (prices) have been set correctly before the over-round is added in. Therefore, before assessing market accuracy or efficiency, it is important to consider that the over-round contained in the average market odds has decreased significantly over time (from 1.106 in 2005/06 to 1.0579 by 2019/20). Assessments of the market performance aggregated over time are therefore sometimes best assessed using normalised odds (so that they sum to 1).

Table 1 shows the observed proportion of home wins, draws and away wins during the pre-Covid period since 2005/06, as well as the mean of the market implied normalised probabilities for the same outcomes. This indicates that on average the market is pricing these outcomes very accurately.

Table 1: Observed Proportion versus Market Implied (Normalised) Probabilities

	Home Win	Draw	Away Win
Observed Proportion	0.463	0.253	0.285
Mean (Implied) Probability	0.448	0.259	0.292

The calibration plots in Figure 1 provide a different perspective on the performance of the market for each of the three outcomes. In this case, each plot was derived by binning the non-normalised market implied probabilities into 100 bins, with approximately 500 matches contained in each bin. For the matches in each bin, the observed proportions of home wins, draws and away wins are then plotted against the associated mean of the market implied probabilities for the relevant outcomes. These were not normalised, as was the case earlier, to retain the over-round since here we also wish to assess the performance of the model from a profit/loss perspective. The line y=x (dashed line) represents the break-even situation for the market, such that points below the line is indicative of the market making an overall profit, whereas points above the line indicate a loss. The fitted regression lines show the overall mean (linear) relationship.

For all three outcomes the fitted line is consistently below the line y=x, suggesting that pre-Covid the market has been profitable overall for all three outcomes in the home, draw and away market. For home and away wins, the fitted slope is also reasonably parallel to y=x; for both cases (using t-tests) there was no evidence of the slopes being different from unity (p=0.24 home and p=0.67 away). Not only does this indicate that the market was equally profitable across the full probability (and hence odds) range for both outcomes, it

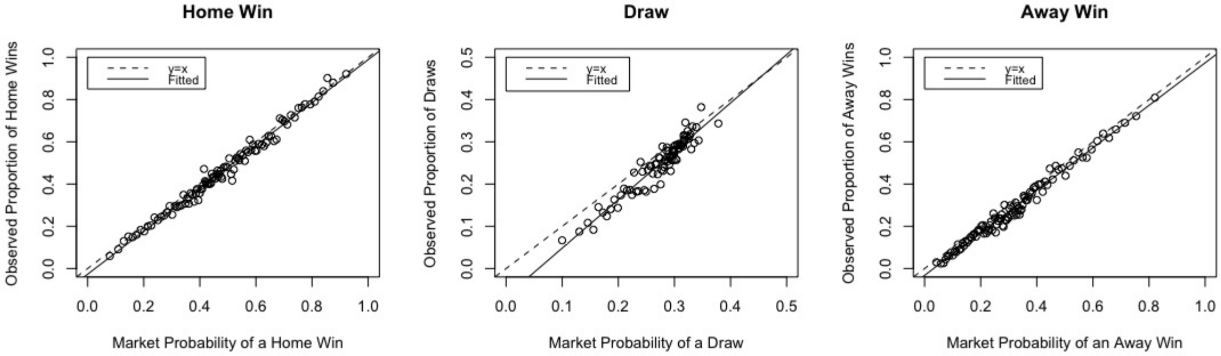


Figure 1: Observed Proportion versus (Non-Normalised) Market Probabilities Pre-Covid

also supports the hypothesis of there being no favourite-longshot bias in the pricing of home and away wins. However, for draws there was evidence of the slope being different from unity (p<0.001), such that the market seems to have been under-pricing draw outcomes (i.e. offering lower than fair odds) at the lower probability (higher odds) end of the market. This also suggests that for matches with an expected lower likelihood of a draw, the over-round is weighted much more heavily on the draw outcome, as opposed to the home or away wins.

Assessments of the change in predictive performance of the market odds over time can be done using scoring rules commonly used to assess probability forecasts. Two of the most common scoring rules, a logarithmic scoring rule (P_1) and the Brier score (BS), are defined as follows:

$$P_1 = \exp\left\{ \frac{1}{N} \sum_{k=1}^{N} \log[P(O_k)] \right\}, \tag{37}$$

$$BS = \frac{1}{N} \sum_{k=1}^{N} [1 - P(O_k)]^2 + [P(NO_{1k})]^2 + [P(NO_{2k})]^2, \tag{38}$$

where $P(O_k)$ represents the market implied probability that match k would result in the eventual observed outcome, $P(NO_{1k})$ and $P(NO_{2k})$ represent the market implied probabilities for the two outcomes, NO_{1k} and NO_{2k}, that were **NOT** observed, such that $O_k, NO_{1k}, NO_{2k} \in S = \{$“home win”,“draw”, “away win”$\}$, whilst N is the number of matches included in the assessment.

Figure 2 shows the values of both measures over time based on a moving window of $N = 10,000$ matches. Larger values of P_1 and smaller values of BS are associated with better predictive performance, and so these plots both illustrate that the market continued to improve in efficiency across the full period of analysis pre-Covid.

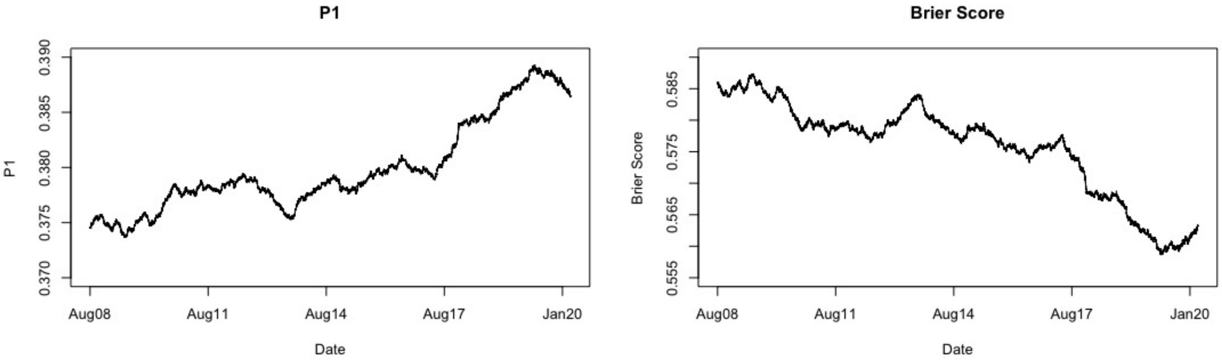


Figure 2: Moving Average Values of P1 and Briers Score Pre-Covid (Rolling Window of 10,000 Matches)

3 Football Betting Market Efficiency pre-Covid versus post-Covid

The pre-Covid period considered here only includes data from the 2017/18 season onwards to allow for at least some of the change over time in the market efficiency that have been discussed above. Figure 3 shows calibration plots similar to those shown earlier, but separates the data into this more recent pre-Covid period and the post-Covid period. For home wins, the fitted line for the post-Covid period (blue dotted line) is consistently below the line pre-Covid (black solid line). However, for draws and away wins the post-Covid fitted lines (blue dotted lines) both lie consistently above the fitted regression line for pre-Covid (black solid line). To examine whether the differences in market performance pre and post-Covid are not due to random chance, three separate binomial logistic regression models were fitted to the individual match data. Home wins, draws and away wins were each treated as three separate binary dependent variables. The relevant match-specific market implied probability was then used as a continuous predictor variable, along with an indicator variable of whether the match occurred pre-Covid (=0) or post-Covid (=1). All models were checked for model adequacy and assumptions.

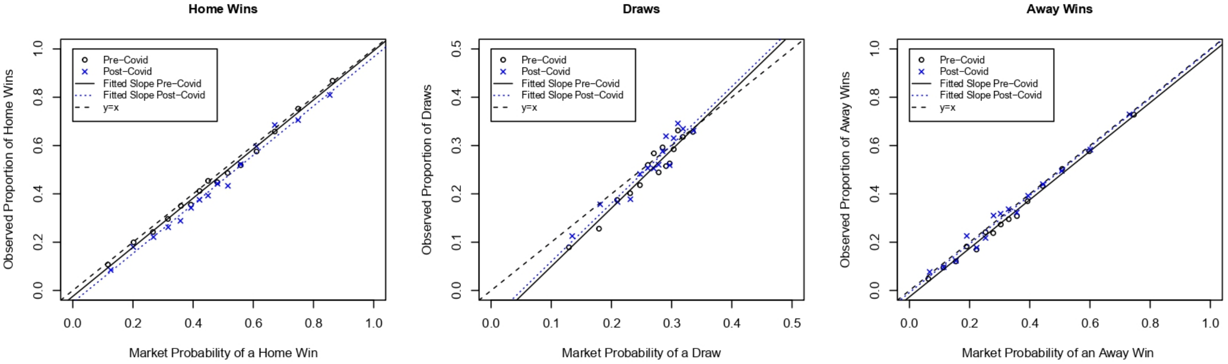


Figure 3: Observed Proportion versus (Non-Normalised) Market Probabilities Pre-Covid versus Post-Covid

Assessments of the interaction term in all three models, supported the hypothesis that any impact of Covid on market performance is the same across the full market probability range (p=0.68 for home wins, p=0.44 for draws and p=0.18 for away wins). The models without the interaction terms are therefore summarised in Table 2. These results would suggest that on average, given equivalent matches pre and post-Covid with similar true probabilities of ending in a home win, draw or away win, the post-Covid market appears to have been under-pricing home wins with relatively lower (shorter) odds and over-pricing away wins and draws with relatively higher (longer) odds compared to pre-Covid. These results would seem to suggest that the market has under-estimated the impact on home advantage, and has not adjusted sufficiently to account for this change.

Table 2: Model Results for Match Outcomes

Dependent	Intercept		Market Probability		Covid (pre=0, post=1)	
Variable	Estimate (s.e.)	p	Estimate (s.e.)	p	Estimate (s.e.)	p
Home Win	-2.45 (0.06)	<0.001	4.76 (0.11)	<0.001	-0.12 (0.04)	0.003
Draw	-3.11 (0.12)	<0.001	7.40 (0.42)	<0.001	0.06 (0.04)	0.20
Away win	-2.60 (0.05)	<0.001	5.02 (0.12)	<0.001	0.10 (0.04)	0.021

The above results would suggest that the post-Covid market should have realised a larger than expected profit on home wins but a smaller profit or even a loss on draws and away wins. To explore this, Figure 4 plots the moving average (over a rolling window of 1,000 matches) for the profit/loss per match, that would have been achieved if bettors had placed unit bets on the home win, draw and away win. This illustrates that from the start of Covid to around half way through the 2020/21 season, there was indeed such an increase in profit on home wins accompanied by falls in profit and even losses mostly with away wins and to some extent draws. However during the later half 2020/21 the market recovered, presumably now having data and information to more accurately measure the impact of Covid on home advantage.

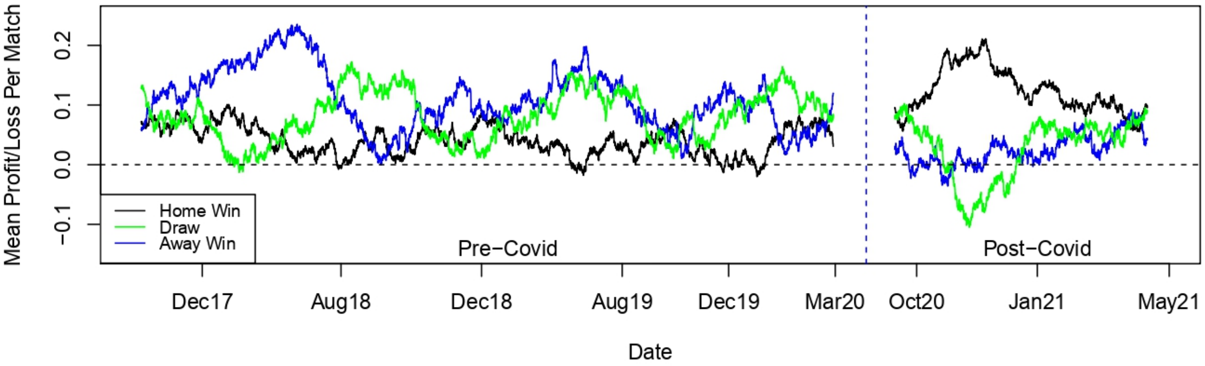


Figure 4: Moving Average Profit/Loss (per match) over Time (Rolling Window of 1,000 Matches)

4 Impact of Covid-19 on Home Advantage

To examine what has changed to reduce the home advantage beyond what had been expected by the market, Table 3 provides comparisons of pre to post-Covid for goals, shots, fouls, yellow cards and red cards. Summary statistics are shown, along with formal assessments of the change pre and post-Covid using GLMs, as indicated in Table 3. All models had the metric as the dependent variable and a binary Covid indicator variable as the independent variable, and all were checked for model adequacy and assumptions.

Table 3: Pre to Post-Covid Comparisons for Various On Field Statistics

	Pre-Covid	Post-Covid	Model Comparisons (Post-Pre)		
Metric	mean (s.d.)	mean (s.d.)	estimate (s.e.)	p	Model
Home Goals	1.56 (1.32)	1.45 (1.26)	-0.07 (0.015)	<0.001	Pois
Away Goals	1.19 (1.17)	1.28 (1.18)	0.07 (0.017)	<0.001	
Goal Difference	0.37 (1.85)	0.17 (1.80)	δ_H =-0.10 (0.028)	<0.001	Skel
			δ_A =0.01 (0.032)	0.79	
Home Shots	12.85 (5.30)	12.02 (5.13)	-0.83 (0.099)	<0.001	Norm
Away Shots	10.30 (4.67)	10.70 (4.77)	0.40 (0.088)	<0.001	
Shot Difference	2.55 (7.74)	1.32 (7.81)	-1.23 (0.145)	<0.001	
Home Fouls	13.24 (4.44)	13.38 (4.23)	0.139 (0.082)	0.090	Norm
Away Fouls	13.59 (4.62)	13.20 (4.35)	-0.389 (0.085)	<0.001	
Foul Difference	-0.35 (5.40)	0.18 (5.24)	0.528 (0.100)	<0.001	
Home Yellow Cards	1.94 (1.40)	1.99 (1.39)	0.023 (0.013)	0.080	Pois
Away Yellow Cards	2.26 (1.45)	2.00 (1.39)	-0.122 (0.013)	<0.001	
Yellow Card Difference	-0.32 (1.76)	-0.01 (1.73)	δ_H = 0.066(0.031)	0.033	Skel
			δ_A = -0.131 (0.029)	<0.001	
Home Red Cards	0.088 (0.283)	0.092 (0.289)	0.049 (0.065)	0.45	Bin
Away Red Cards	0.117 (0.321)	0.099 (0.299)	-0.184 (0.061)	0.003	
Red Card Difference	-0.029 (0.418)	-0.007 (0.405)	0.139 (0.049)	0.005	Ord

Models (GLMs): Pois=Poisson, Skel=Skellam, Norm=Normal, Bin=Binomial Logistic, Ord=Ordinal Logistic.
δ_H and δ_A: Covid effects on (log) home and away rates for the two parameters in the Skellam models.

Home advantage measured in terms of goal difference (home-away) has more than halved, falling from a mean difference of +0.37 goals to a mean difference of just +0.17 goals, in favour of the home team. There were approximately 7% more goals by the away team which represented a statistically significant increase (p<0.001), with 7% fewer goals scored by the home team (p<0.001). However, the results of the Skellam GLM suggests that the decrease in goal difference was actually due to a fall in the home scoring rate (p<0.001) and not any change in away goals (p=0.79). This suggests that any increases in away goals occurred in matches which tended to be higher scoring for both teams. The advantage home teams had in terms of shot difference (home-away) has also almost halved, from just over 2.5 shots extra per match, down to less than 1.5 shots extra per match (p<0.001). Compared to pre-Covid, on average there were 0.83 shots less per match by the home team post-Covid (p<0.001), but an increase of 0.40 shots per match on average by the away team (p<0.001).

Home advantage in terms of disciplinary measures has also seen some startling changes. Pre-Covid, the away team had recorded on average 0.35 more fouls per match than the home team, whereas post-Covid this has swung completely around with the home team now recording on average 0.18 more fouls per match than the away team (p<0.001). However, perhaps most striking of all, is the almost complete disappearance of the additional yellow and red cards that away teams used to incur, over and above those received by the home teams. Pre-Covid the away team used to receive on average an extra 0.32 yellow cards and 0.029 red cards per match than the home team (approximately 1 extra yellow every 3 matches and 1 extra red card per season). However, post-Covid these difference have all but diminished to just 0.01 extra yellow cards and 0.007 red cards per match (1 extra yellow card every 2-3 seasons and 1 extra red card every 3-4 seasons).

5 Discussion

The betting market makers have clearly struggled more post-Covid, offering odds on home wins that have been too high, presumably as a result of under-estimating the size of the impact that Covid (including the absence of fans) will have had on home advantage. Covid has seen home advantage in terms of goal difference falling by a half, with the results suggesting this has been driven by the home team scoring fewer goals, rather than the away team scoring more goals. Shots have seen similar changes, with the home team taking almost 1 shot less per match, whilst the away team has taken almost 0.5 shots more per match. There was also evidence that referees are indeed influenced by the home crowd with fewer fouls being awarded against the away teams, and the almost complete disappearance of the additional yellow and red card sanctions that the away teams used to incur over and above those received by the home team. This would seem to provide strong evidence of a referee bias against the away team reported in previous studies.

References

[1] Boyko, R., Boyko, A. and Boyko, M. (2014) *Referee bias contributes to home advantage in English Premiership football* Journal of Sport Sciences **25** (11), 1185-1194. https://doi.org/10.1080/02640410601038576

[2] Buraimo, B, Forest, D. and Simmons, R. (2010) *The 12th man?: refereeing bias in English and German soccer* Journal of Sport Sciences **173** (2), 431-449. https://doi.org/10.1111/j.1467-985X.2009.00604.x

[3] Ponzo, M. and Scoppa, V. (2018) *Does the Home Advantage Depend on Crowd Support? Evidence from Same-Stadium Derbies* Journal of Sports Economics **19**(4) 562-582. DOI: 10.1177/1527002516665794

[4] Singleton, C., Schreyer, D. and Reade, J. (2020) *As football returns in empty stadiums, four graphs show how home advantage disappears* The Conversation. https://theconversation.com/as-football-returns-in-empty-stadiums-four-graphs-show-how-home-advantage-disappears-138685

[5] Reade, J. (2021) *In European Soccer, Home-Field Advantage Has Survived Covid-19, But Not By Much*. https://www.forbes.com/sites/jamesreade/2021/03/19/the-home-field-advantage-has-survived-covid-19-but-not-by-much/?sh=4fb736ea5400

Football without fans — further investigations

J. James Reade*

*Department of Economics, University of Reading + email address: j.j.reade@reading.ac.uk

Abstract

A range of papers across a number of fields have investigated the impact of football being played around the world without fans during the Covid-19 Global Pandemic. The broad conclusion is that home advantage has been reduced but not completely eradicated, and that there is significant variation across leagues. The mechanism seems clear, namely that it is referee decisions that are impacted by the absence of crowds.

As such, it remains that some home advantage exists even without fans. Common explanations for the remainder of the home advantage are the familiarity associated with being in home surrounds, and the fatigue associated with travel.

In this paper a dataset detailing each on-the-ball event in football matches across 20 competitions worldwide over the last two years is used to investigate further the home advantage in games played both with and without fans.

1 Introduction

The Covid-19 pandemic has profoundly affected the global economy. One of the many visible aspects of the Covid-19 pandemic has been sporting events taking place without spectators present. Prior to the pandemic, [5] document that in elite European football, there had been merely 160 matches played without fans out of over 30,000 matches since 2003, yet since the initial suspension of all sporting competition in Spring 2020, there have been thousands of matches played without crowds. [1] find 1,498 matches across 6,481 matches in 17 countries over the remainder of the 2019/2020 seasons around the world.

The impact appears stark, too: in many leagues the advantage that playing at home bestowed upon the home team — that they would win more often than they ought to given the relative strengths of the competing teams — completely evaporated. In England's Premier League in the 2020/2021 season, out of 380 matches 153 (40%) finished as away wins while only 144 (38%) finished as home wins. In the 2018/2019 season, the last complete season before the pandemic, there was 180 home wins (47%) to 129 away wins (34%).

[1], one of many papers investigating this pattern, found that the overall impact on outcomes was small, to the extent that once controlled for by a range of explanatory variables, it wasn't statistically significant. But it was there, nonetheless, and clearly visible in data plots.

In this paper, we investigate the impact of an absence of fans using a highly detailed dataset. We collect data from the popular website www.whoscored.com, which gives details second-by-second actions from football matches around the world. We use this to consider the particular types of actions for which outcomes differed in the two categories of matches.

We find evidence to support both a fan-influenced referee and player mechanism for influencing home advantage. In Section 2 out methodology is set out, in Section 3 we present details on our data, and in Section 4 we present our results. Section 5 concludes.

2 Method

Sport, and football, can be thought of in terms of a contest function [6]:

$$p_{ij} = \frac{e_i(c_i)^\gamma}{e_i(c_i)^\gamma + e_j(c_j)^\gamma}, \tag{39}$$

where:

- p_{ij}: Probability team i wins contest against team j.
- e_i (e_j): Effort level team i (j) expends.
    - Effort a function of costs c_i (c_j).
- γ: Parameter dictating how contest translates effort into likely outcomes.

The home advantage can be framed in this context. It can be that the cost of producing effort is higher for the visiting team: the cost of travel, and the impact of fatigue, for example.

With no fans in the stadium, home teams found the cost of producing effort increased. The reason that the cost increased though is not clear. One hypothesis is that in normal times, the home team benefits from refereeing decisions that reduce the cost of producing the required level of effort in matches. For example, referees are more likely to call fouls against the away team [4], and are more likely to award more injury time if the home team is losing [2, 3].

To investigate this further, we consider a number of outcome measures from football matches over the last two years, and consider the extent to which games played without fans have differed in nature to those with fans.

We focus on the referee mechanism. We firstly consider the awarding of cautions (yellow and red cards), before considering the underlying playing patterns. We consider firstly fouls, before moving to consider more general patterns of play. We consider the kinds of actions on the field that can often lead to fouls: aerial duels, dispossessions, dribbles, interceptions, and tackles.

The dispensing of cautions is a referee decision, yet it is based on both the behaviour of the referee and also players. Fouls are a different measurer of player behaviour, but again the foul must be called by the referee.

In all cases, we run a simple linear regression model:

$$y_{ijt} = \beta_0 + \beta_1 elo_{it} + \beta_2 elo_{jt} + \beta_3 att_{ijt} + \beta_4 closeddoors_{ijt} + e_{ijt}. \tag{40}$$

We restrict attention to football since August 2018.

3 Data

We collect data from two sources. Our first source is www.football-data.co.uk, a source of match results and bookmaker prices back many years, for a range of European league competitions. Our second source is www.whoscorerd.com, a website that presents hugely detailed in-match data about a wide range of football matches around the world.

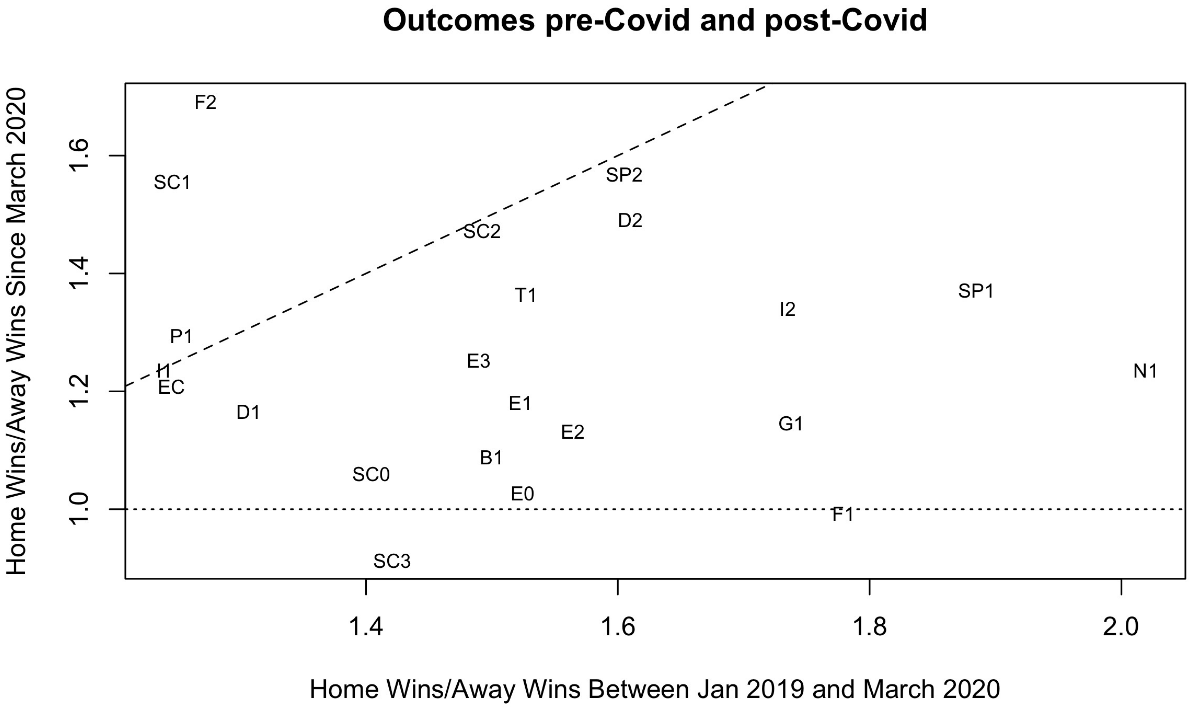


Figure 1: Ratio of home wins to away wins, pre- and post-Covid-19 initial shutdown.

4 Results

Analysis of the headline data on outcomes

The regressions in Table 2 are associated with player and referee behaviour. The first three columns correspond to fouls by each team in matches. Post-Covid-19, almost half a foul more has been called against home teams, and 0.2 fouls fewer against away teams. However, from the second set of three columns, no more yellow cards have been awarded to home teams despite more fouls being called. Around 0.4 fewer yellow cards have been given to away team players.

In sum, this means that while there's been a slight, insignificant increase in the number of fouls a home player can commit before receiving a yellow card of 0.162. For away players, however, this has increased substantially — visiting players can now commit almost an entire extra foul before recieving a yellow card.

If we assume all fouls are identical, then this suggests that the Covid impact is via referees rather than via players.

We can go into more depth by using Whoscored data. In Table 3 we look at in-match events thaat signal the break-up of play, since these are likely the kinds of actions that lead to fouls being called.

We look at aerial duels, dispossessions, interceptions, and tackles. We find almost no impact on the number of aerial duels by either side, a decrease in dispossessions by both teams, an insignificant and small

Table 1: Regressions on match outcomes pre-/post-Covid. Data: www.football-data.co.uk

	Dependent variable:		
	outcomeH	outcomeD	outcomeA
	(1)	(2)	(3)
elopredict	0.779***	−0.072**	−0.707***
	(0.027)	(0.014)	(0.025)
post.covid	−0.021**	−0.015**	0.037**
	(0.005)	(0.003)	(0.007)
Observations	22,223	22,223	22,223
R^2	0.073	0.005	0.072
Adjusted R^2	0.072	0.004	0.071
Residual Std. Error (df = 22197)	0.477	0.440	0.445
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Regressions on fouls and cards pre-/post-Covid-19. Data: www.football-data.co.uk

	Dependent variable:								
		Fouls			Yellows			Fouls per Yellow	
	Home	Away	Diff	Home	Away	Diff	Home	Away	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
elopredict	−2.180**	0.694	−2.874*	−1.010***	0.412**	−1.423***	1.312**	−0.900	2.212*
	(0.365)	(0.484)	(0.834)	(0.058)	(0.061)	(0.102)	(0.235)	(0.317)	(0.575)
post.covid	0.562**	−0.202***	0.764**	0.076	−0.379***	0.455***	0.162	0.823***	−0.662***
	(0.085)	(0.011)	(0.088)	(0.041)	(0.026)	(0.018)	(0.088)	(0.078)	(0.056)
Observations	20,245	20,245	20,245	22,198	22,198	22,198	20,245	20,245	20,245
R^2	0.166	0.138	0.015	0.109	0.098	0.030	0.024	0.029	0.007
Adjusted R^2	0.165	0.137	0.014	0.108	0.097	0.029	0.023	0.028	0.006
Residual Std. Error	3.828 (df = 20220)	3.963 (df = 20220)	5.239 (df = 20220)	1.273 (df = 22172)	1.327 (df = 22172)	1.688 (df = 22172)	4.156 (df = 20220)	4.119 (df = 20220)	5.585 (df = 20220)
Note:									*p<0.1; **p<0.05; ***p<0.01

increase in the number of interceptions, and a small but significant decrease (aabout 1 per team per match) of tackles.

Overall, there isn't an increase in the kinds of actions that lead to fouls, and yet there are more fouls given against the home team.

Table 3: Regressions on different types of play relating to the break up of possession. Data: www.whoscored.com

	Dependent variable:							
	Aerial		Dispossessions		Interceptions		Tackles	
	home	away	home	away	home	away	home	away
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
attendance	0.00000	0.00000	0.00000	0.00000	0.00001**	0.00001	0.00004***	0.00002**
	(0.00002)	(0.00002)	(0.00001)	(0.00001)	(0.00001)	(0.00000)	(0.00001)	(0.00001)
closed.doors	0.010	0.010	−0.697***	−1.140***	0.203	0.247	−0.990***	−1.142***
	(0.526)	(0.526)	(0.179)	(0.156)	(0.165)	(0.164)	(0.244)	(0.314)
elopredict	1.779	1.780	0.292	−1.378***	−2.075***	2.069***	−2.543***	1.689**
	(2.019)	(2.019)	(0.453)	(0.433)	(0.490)	(0.475)	(0.831)	(0.804)
Observations	12,250	12,250	12,208	12,208	12,208	12,208	12,208	12,208
R^2	0.545	0.545	0.178	0.182	0.175	0.185	0.299	0.297
Adjusted R^2	0.522	0.522	0.136	0.141	0.133	0.143	0.264	0.261
Residual Std. Error	11.341 (df = 11658)	11.341 (df = 11658)	3.693 (df = 11616)	3.727 (df = 11616)	3.932 (df = 11616)	4.040 (df = 11616)	6.317 (df = 11616)	6.375 (df = 11616)
Note:	*p<0.1; **p<0.05; ***p<0.01							

We can also look at passing actions in a match, and do so in Table 4. We find that passing has increased, and successful passing in particular for the away team. The home team makes significantly fewer key passes.

Table 4: Regression results for different types of passes. Data: www.whoscored.com

	Dependent variable: Passes							
	All		Accurate		Successful		Key	
	home	away	home	away	home	away	home	away
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
attendance	−0.00002	−0.0001	−0.0001	−0.0001	−0.001	−0.001	−0.00000	−0.00000
	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.001)	(0.001)	(0.00001)	(0.00000)
closed.doors	7.581*	12.306***	7.516*	13.099***	−26.666	116.094***	−0.961***	−0.066
	(3.868)	(4.094)	(3.919)	(4.162)	(28.409)	(33.599)	(0.155)	(0.159)
elopredict	131.852***	−118.193***	131.235***	−114.911***	1,031.605***	−949.521***	0.638	−1.377***
	(11.939)	(12.329)	(12.413)	(12.325)	(101.301)	(94.122)	(0.450)	(0.408)
Observations	12,259	12,259	12,259	12,259	12,259	12,259	12,259	12,259
R^2	0.565	0.540	0.604	0.576	0.509	0.498	0.302	0.265
Adjusted R^2	0.543	0.517	0.584	0.555	0.484	0.473	0.267	0.228
Residual Std. Error (df = 11667)	80.406	78.649	80.228	78.602	693.805	697.590	3.641	3.392
Note:	*p<0.1; **p<0.05; ***p<0.01							

While teams are passing more, it is likely then that they are running with the ball less. However, from Table 5 that impact is tiny and marginally significant on total dribbles by the home team, although they have many fewer successful dribbles.

Table 5: Regression results for different types of running with the ball. Data: www.whoscored.com

	Dependent variable: Dribbles			
	All		Successful	
	home	away	home	away
	(1)	(2)	(3)	(4)
attendance	0.00002**	0.00003***	0.001	0.001***
	(0.00001)	(0.00001)	(0.0005)	(0.0004)
closed.doors	−0.441*	0.152	−24.347*	6.535
	(0.247)	(0.193)	(12.774)	(10.743)
elopredict	1.390**	−1.165*	41.430	−96.099***
	(0.663)	(0.651)	(36.053)	(32.638)
Observations	12,208	12,208	12,208	12,208
R^2	0.311	0.314	0.265	0.268
Adjusted R^2	0.276	0.279	0.228	0.231
Residual Std. Error (df = 11616)	5.198	5.004	288.335	281.063
Note:		*p<0.1; **p<0.05; ***p<0.01		

5 Conclusions

Football without fans has been different during Covid-19. However, the simple narrative that home advantage has been reduced glosses over the mechanism, and also the differences across countries and leagues.

In this paper we look at detailed in-match data to investigate this further.

This deeper investigation into the impact of ghost games suggests that play has become more passing oriented, with the away team more able to retain the ball. A possible reason for this is that without home fans encouraging the home team to close down the visiting team, there is less pressure on visiting players.

There are thus fewer opportunities for the types of passage of play that often lead to fouls being awarded. Nonetheless, fouls by the home team has increased slightly, and fewer yellow cards have been awarded to visiting players.

It suggests an interplay between the impact of the referee and the changing style of the game without fans.

References

[1] A. Bryson, P. Dolton, J.J. Reade, D. Schreyer, and C. Singleton. Causal effects of an absent crowd on performances and refereeing decisions during covid-19. *Economics Letters*, 2020.

[2] T.J. Dohmen. The Influence of Social Forces: Evidence from the Behavior of Football Referees. *Economic Inquiry*, 2008. doi: 10.1111/j.1465-7295.2007.00112.x. URL `http://www.blackwell-synergy.com/doi/abs/10.1111/j.1465-7295.2007.00112.x`.

[3] L. Garicano, I. Palacios-Huerta, and C. Prendergast. Favouritism Under Social Pressure. *Review of Economics and Statistics*, 87(2):208–216, 2005.

[4] A.M. Nevill, N.J. Balmer, and A.M. Williams. The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, 3(4):261–272, 2002.

[5] J.J. Reade, D. Schreyer, and C. Singleton. Eliminating supportive crowds reduces referee bias. Discussion Paper em-dp2020-14, Department of Economics, University of Reading, 2020.

[6] S. Szymanski. The economic design of sporting contests. *Journal of economic literature*, 41(4):1137–1187, 2003.

The Financial Impact of Financial Fair Play Regulation: Evidence from the English Premier League

Mobolaji Alabi*

*ICMA Centre, University of Reading, UK + email address: m.o.alabi@pgr.reading.ac.uk

Abstract

In response to the ailing financial situation of European football clubs, UEFA introduced Financial Fair Play (FFP) regulation in 2011 to guide clubs towards profitability and sustainability. The existing literature is inundated with the impacts of FFP on footballing competitiveness. However, this study focuses on FFP's impact on financial performance. The English Premier League (EPL) – with the highest losses and debt level in Europe pre-FFP – is an excellent scene to assess the effectiveness of FFP. We measure the impact of FFP by adopting the Difference-in-Differences strategy. We find statistical significance for improved profitability; however, financial sustainability (solvency) is yet to experience similar improvements. We believe our findings and subsequent regulatory recommendations can further improve the financial performance of football clubs.

Keywords: Financial Fair Play; English Premier League; Football Finance; and UEFA.

1 Introduction

In 2010, football clubs in the top five European leagues – English, Spanish, German, French and French – generated €12.bn in revenue (6% increase) yet, financial losses for the year were €1.6bn (33% increase). Against the backdrop of financial losses in European football, UEFA announced the introduction of FFP regulation in 2011 to curtain the poor financial health of football clubs. The cornerstone of FFP is the break-even requirement (BER). By mandating football clubs to keep (relevant) expenses to not more than €15m above (relevant) income over three years, BER aims to introduce the concept of "living within means" in European football [UEFA, 2011].

European football is unique. Unlike other jurisdictions, the desire for on-field success in Europe vastly outweighs any other objective. Board-level conversations only tilt to the club's finances as a recourse to stem poor on-field performances or maintain recent success [Sloane, 1971]. The financial performance of the club is a means to end, rather than an end. The high correlation between on-field success and player wages [Szymanski, 2003] is why 78 clubs in Europe spent more than they earned on player wages to pursue success in 2006 [Franck and Lang, 2014].

Wealthy owners and bank loans subsidised the revenue shortfall, thereby increasing the clubs' indebtedness. This safety net provided a financial buffer for excessive risk-taking, and football was dubbed as a "too popular to fail" industry [Rothenbücher et al., 2010]. However, high-profile club failures – for example, Leeds United, Parma, Portsmouth and Deportivo La Coruna – necessitated regulatory intervention.

Recently, studies in football have attempted to assess the impact of FFP in European football. The nomenclature of FFP alludes to a level playing field for football clubs. Evidence suggests that competitive balance in European football is yet to improve, with big clubs still dominating on-field success [Birkhäuser et al., 2019]. Nonetheless, FFP's explicit aim is financial performance. Recent studies [Franck, 2018, Caglio et al., 2019] found improvement in financial performance across the five top European leagues. [Ahtiainen and Jarva, 2020] studied the same leagues and found a significant positive impact only in Spain.

The level of competition, the magnitude of revenue, ownership structure, economic position and regulatory regimes differ in Europe. Thus, a country-specific study might provide more insight into the impact of FFP. Therefore, the primary aim of this study is to assess the impact of FFP in the English Premier League (EPL) because; the average losses and debt reported by English clubs are the highest in Europe [Caglio et al., 2019], revenue and player wages in England is the highest in Europe [UEFA, 2011] and there is evidence of improved competitive balance in England [Wilson et al., 2018]. We focus on profitability and financial sustainability in the EPL, post-FFP's introduction.

This paper is organised as follows. Section 2 develops the theoretical framework for assessing the impact of FFP and describes the data. Section 3, the methodology is detailed. Section 4 shows the regression results and conclusion in Section 5.

2 Theoretical Framework and Data

The FFP regulation requires clubs in Europe to submit reports separate from statutory financial statements. This is because the financial reporting standards do not require companies to report measures such as BER. The submission required by UEFA is an additional disclosure of information for measuring clubs' adherence to FFP. By disclosing financial information relating to BER, UEFA expects clubs' to align to the objective of FFP or face sanctions. This expectation is referred to as the real effect hypothesis in finance and accounting [Kanodia and Sapra, 2016].

We collected the financial statements for 37 English football club based on their participation in the (EPL) between 2005 and 2019. Our criteria – including clubs that participated at least twice in the sample period – excluded only one football club, Blackpool, from our sample. We obtained the financial statements of the football clubs from either the football clubs' official website or filling with the Companies House. Based on UEFA's definition of BER, we extracted the financial data required to proxy the BER. UEFA defines BER as the difference between relevant income and relevant expenses.

To capture financial sustainability, we adopted net cash-flow-from-operation to total debt ratio (CFTD). English football clubs are known to be heavily levered, and though high cash inflow permitted this, recently, football clubs' net cash flows are sometimes negative. As such, the conventional debt to cash flow measure might skew the results obtained. The higher the CFTD, the higher a football club's ability to meet its financial obligations, vice versa.

Table 1: Descriptive Statistics

English football clubs (555 observations)	Mean	Std.Dev	Max	Min	Median	Obs	Coverage
Commercial	24.4	43.4	276.1	0.4	8.6	474	85%
Matchday	20.7	26.8	154.3	1.8	10.4	474	85%
Broadcast	50.5	50.1	260.8	0.0	38.2	474	85%
Relevant income	101.9	117.3	655.1	4.5	59.9	522	94%
Wages and salaries	58.9	58.1	332.4	3.6	38.4	522	94%
Players sales profit/(loss)	10.3	16.6	123.9	-12.7	4.3	522	94%
Relevant expense	106.3	111.4	636.9	7.1	66.2	522	94%
Break-even-requirement (BER)	-4.3	28.6	141.6	-191.5	-4.0	522	94%
Net transfer received (paid)	-15.6	33.5	45.3	-249.7	-4.7	522	94%
Total debt	152.6	223.7	1726.0	3.7	78.2	522	94%
Debt to assets	145%	149%	1337%	19%	103%	522	94%
Cash flow (from operations)	11.2	38.4	245.0	-82.6	0.9	432	78%
Cash flow to total debt (CFTD)	0.02	0.25	1.83	-2.00	0.01	432	78%
Wages as % of revenue	85.0%	34.9%	253.7%	37.3%	75.7%	522	94%
players expense as % of revenue	105.4%	41.4%	353.5%	43.7%	95.4%	522	94%

Note: All figures are in millions except for percentages. The figures represent the variables required for assessing the impact of FFP. The additional variables wages as a % of revenue and player expenses as a % of revenue were collected because of a UEFA guide to clubs to keep these percentages at the 70% mark.

3 Methodology

We adopt the difference-in-differences (DiD) approach to estimate the causal effect of FFP. DiD is a quasi-experiment research strategy that isolates and identifies a planned intervention's treatment effect by estimating counterfactual between two groups (Target and Control groups). For this paper, the introduction and enforcement of FFP in 2011 is our treatment.

We define targeted (treated) football clubs as those that have participated in UEFA competitions in at least 10 out of 15 seasons over our sample period. This is because FFP applies to only clubs that qualify and participate in either of UEFA's club competitions.

An important assumption for the DiD estimation is the parallel trend assumption. It stipulates that in the absence of a treatment/intervention, the existing differences and uniformity of trends for the target and control groups would persist in the post-treatment period.

As we cannot observe what would have happened to the treated group post-intervention, DiD relies on the existing differences and trends in the groups to estimate the counterfactual [Rambachan and Roth, 2019].

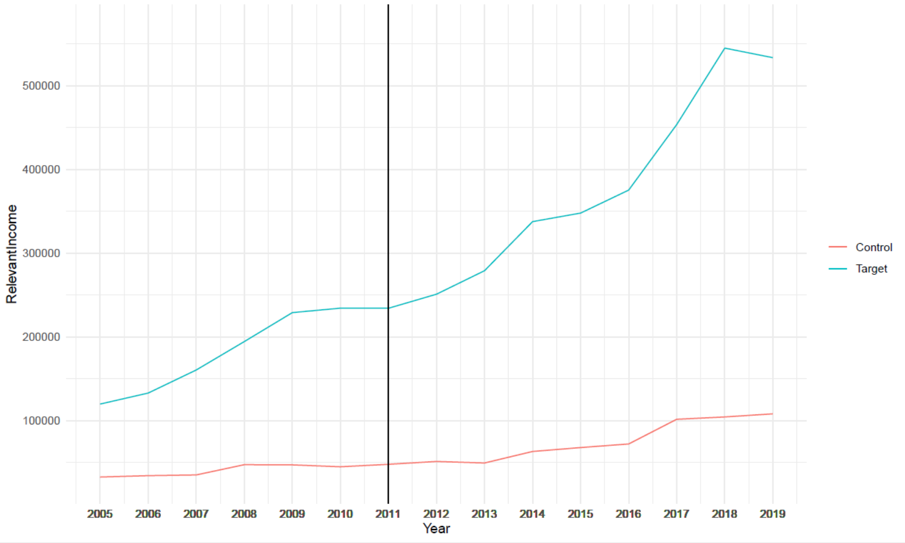


Figure 1: The black line in 2011 signifies the introduction of FFP. Relevant income for the control and target group are similar in trends pre-2011

It is impossible to be sure that similar trends would subsist; however, we can gain comfort and support for the assumption. Inspecting the pre-intervention data should not reveal serious deviation in trends for both groups. Apparent and continuous pre-intervention trend deviation does not lend support for the parallel trend assumption.

Figures 1 and 2 above lends support to the parallel trends assumption.

Regression model and research design

Below are the regression models for assessing the impact of FFP.

$$Y_{it} = \beta_1 TARGET_i + \beta_2 POST_t + \beta_{DiD} TARGET_i \times POST_t + CONTROLS + \varepsilon_{it} \tag{41}$$

$$Y_{it} = \beta_{DiD} TARGET_i \times POST_t + CONTROLS + FE + \varepsilon_{it} \tag{42}$$

For all instances where we use the model, Yit is the outcome variable, TARGET and POST are dummy variables that take the value of 1 for clubs treated by FFP and every year after 2011. TARGET x POST is the interaction term between TARGET and POST. βDiD is the DiD and the causal effect coefficient. βDiD estimates the mean difference in outcome variable between targeted and control groups before and after the introduction of FFP.

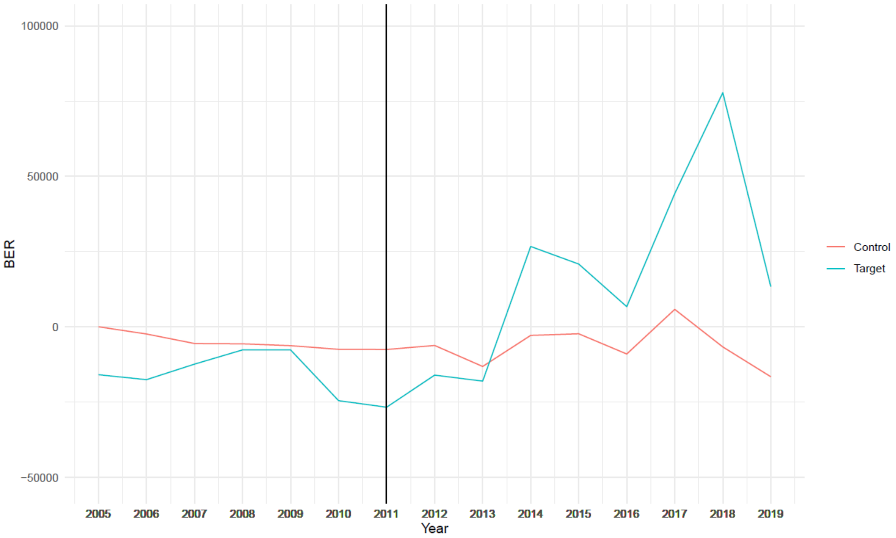


Figure 2: : The black line in 2011 signifies the introduction of FFP. BER for the control and target group are similar in trends pre-2011

4 Results and Discussions

Profitability

The coefficient of interest β_{DiD} is positive, economically and statistically significant in Table 2. The BER β_{DiD} for targeted clubs increased by £38.85m (£39.61m with club and year fixed effect) more than they increased for control clubs after the introduction of FFP.

Table 2: BER, Relevant Income and Relevant Expenses Regressions

	BER		Relevant Income		Relevant Expenses	
	(1)	(2)	(3)	(4)	(5)	(6)
POST(β_1)	-1.23	–	35.75***	–	36.99***	–
	(2.08)	–	(7.44)	–	(6.80)	–
TARGET(β_2)	-13.15	–	145.00***	–	158.15***	–
	(16.23)	–	(25.81)	–	(24.69)	–
POST * TARGET (β_{DiD})	38.85***	39.61***	168.17***	167.87***	129.32***	128.26***
	(8.94)	(9.12)	(26.50)	(26.93)	(22.09)	(22.44)
Constant (β_0)	-4.98***	–	41.60***	–	46.58***	–
	(1.08)	–	(4.43)	–	(4.98)	–
Time fixed effect	–	✓	–	✓	–	✓
Firm fixed effect	–	✓	–	✓	–	✓
Observations	522	522	522	522	522	522
R2	0.083	0.404	0.717	0.857	0.706	0.878
Within R2	–	0.100	–	0.336	–	0.276

Robust standard errors are clustered at club level. All numbers in the table are presented in millions of £. Significance levels denoted as *p<0.1, **p<0.05, and ***p<0.01.

Financial sustainability

The coefficient of interest in columns 1 and 2 of Table 3 suggests that CFTD has improved marginally, but the improvement is not significant. This suggests that FFP is yet to impact the financial sustainability of EPL football clubs significantly.

Table 3: CFTD, Cash flow and Total Debt Regression

	CFTD		Cash flow		Total Debt	
	(1)	(2)	(3)	(4)	(5)	(6)
POST (β_1)	0.021	–	2.65	–	34.63***	–
	(0.037)	–	(2.53)	–	(10.25)	–
TARGET(β_2)	0.133**	–	29.34*	–	329.02***	–
	(0.052)	–	(15.46)	–	(96.70)	–
TARGET x POST (β_{DiD})	0.016	0.026	39.69***	38.91***	170.78**	167.06**
	-0.058	(0.056)	(10.21)	(9.88)	(73.90)	(74.60)
Constant(β_0)	-0.022	–	-0.428	–	60.95***	–
	(0.019)	–	(.792)	–	(8.52)	–
Time fixed effect	–	✓	–	✓	–	✓
Club fixed effect	–	✓	–	✓	–	✓
Observations	432	432	433	433	522	522
R2	0.052	0.312	0.339	0.612	0.543	0.868
Within R2	–	0.001	–	0.095	–	0.130

Robust standard errors are clustered at club level. All numbers in the table are presented in millions of £. Significance levels denoted as *p<0.1, **p<0.05, and ***p<0.01.

5 Conclusions

From the regression results, we see that FFP's introduction has improved the profitability of target football clubs, but we do not find evidence of a similar effect on financial sustainability. A possible explanation for this situation is that football clubs, to remain competitive, spend the profit made in prior periods in the subsequent years, albeit at a reduced rate, post-FFP. At a minimum and in line with one of FFP's objectives, football clubs are exhibiting rationality in their spending – that is, spending within their revenue. Nevertheless, with football clubs increasingly paying for players acquisition via instalments – which has increased short term debt as indicated in Figure 3– financial sustainability is yet to be improved. An increase in short-term debt – consisting mainly of money owed to other clubs – exposes football clubs to higher vulnerability.

The failure of one club could potentially create a domino effect in the football industry. The interconnectedness of football clubs is similar to the banking industry. Therefore, We recommend that UEFA introduce a Basel III style capital requirement which we believe will further strengthen the clubs' financial position for the future.

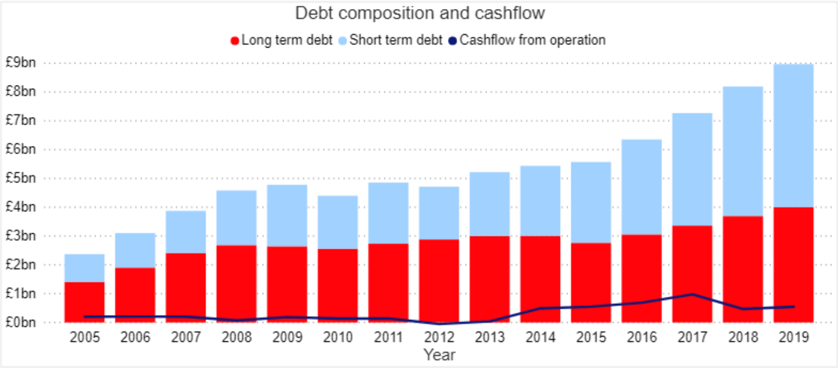


Figure 3: Debt and cashflow graph for all 37 clubs in our data set. A significant portion of short-term debt is money owed to other football clubs for transfer of players while long-term debt is money owed to providers of capital

References

[Ahtiainen and Jarva, 2020] Ahtiainen, S. and Jarva, H. (2020). Has UEFA's financial fair play regulation increased football clubs' profitability? *European Sport Management Quarterly*, 0(0):1–19.

[Birkhäuser et al., 2019] Birkhäuser, S., Kaserer, C., and Urban, D. (2019). Did UEFA's financial fair play harm competition in European football leagues? *Review of Managerial Science*, 13(1):113–145.

[Caglio et al., 2019] Caglio, A., Laffitte, S., Masciandaro, D., and Ottaviano, G. (2019). Does Financial Fair Play Matter? The Real Effects of UEFA Regulation for European Football Clubs.

[Franck, 2018] Franck, E. (2018). European Club Football after "Five Treatments" with Financial Fair Play—Time for an Assessment. *International Journal of Financial Studies*, 6(4):97.

[Franck and Lang, 2014] Franck, E. and Lang, M. (2014). A theoretical analysis of the influence of money injections on risk taking in football clubs. *Scottish Journal of Political Economy*, 61(4):430–454.

[Kanodia and Sapra, 2016] Kanodia, C. and Sapra, H. (2016). A Real Effects Perspective to Accounting Measurement and Disclosure: Implications and Insights for Future Research. *Journal of Accounting Research*, 54(2):623–676.

[Rambachan and Roth, 2019] Rambachan, A. and Roth, J. (2019). An honest approach to parallel trends. *Unpublished manuscript, Harvard University.[99]*.

[Rothenbücher et al., 2010] Rothenbücher, J., Mesnard, X., Rossi, L., Garcia Ramos Lucero, M., and Hembert, E. (2010). The A . T . Kearney EU Football Sustainability Study - Is European football too popular to fail ? pages 1–8.

[Sloane, 1971] Sloane, P. (1971). The Economics of Professional Football: The Football Club as a Utility Maximiser. *Scottish Journal of Politiical Economy*, 17(June):121–145.

[Szymanski, 2003] Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature*, 41(4):1137–1187.

[UEFA, 2011] UEFA (2011). The European Club Footballing Landscape - Club licensing benchmarking report financial year 2010. page 66.

[Wilson et al., 2018] Wilson, R., Ramchandani, G., and Plumley, D. (2018). Mind the gap: an analysis of competitive balance in the English Football League system. *International Journal of Sport Management and Marketing*, 18(5):357.

On detecting structural breaks in cricket run-scoring through sequential time-switched hypothesis tests

Moinak Bhaduri*

*Department of Mathematical Sciences, Bentley University, Massachusetts + email address: mbhaduri@bentley.edu

Abstract

Viewed retrospectively, locating the moments in a cricket innings around which run-scoring patterns changed holds immense relevance in justifying “turning-points” of a match. This work, at its core, is devoted to such change-finding. The progression of an innings is modelled through a self-exciting point process, and a sequence of tests will be deployed to pinpoint structural breaks in both the first and the second-order intensities. Certain novel time-switched statistics will be the key ingredients. Estimation reliability will be affirmed through bootstrapped intervals, and improvements will be quantified through comparisons with several competing change-detection methods. We describe how, without sounding too many false alarms, the estimated change points correspond both to obvious/expected structural breaks (the dismissal of a batsman, the lifting of power-plays, etc.) and to those that are more subtle. We demonstrate how the proximities (quantified through the Hausdorff metric) of identified changes can be exploited to address more complex questions – how, for instance, to quantify a largely qualitative notion of “interesting-ness” (of one match or an entire tournament). Change point-based clustering tools will also be offered.

1 Introduction

Cricket matches, especially those of limited overs, played between two evenly-matched teams, are a delight to witness if the tussle is engaging, and more crucially, stays engaging throughout its evolution. What arrests a spectator’s attention is not so much the volume of runs scored or the personal milestones achieved by a batsman or bowler, not even the enormity of the occasion, but the number of twists and turns a match goes through. In recent times, newer metrics to condense cricket matches are being offered, graphics are being polished, fresher ways of promoting the game are being examined. The dated “batting average", “economy rate", “runs per over", etc., have made way for the more carefully crafted “pressure index", “market valuation of players", “wicket weights of different batting positions", and the like. Saikia et al. (2019) is an excellent resource. Despite the progress, this notion of “interesting-ness" has been hard to encapsulate through one (or even a handful) number. This work suggests an option.

Run-scoring is modelled through a stochastic process, driven by a random intensity. Changes in run-accumulation patterns are identified through an algorithm proposed recently by Bhaduri (2018). The proximity of the identified change-points is shown to represent “intersting-ness", in a way. Section 2 lays out the mathematical construct through which we will describe the operation of run-scoring. It offers, in addition, a fresh way of estimating changes in that pattern, and lists competing methods designed to execute similar jobs. Section 3, through simulations, establishes the superiority of the proposed technique. Section 4 implements the fresh change detection methods on run-scoring and resource utilisation differentials during the England-New Zealand World Cup final, played at Lord's Cricket Ground on July 14, 2019.

2 Theory and methods

2.1 Hawkes process

A continuous time stochastic process $\{N(t)\}_{t\geq 0}$ is routinely deployed to model a bunch of ordered arrival points $t_1 < t_2 < ... < t_n < ...$ representing the global occurrence times of some random phenomenon of interest. $\{N(t)\}_{t\geq 0}$ is also referred to as a counting process with the understanding that $N(t)$ at a given time t, will count the number of observations in $(0,t]$. Please notice the (almost sure) strict ordering in the arrival times, ensuring a simple point process, the type we are examining now, as opposed to an explosive one. Additionally, a function $\lambda(.)$, termed the intensity, given through

$$\lambda(t) = \lim_{h\to 0} \frac{P\{N(t,t+h] \geq 1\}}{h}, t \geq 0, \tag{43}$$

is taken to exist, quantifying the instantaneous probability of observing at least one shock. A $\lambda(.)$ free of time leads to a stationary point process. An increasing $\lambda(.)$ leads to a deteriorating process (where shocks occur more and more frequently as time goes on), while a decreasing $\lambda(.)$ implies an improving sequence (with shocks happening less and less frequently). Regularity conditions on $\lambda(.)$ lead to the independent increment property and ultimately, to a Poisson process, i.e., when $N(t) \sim \text{Pois}(\int_0^t \lambda(x)dx)$, with probability calculations done through

$$P[N(t) = n] = \exp(-\int_0^t \lambda(x)dx) \frac{\{\int_0^t \lambda(x)dx\}^n}{n!}, n = 0,1,2,3.... \tag{44}$$

A detailed description can be had from Rigdon and Basu [14], for instance. Our change point detection proposals, elaborated in Bhaduri (2018) [3] survey processes of the above kind, with purely deterministic choices of $\lambda(.)$. Such a choice, however, assumes the (conditional) intensity is independent of the history - a condition relaxed through a stochastic choice of $\lambda(.)$. A random, data-dependent intensity enables one event to influence another following, an apt requirement to model such events like earthquakes where one major shock inflates the occurrence probabilities of several aftershocks over a close neighbourhood. We opt for

$$\lambda_Y(t) = \lambda_0(t) + \sum_{i=1+max(0,N(t)-r)}^{N(t)} \omega(t - t_i). \tag{45}$$

This choice leads to a Hawkes process (Hawkes (1971) [15]) where the intensity is composed of one (typically) data independent baseline $\lambda_0(.)$, modelling the rate of occurrence of the "major" events, and a data-dependent memory kernel $\omega(.)$, modeling how much of an influence one shock has on the "minor" aftershocks to follow. Hawkes (1971) [15] chose the exponential memory of the form $\omega(u) := \alpha \exp(-\beta u)$, where $\beta > 0$ controls the rate of "forgetting the past". We stick to the exponential choice with $\alpha < \beta$ to ensure stability.

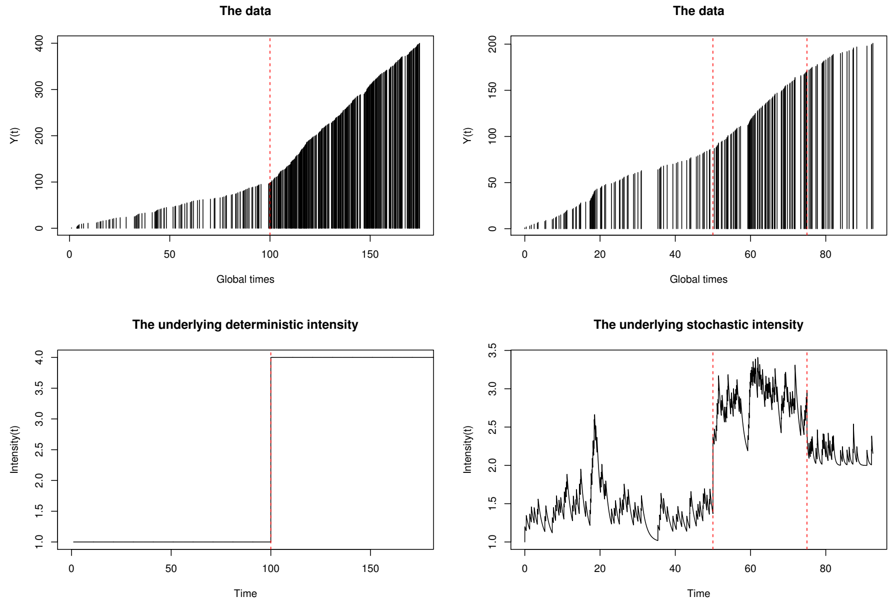


Figure 1: Data reaction to changes under deterministic (left panel) and random (right panel) intensities

Hawkes processes of this type have been used to analyse the arrival times of spam emails (Prince and Heard (2020) [17]), modelling intensity bursts in financial data sets [18], and various others.

Time independent choices of $\lambda_0(.)$ and $\omega(.)$ lead to a stationary Hawkes process. For our analyses, to create a non-stationary version, we have corrupted both the baseline intensity and the memory kernel in our simulation studies (section 3 below) at pre-determined locations and measured the performance of our proposals against established competitors. Figure 1 demonstrates the differences between the environments implied by deterministic and stochastic intensities. A prominent jump in the deterministic step intensity from 1 to 4 at $t = 100$ leads to an obvious heavier crowding on the arrival times while changes (in the baseline $\lambda_0(t)$ from 1 to 2 at $t = 50$ and in the memory kernel $\omega(t)$ from 0.8 to 2.8 at $t = 75$) in the stochastic intensity lead to less discernible changes in the data leading to a more complex change detection exercise.

2.2 Change detection

2.2.1 Known tools

For a thorough discussion on change detection in a longitudinal context, we direct interested readers to some of our previous work: Bhaduri (2018) [3], Zhan et al. (2019), [22], Bhaduri and Ho (2018) [11],Bhaduri and Zhan (2018) [10], Ho and Bhaduri (2017) [8], Bhaduri, Zhan and Chiu (2017) [12], Bhaduri et al. (2017) [13], Ho et al. (2016) [7], Ho and Bhaduri (2015) [21], Tan, Bhaduri, and Ho (2014) [20]. These articles elaborate on the batch and sequential detection scenarios and each one of the CPM-based options to follow. For our current short communication, we briefly touch upon Hawkins et al. (2003)'s [16] general approach. Given a bunch of discrete-time variables X_is, change-locations τ_is may update the underlying probability distributions

Table 1: Choices for the two-sample $D_{k,n}$ statistics under the CPM framework (Hawkins et al. (2003))

Competitor	Construction	Choice
CPM-Exp (Ross (2014))	$M_{k,n} = -2log(\frac{L_0}{L_1})$	$D_{k,n} = M_{k,n}$
CPM-Adjusted Exp (Ross (2014))	$M^c_{k,n} = \frac{M_{k,n}}{E(M_{k,n})}$	$D_{k,n} = M^c_{k,n}$
CPM-Mann-Whitney (Hawkins, Deng (2010))	$U_{k,n} = \sum_{i=1}^{k} \sum_{j=k+1}^{n} sgn(X_i - X_j)$	$D_{k,n} = U_{k,n}$ (scaled)
CPM-Mood (Ross et al. (2011))	$M = \sum_{X_i}((\sum_{i\neq j}^{n} I(X_i \geq X_j)) - \frac{n+1}{2})^2$	$D_n = M$ (standardized)
CPM-Lepage (Ross et al. (2011))	$L = U^2 + M^2$	$D_n = L$
CPM-Kolmogorov-Smirnov (Ross, Adams (2012))	$M_{k,n} = sup_x \lvert \hat{F}_{S_1}(x) - \hat{F}_{S_2}(x) \rvert$	$D_{k,n} = M_{k,n}$
CPM-Cramer-von-Mises (Ross, Adams (2012))	$M_{k,n} = \int_{-\infty}^{\infty} \lvert \hat{F}_{S_1} - \hat{F}_{S_2} \rvert dF_t(x)$	$D_{k,n} = M_{k,n}$

as:

$$X_i \sim \begin{cases} F_0 & \text{if } i \leq \tau_1 \\ F_1 & \text{if } \tau_1 < i \leq \tau_2 \\ F_2 & \text{if } \tau_2 < i \leq \tau_3 \\ \ldots \end{cases}$$

A detection problem, therefore, comprises of choosing one of

$$H_0 : X_i \sim F_0(x;\theta_0), i = 1,2,..,n \tag{46}$$

$$H_1 : X_i \sim \begin{cases} F_0(x;\theta_0), & i = 1,2,...,k \\ F_1(x;\theta_1), & i = k+1,k+2,..n \end{cases} \tag{47}$$

With a given sample size n, some statistic $D_{k,n}$ is constructed, that measures the “difference" between the pre- and post-chang blocks for an arbitrary choice of an initial change estimate at k. These $D_{k,n}$s, in turn, lead to

$$D_{k,n} \Rightarrow D_n = \max_{k=2,3,..,n-1} D_{k,n} \tag{48}$$

and a change is signaled through

$$\phi(D_n) = \begin{cases} 1 & \text{if } D_n > h_n \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

where the threshold h_n is estimated from the null-distribution of D_n, with the estimated change location at

$$\hat{\tau} = argmax_{k=2,3,..,n-1} D_{k,n} \tag{50}$$

Different choices of $D_{k,n}$ lead to different options, working well under different assumptions (changes, for instance, only in the mean or the variance structure, the trend, etc.). Tables 1 and 2 lay them out while the references above provide details. While comparing these options with our proposals below in a continuous time point process setting, we take the X values as the inter-event times, which leads to a discrete time series.

2.2.2 Our proposal

Our approach (Bhaduri (2018) [3]) towards detecting changes offers an algorithm that can operate on continuous time (i.e., the conversion to discrete-time Xs is not needed). Essentially, a test involving a block of n neighbouring event times needs to be conducted. If this test signals stationarity and if a similar test

Table 2: Other parametric options based on likelihood ratio tests, energy divergence, and trend tests

Competitor	Working
E-divergence (Matteson, James (2013, 2014))	$D(X,Y;\alpha) = \int_{R^d} \lvert\phi_X(t) - \phi_Y(t)\rvert^2 (\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma((d+\alpha)/2)} \lvert t\rvert^{d+\alpha})^{-1} dt > C$
Parametric (Chen, Gupta (2011))	$L_k = -2log \frac{L_0(\hat{\lambda})}{L_1(\hat{\lambda},\hat{\lambda}')} < C$
Pettitt (Pettitt (1979))	$K_T = max_{1 \leq t \leq T} \lvert \sum_{i=1}^{t} \sum_{j=t+1}^{T} sgn(X_i - X_j) \rvert > C$
Buishand (Buishand (1982))	$U = \frac{1}{n(n+1)} \sum_{k=1}^{n-1} (\frac{S_k}{D_x})^2$, where $S_k = \sum_{i=1}^{k} (X_i - \bar{X}), D_x = sd(X)$

Table 3: Test statistic proposals for multiple testing (Bhaduri (2018))

Proposal	Critical regions	Comments
$Z = -2\sum_{i=1}^{n} log(t_i/t_n)$	$Z \leq \chi^2_{1-\alpha/2,2n-2}$ or $Z \geq \chi^2_{\alpha/2,2n-2}$	UMPU (Bain and Engelhardt (1991)) in power law setting: $\lambda(t) = \frac{\beta}{\theta}(\frac{t}{\theta})^{\beta-1}$, $t > 0$
$Z_B = -2\sum_{i=1}^{n} log(1 - t_i/t_n)$	$Z_B \leq \chi^2_{1-\alpha/2,2n-2}$ or $Z_B \geq \chi^2_{\alpha/2,2n-2}$	More powerful than Z in detecting increasing step intensities (Ho (1993)). Under further analysis.
$R := max(Z, Z_B)$	$R \geq c^\alpha_R$	Powerful under deterministic intensities (Bhaduri (2018)).
$L := min(Z, Z_B)$	$L \leq c^\alpha_L$	Powerful under deterministic intensities (Bhaduri (2018)).

involving a block of $n+1$ neighbouring event times (the n-many from the previous stage and the immediate next) signals non-stationarity, we estimate a change-point between the n-th and the $n+1$-th event times. More formally, it runs thus:

- Set series of hypotheses: $\{H_1, H_2, ...H_m\}$, p-values: $p_1, p_2, ...p_m$.
- H_i tests stationarity on the first $i+1$ events.
- Order the p-values: $p_{(1)} < p_{(2)} < ... < p_{(m)}$.
- Set $S_i := \{k : p_{(k)} < \frac{k}{m}\alpha\}$
- $$\hat{\tau}_i := \begin{cases} \min\{k : p_{(k)} < \frac{k}{m}\alpha\}, & S_i \neq \emptyset \\ \infty, & S_i = \emptyset \end{cases}$$

Once the earliest significant test (if any) is detected, the algorithm may be restarted with the detected change point as the fresh time origin to discover subsequent changes, if any. The technique is, therefore, free of the “at-most-one-change-point (AMOC)" assumption under which several parametric proposals work. Declaring significance through ordered p-values is done cautiously since performing multiple correlated tests is known to inflate the type-I error probability. We follow the false discovery rate control suggested by Benjamini and Hochberg (1995) [9]. The actual testing is done through an array of novel statistics offered by Bhaduri (2018) [3]. Their definitions and crucial properties are summarized in Table 3. A deteriorating sequence inflates the value of Z_B and deflates the value of Z. Randomized versions of the two through the maximum and the minimum signal general non-stationarities (i.e., both improvement and deterioration) through significantly large or small values. The critical thresholds c^R_α and c^L_α are summarized in Bhaduri (2018) [3].

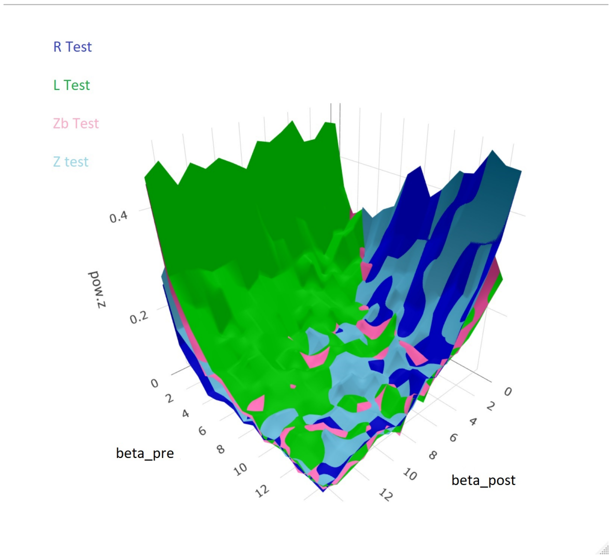


Figure 2: Power study among Z, Z_B, R, and L showing asymmetry

3 Simulation studies

As a preliminary test to identify which one of the four statistics shown in Table 3 has the highest classification accuracy, we have conducted a power study summarized in Figure 2 where the immigrant intensity $\lambda_0(.)$ was taken to be time-inhomogeneous and the only change was brought through the offspring kernel $\omega(.)$. The pre- and post-change β values are placed along the x and the y axes, while the power of each test, the estimated probability of correctly identifying a non-stationary sequence as a non-stationary sequence, is plotted along the z axis. The power surface appeals to intuition. Along the diagonal, when the pre- and the post-change recollections are reasonably identical, detection gets tougher, leading to lower power. In contrast, along the edges and the corners, when the difference is stark, the power rises. We found there is no one statistic that shows the uniformly best power (although the minimum based L test occupies a larger region) and about the diagonal, there exists an asymmetry in the power surfaces.

Next, we investigated, through Figures 3 through 5, the closeness of the change points estimated by our sequential proposals and their competitors to the true ones through a more massive simulation study conducted with 10^4 runs. Each dot signifies a global time (plotted along the vertical axis) of change detection, and under each setting, we have conducted tests on both failure truncated (i.e., when we wait for a specified *number* of shocks, regardless of the time it takes to wait that long) and time truncated (i.e., when we wait for a specified time, regardless of the number of events seen by then) cases. Figure 3 shows the average run length comparisons under the assumption of no change. The heaviest crowding is observed towards the end of the process. This, owing to how non-detections are expressed (please see the previous section), confirms that all our proposals and most of the rest pick up stationarity adequately. The next scenario, graphed in Figure 4 shows the results under a true change in the memory kernel's β value from 3.8 to 0.8 at time 100 (shown through the broken horizontal line), with the baseline intensity held constant. Most of our sequential proposals, especially L and R generate estimations that crowd around this true change point. Another observation is that

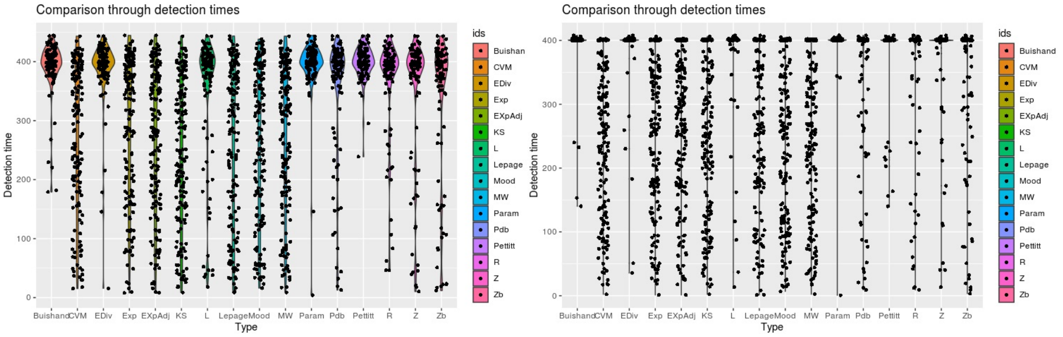


Figure 3: Change identification under stationarity: left panel - failure truncation, right panel - time truncation

quite a few of our competitors estimate a change point prior to the true change - clearly a false alarm. The sequential proposals are largely free of such a problem. Finally, in Figure 5, we bring about changes in both the baseline intensity (from 1 to 2) at time point 50 (shown through the heavy broken horizontal line) and the memory kernel (from 3.8 to 0.8) at time points 75 or 100 (shown through the fainter broken horizontal line). Again, we find it is our sequential proposals that demonstrate the strongest and clearest clustering around these two true change points without sounding too many false alarms.

A natural question to ask at this stage is how will these estimators react to a large influx of data? Asymptotic consistency of these types of estimators, as pointed out by Troung et al. (2020) [23] can be examined through:

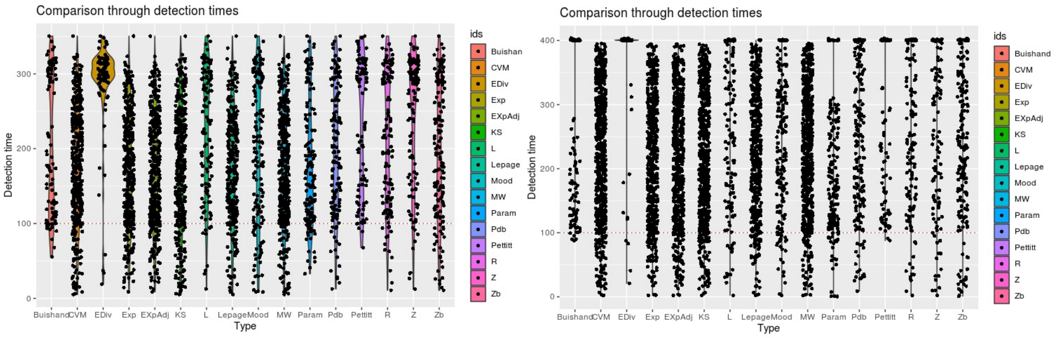
- i) $P(|\hat{\tau}| = K)$
- ii) $\frac{1}{T}||\hat{\tau} - \tau^*||_\infty$

where $||\hat{\tau} - \tau^*||_\infty := \max\{\max_{\hat{t}\in\hat{\tau}} \min_{t^*\in\tau^*} |\hat{t} - t^*|, \max_{t^*\in\tau^*} \min_{\hat{t}\in\hat{\tau}} |\hat{t} - t^*|\}$, with τ^* representing the set of true change points, $\hat{\tau}$ representing the set of estimated change points, and k representing the size of τ^* (i.e., the true number of changes). A change point algorithm is said to be asymptotically consistent if the first probability converges to 1 and the second norm converges to 0 as the terminal time of the process is pushed to ∞. The norm in (ii) represents the Hausdorff norm needed to quantify the “distance" between two sets not necessarily of the same size. It isn’t hard to verify this distance penalizes overestimation quite harshly.

With our simulation cases, calculation of the long-run probabilities of the type (i) is shown through Figure 6. We observe (especially in the time-truncated scenario) this asymptotic probability for most of our competitors drop fast (primarily due to their sounding too many false alarms), while ours either hold steady or increase. This suggests with our sequential offerings, if one waits sufficiently long (either in terms of time or in terms of data), the right number of change pints will be picked, i.e., one won’t over or underestimate with a high chance.

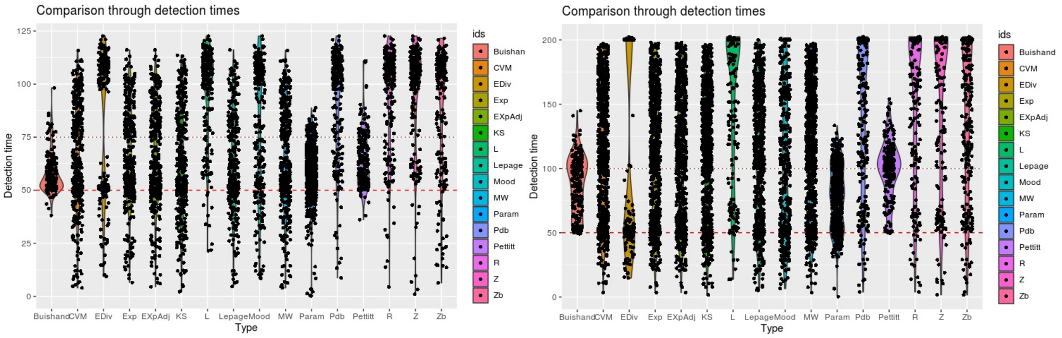
4 Applications to cricketing scenarios

We focus on the much-discussed World Cup 2019 final match, played on July 14, 2019 at Lord’s between England and New Zealand. Batting first, utilising their full quota of fifty overs, New Zealand piled up 241 for the loss of eight wickets. Chasing, England reached that exact total at the end of their fifty, causing a


Comparison through detection times

Comparison through detection times

Figure 4: Change identification under $\beta : 3.8 \to 0.8$: left panel - failure truncation, right panel - time truncation


Comparison through detection times

Comparison through detection times

Figure 5: Change identification under $\lambda_0 : 1 \to 2, \beta : 3.8 \to 0.8$: left panel - failure truncation, right panel - time truncation

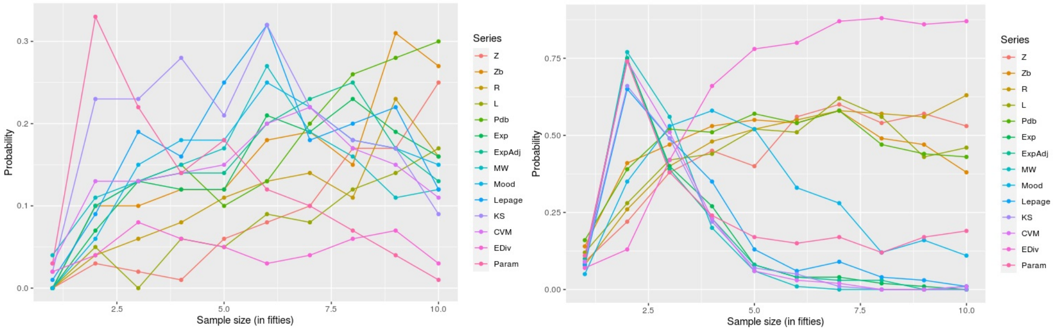


Figure 6: Checking asymptotic consistency with $K = 2$, left panel - failure truncation, right panel - time truncation

super-over, which too, ended in a tie. England, eventually, lifted the trophy due to scoring more boundaries. Ball-by-ball data are collected from www.cricsheet.org.

4.1 Changes in run-scoring while chasing and setting targets

To match the theory laid out in the previous sections, we treat, within an innings, a run scored as an event. Since runs are not always accumulated through isolated singles, we spray (uniformly) the total runs (maximum 7) collected off a delivery between the previous delivery and the current one. This allows the generation of a point process over a (reasonably) continuous time domain.

Run-collections, by both New Zealand and England, over time (played here through, roughly, 300 x 7 = 2100 deliveries) are graphed in Figure 7. The bidirectional algorithm introduced by Bhaduri (2018) described in sections 2 and 3, applied to these patterns will identify structural shifts within these innings at times represented through the vertical separators. Run-scoring rates change (statistically) significantly as one crosses over these separators. Some changes are expected: the one when New Zealand reached 182, for instance. This is a time when Neesham got dismissed. Others, however are subtle. The one around New Zealand's cumulative total of 20 corresponds to the time when J. Archer and C. Woakes started bowling better lines. "Interesting-ness" within an innings and over an entire match may be conveyed through the number and the proximities of these identified change locations.

If we agree to measure closeness between the English and the Kiwi change points through the Hausdorff metric shown in section 3, the heatmap in Figure 9 results, summarizing the scenario when different tools are implemented across the innings. It is interesting to note that if the newer R and L-based algorithms are used on both, this distance turns out to be more minimal indicating an exciting contest, than if any of these options are replaced by primitive alternatives.

Figure 8 clusters detection methods on both innings and shows which methods generate similar change points in each case. We note that for the English innings, all our proposals are grouped together. This agreement is, in part, due to the fact that England were chasing and with a clearly defined target (unlike New Zealand), changes in their progression was more detectable.

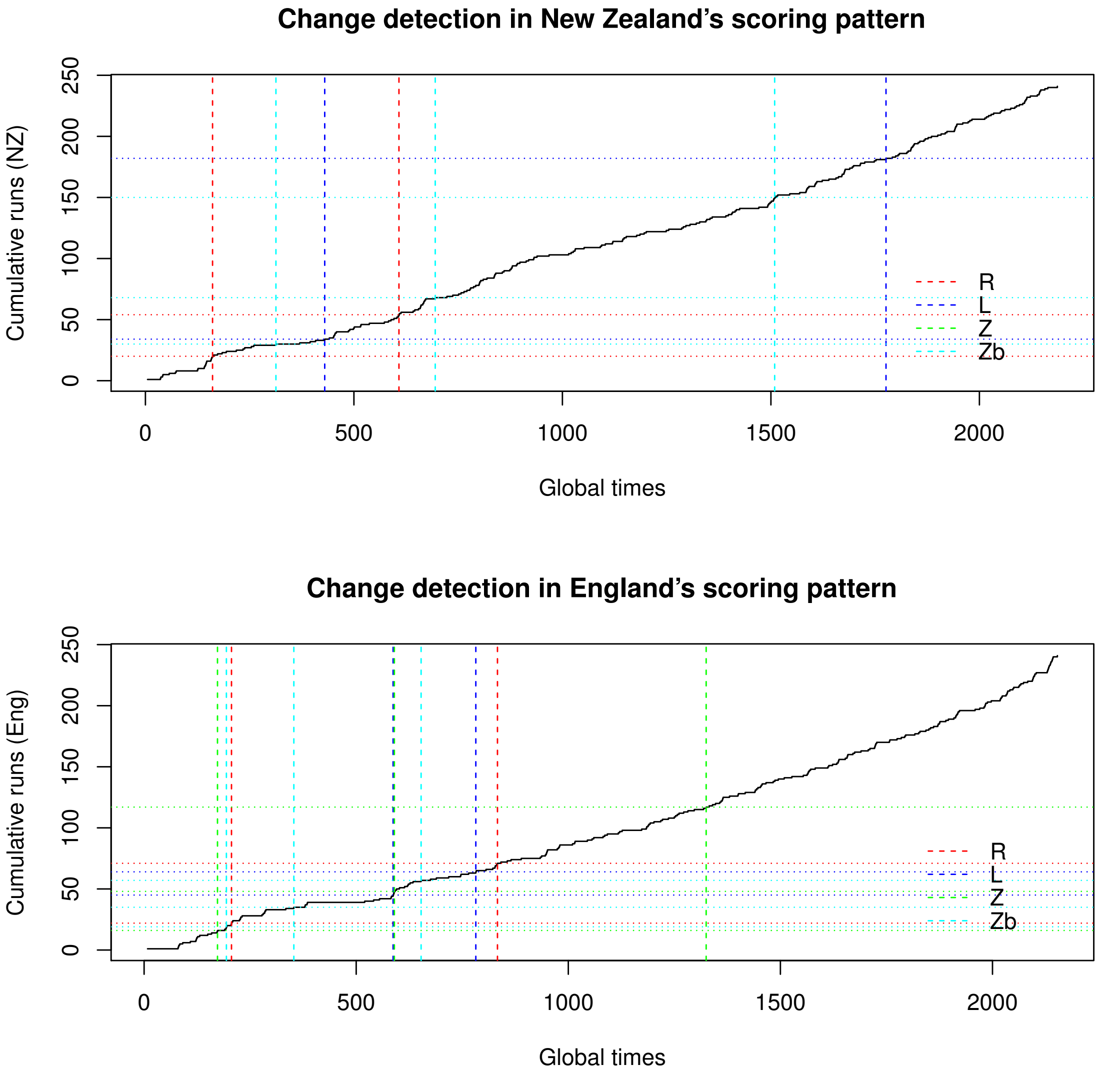


Figure 7: Structural breaks in run-scoring. Changes identified using Z, Z_B, R and L-based techniques (Bhaduri (2018)).

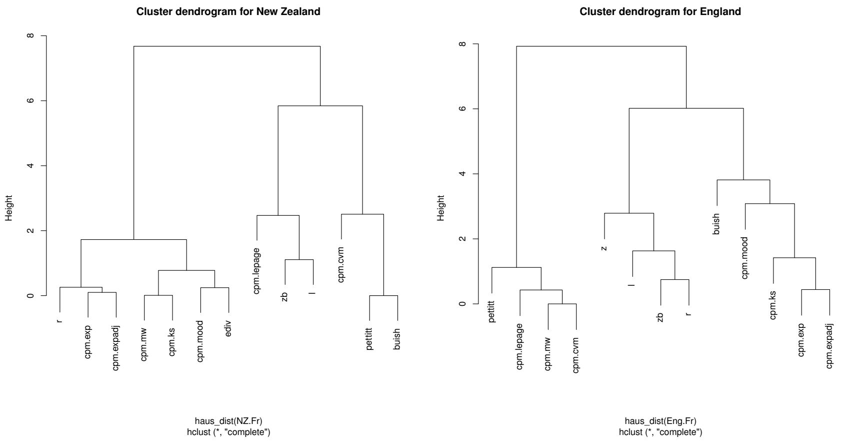


Figure 8: Hausdorff distance-based cluster trees finding which detection techniques identify similar change points.

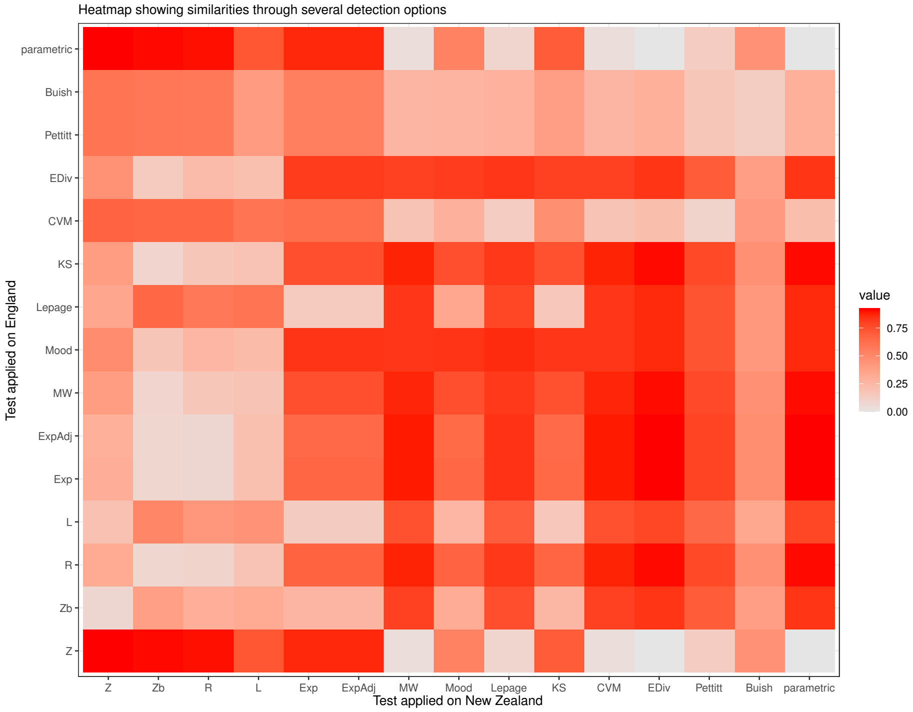


Figure 9: Heatmap demonstrating "interesting-ness" using different change detection options applied on each innings. Low heats indicate smaller Hausdorff distances, implying an engaging contest.

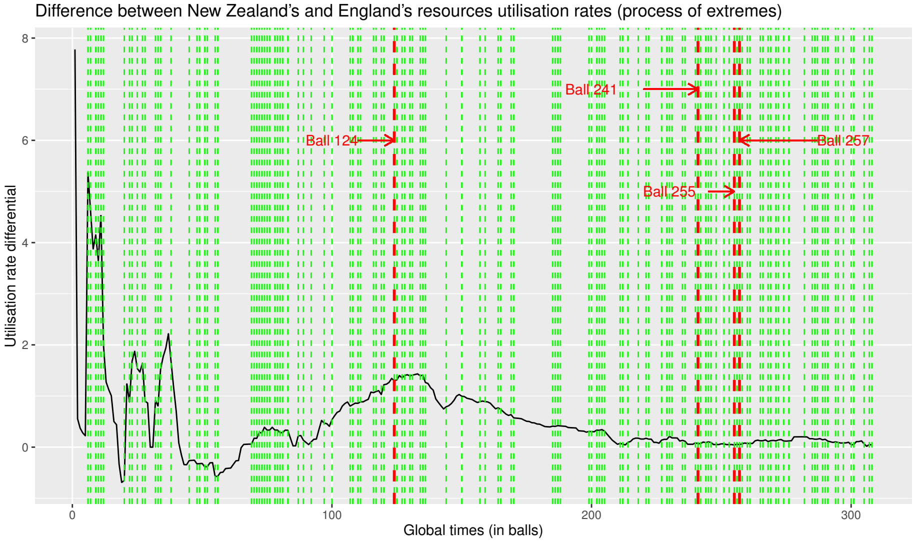


Figure 10: Utilisation rate differential tracked over time (convention: New Zealand - England). Changes identified using R and L-based techniques (Bhaduri (2018)).

4.2 Twists in resource utilisation differentials

The hitting of boundaries or the falling of wickets are known to exert different impacts at different stages of an innings. The concept of resource utilisation was introduced by Duckworth and Lewis (1998) and polished later by Stern (2016), to address this issue. Structurally similar to traditional run-rates, it is defined through $S(u,t)/R(u,t)$ where $S(u,t)$ is the cumulative amount of runs collected at the end of t overs and u wickets, and $R(u,t)$ is the resources used up to reach this point.

Figure 10 shows the difference in these rates (New Zealand - England) as a function of balls bowled. The peaks and valleys of this curve could be of interest. A peak (triggering a downfall) represents a period favourable to England, while a valley (heralding a rise), a phase favourable to New Zealand. We isolate these extremes, these turning points (shown through the green slices), to construct our shock sequence - the sequence that can be described through a point process. The green bands cluster heavily towards the end of the fifty-over period, indicating the match remained competitive till the very end. Some of the change points isolated from the green filtered point process (shown in Fig. 10 through the red separators), the one after delivery 124, for instance, correspond to the breaks identified from the run-sequence shown in Figure 7, matching with those around “time" point 868 (= 124x7, on the sprayed scale), while others, such as the one after ball 241, do not.

5 Conclusions

This work shows a way in which one can talk, in an objective manner, how interesting a cricket match has been. Shifts in run-scoring and in resource utilisation are identified through recent change-detection algorithms. Such shifts, as our analysis on one specific one-day international demonstrates, may or may not occur at expected times (the dismissal of a batsman, the lifting of power-plays, etc.). The number and proximity of these changes are shown to capture the inherent “interesting-ness". The method can be applied in addition to the Empirical Recurrence Rates Ratio-based method introduced in a cricketing context by Bhaduri (2020). Generalisations to cover an entire tournament can also be made.

References

[1] Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7, no. 1, 1–26 (1979).

[2] Braun, Willard & Kulperger, R.J. A bootstrap for point processes. Journal of Statistical Computation and Simulation. 60. 10.1080/00949659808811878. (1998).

[3] Bhaduri, M (2018) *Bi-Directional Testing for Change Point Detection in Poisson Processes* UNLV Theses, Dissertations, Professional Papers, and Capstones. 3217.

[4] Bhaduri, M (2020) *Quantifying the interesting-ness of a limited-over cricket match through Empirical Recurrence Rates Ratio-based change-detection analysis*, Proceedings of the 15th Australasian Conference on Mathematics and Computers in Sport. ISBN: 978-0-646-82267-9.

[5] Duckworth, F. C. and Lewis, A. J. (1998) *A fair method of resetting the target in interrupted one-day cricket matches.* Journal of the Operational Research Society, 49(3), 220– 227.

[6] Stern, S. E. (2016) *The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates.* Journal of the Operational Research Society, 67(12), 1469– 1480.

[7] Ho, C.-H., Zhong, G., Cui, F., & Bhaduri, M. Modeling interaction between bank failure and size. *Journal of Finance and Bank Management* **4**, 15–33 (2016).

[8] Ho, C.-H. & Bhaduri, M. A Quantitative Insight into the Dependence Dynamics of the Kilauea and Mauna Loa Volcanoes Hawaii. *Mathematical Geosciences* **49**, 893–911 (2017).

[9] Benjamini, Y., & Hochberg, Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300. (1995).

[10] Bhaduri, M. & Zhan, J. Using Empirical Recurrence Rates Ratio for Time Series Data Similarity. *IEEE Access* **6**, 30855–30864 (2018).

[11] Bhaduri, M. & Ho, C.-H. On a Temporal Investigation of Hurricane Strength and Frequency. *Environmental Modeling & Assessment* **24**, 495–507 (2018).

[12] Bhaduri, M., Zhan, J. & Chiu, C. A Novel Weak Estimator For Dynamic Systems. *IEEE Access* **5**, 27354–27365 (2017).

[13] Bhaduri, M., Zhan, J., Chiu, C. & Zhan, F. A Novel Online and Non-Parametric Approach for Drift Detection in Big Data. *IEEE Access* **5**, 15883–15892 (2017).

[14] Rigdon, S.E., and Basu, A.P. (2000). *Statistical Methods for the Reliability of Repairable Systems*, Wiley series in Probability and Statistics, John Wiley and Sons.

[15] Hawkes, A. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. Biometrika, 58(1), 83-90. doi:10.2307/2334319 (1971)

[16] Hawkins, D.M., Qiu, P.H., and Kang, C.W. *The Changepoint Model for Statistical Process Control.* Journal of Quality Technology, 35(4):355–366, (2003)

[17] Prince-Williams, M, Heard, N. A. *Nonparametric self-exciting models for computer network traffic*, Stat Comput 30, 209–220, (2020)

[18] Rambaldi, Marcello and Filimonov, Vladimir and Lillo, Fabrizio (2018) *Detection of intensity bursts using Hawkes processes: An application to high-frequency financial data*, Phys. Rev. E, 97(3), 032-318

[19] Saikia, H., Bhattacharjee D., Mukherjee, D. (2019). Cricket performance management: mathematical formulation ad analytics, Springer.

[20] Tan, S., Bhaduri, M. & Ho, C.-H. A Statistical Model for Long-Term Forecasts of Strong Sand Dust Storms. *Journal of Geoscience and Environment Protection* **02**, 16–26 (2014).

[21] Ho, C.-H. & Bhaduri, M. On a novel approach to forecast sparse rare events: applications to Parkfield earthquake prediction. *Natural Hazards* **78**, 669–679 (2015).

[22] Zhan, F. *et al.*. Beyond Cumulative Sum Charting in Non-Stationarity Detection and Estimation. *IEEE Access* **7**, 140860–140874 (2019).

[23] Truong, C, Oudre, L, Vayatis, N *Selective review of offline change point detection methods*, Signal Processing, 167, 107-299. (2020)

More on verification of probability forecasts for football outcomes: score decompositions, reliability, and discrimination analyses.

Jean-Louis Foulley

IMAG, Université de Montpellier, France : foulley-jl@gmail.com

Abstract

Forecast of football outcomes in terms of Home Win, Draw and Away Win relies largely on ex ante probability elicitation of these events and ex post verification of them via computation of probability scoring rules (Brier, Ranked Probability, Logarithmic, Zero-One scores). Usually, appraisal of the quality of forecasting procedures is restricted to reporting mean score values. The purpose of this article is to propose additional tools of verification, such as score decompositions into several components of special interest. Graphical and numerical diagnoses of reliability and discrimination and kindred statistical methods are presented using different techniques of binning (fixed thresholds, quantiles, logistic and iso regression). These procedures are illustrated on probability forecasts for the outcomes of the UEFA Champions League (C1) at the end of the group stage based on typical Poisson regression models with reasonably good results in terms of reliability as compared to those obtained from bookmaker odds and whatever the technique used. Links with research in machine learning and different areas of application (meteorology, medicine) are discussed.

1 Introduction

The list of areas opening to forecast would be exceedingly long if one wishes to draw it up in details: from economic inflation, employment rates, weather and climate change, medical diagnosis and biological tests to media and entertainment market as well as gambling & sporting events among others. Forecasting consists first of producing forecasts from available data and methodologies, and second to assess their quality. Forecasts are basically of two types: pointwise or by means of probability. Quantify the uncertainty of a forthcoming event highlights the superiority of probability forecasts over categorical ones even economically (Winkler & Murphy, 1979; Dawid, 1986). As far as football matches are concerned, outcomes are in terms of either results such as Win, Draw or Loss (WDL)-also abbreviated as Home Win, Draw and Away Win (H,D,A respectively) for two-legged matches- i.e., categorical data, or in terms of scorelines {Y(A), Y(B)} goals in match (A vs B) i.e., pairs of integers: see the review by Read et al. (2021).

Here we will be concerned with probability forecasts of WDL (HAD) results. In this area, much effort has been devoted to statistical approaches to forecasting, especially by modelling outcomes of matches (Scarf and Selliti-Rangel, 2019). There is a growing demand for relevant and acute probabilistic forecasting due to the large audience for TV broadcasting of association football matches and the related huge betting markets. Traditionally in football, one summarizes HDA forecast performance by a synthetic criterium such as the mean square error known as Brier's score (1950). Although this statistic brought useful information, it is only an average measure of the overall accuracy of such predictions. Analytical attributes of this accuracy can be brought out such as Reliability, Resolution, Discrimination, Refinement to shed more light on some qualities or deficiencies of issued forecasts. The purpose of this note is mostly pedagogical with the aim to review the main tools available to that respect, and to illustrate them on data of a well-known club competition with the UEFA champions League (also known in brief as C1) forecasted by simple Poisson regression models. Focus is on a distribution-oriented approach based on the joint distribution of elementary forecast, binary outcome pairs and its factorization. We will deal with several decompositions of the Brier score (BRS) applied to each binary outcome considered separately namely 1) the Murphy decomposition derived from the so-called Calibration-Refinement factorization (CR), and 2) the Likelihood-base (LB) and Yates's decompositions. In parallel to the CR decomposition, we will present a graphical device known as the reliability diagram which allows diagnostic of strength or deficiency in this component. In each case, comparison will be made between forecasts derived from a simple Poisson loglinear model (POI), and to Bookmaker Odds implied probabilities (ODD). Finally, in the last (4th) section, we discuss the main points raised by using these decompositions and possible implications for improving forecast efficiency.

2 Decomposition of the Brier score

2.1 Basic theory

Let X be the binary outcome of the event H with probability q and P the random variable probabilistic forecast of X taking values p. Taking as scoring rule, the quadratic or Half-Brier Score defined as the loss function $S(P,X)=(P-X)^2$, and using the conditioning de conditioning rule, the Murphy (1973) decomposition of its expectation can be written as

$$\mathbb{E}[S(P,X)]=\mathrm{Var}(X)-\mathrm{Var}_P\left[\mathbb{E}_X(X\mid P)\right]+\mathbb{E}_P\left\{\left[\mathbb{E}_X(X\mid P)-P\right]^2\right\}, \qquad (1)$$

where expectation is taken with respect to the pairs of forecasts and outcomes P, X.
Examination of this formula immediately identifies the 3 components of this decomposition:

1) Uncertainty (UNC) equal to $\mathrm{Var}(X)=q(1-q)$, the variance of the outcome that is out of control of the forecaster,
2) Resolution (RES) equal to $\mathrm{Var}_P\left[\mathbb{E}_X(X\mid P)\right]$ referring to the variability between the conditional expectations of the observed outcomes given their forecasts,
3) Reliability (REL) or (Mis) Calibration equal to $\mathbb{E}_P\left\{\left[\mathbb{E}_X(X\mid P)-P\right]^2\right\}$ measuring the average squared differences between the conditional expectation of the outcome and its forecast.

Murphy and Winkler (1987) also gave the dual decomposition of (1)

$$\mathbb{E}[S(P,X)]=\mathrm{Var}(P)-\mathrm{Var}_X\left[\mathbb{E}_P(P\mid X)\right]+\mathbb{E}_X\left\{\left[\mathbb{E}_P(P\mid X)-X\right]^2\right\}. \qquad (2)$$

This decomposition is known as the Likelihood-base factorization as opposed to the Calibration-Refinement factorization of formula (1) referring to the 2-way decomposition of the joint distribution of (P,X). This formula leads to identify three components: i) Refinement (REF) equal to $\mathrm{Var}(P)$, the variance of probabilistic forecasts also known as Sharpness, ii) Discrimination (DIS) equal to $\mathrm{Var}_X\left[\mathbb{E}_P(P\mid X)\right]$ i.e., the variance between conditional distributions of forecasts given the outcomes X, iii) Conditional bias type 2 as called by Bradley et al, (2003) equal to $\mathbb{E}_X\left\{\left[\mathbb{E}_P\left(P\mid X\right)-X\right]^2\right\}$ which is the dual expression of resolution. A special case of interest is the one with P having a probability mass function (pmf) concentrated at $p=\mathbb{E}\left(X\right)$, the mean of the marginal distribution of the binary outcome (the so-called climatological forecast). Then the forecast has no refinement $(\mathrm{Var}\,P=0)$, no discrimination as well, necessarily no resolution, even though it is perfectly calibrated. In fact, it is the only forecast being both reliable with no discrimination (Bröcker, 2012). In that case, $\mathbb{E}[S\left(P,X\right)]$ reduces to its uncertain component UNC as an upper reference value for the expected Brier score. That is the reason why the decomposition in (1) is often expressed as fractions of UNC and the complement to one of the scaled Brier Score (BS) as a Brier Skill Score

$$BSS=1-BS/BS_{ref}=\left(REL\text{-}RES\right)/UNC\,. \tag{3}$$

This formula clearly emphasizes the trade-off between these two components with the aim of increasing resolution without neglecting reliability.

There is another decomposition by Yates (1982) deserving as much attention as it defines influential components of the expected quadratic score which are easier to understand:

$$\begin{aligned}\mathbb{E}[S\left(P,X\right)] &= \mathrm{Var}\left(X\right)-2\,\mathrm{Cov}\left(P,X\right)+\mathrm{Var}_X\left[\mathbb{E}_P\left(P\mid X\right)\right]\\ &\quad+\mathbb{E}_X\left[\mathrm{Var}_P\left(P\mid X\right)\right]+\left[\mathbb{E}\left(P\right)-\mathbb{E}\left(X\right)\right]^2.\end{aligned} \tag{4}$$

This formula stems from the basic expression $\mathbb{E}[S\left(P,X\right)]=\mathrm{Var}\left(P-X\right)+\left[\mathbb{E}\left(P\right)-\mathbb{E}\left(X\right)\right]^2$ with $\mathrm{Var}(P)=\mathrm{Var}_X\left[\mathbb{E}_P\left(P\mid X\right)\right]+\mathbb{E}_X\left[\mathrm{Var}_P\left(P\mid X\right)\right]$ and $\mathbb{E}\left(P\mid X\right)$ being the best predictor of P given X. As this predictor is a linear one: $\mathbb{E}\left(P\mid X\right)=a+bX$ with $a=\mathbb{E}\left(P\mid X=0\right)$, $b=\mathbb{E}\left(P\mid X=1\right)-\mathbb{E}\left(P\mid X=0\right)$, $\mathrm{Var}_X\left[\mathbb{E}_P\left(P\mid X\right)\right]=b^2\,\mathrm{Var}\left(X\right)$, $\mathrm{Cov}\left(P,X\right)=b\,\mathrm{Var}\left(X\right)$. These last two expressions highlight the key role of the regression coefficient b of P on X equal to the difference in expected value of forecasts pertaining to future positive outcomes from those out of negative ones.

Yates (1982) emphasized the different influence of the two components of $\mathrm{Var}(P)$. The first one, which is the between class variance $\mathrm{Var}_X\left[\mathbb{E}_P\left(P\mid X\right)\right]$, he qualified as “VarPmin”, is beneficial. The second one, which is the within class variance $\mathbb{E}_X\left[\mathrm{Var}_P\left(P\mid X\right)\right]$, called “ΔVarP” or “Scattered Variance” representing the lack of sharpness of the distributions of $P\mid X=x$, is detrimental. The last term in (4) measures marginal bias and was called “Calibration or Reliability-in-Large” (CIL or RIL). In short, this decomposition is written as: UNC- 2COV+ VarPmin+ ΔVarP+ RIL. Actually, Yates’ and the LB decompositions are closely related although not subject to the same interpretation, with REF=VarPmin+ΔVarP, DIS= VarPmin and CB2=UNC-2COV+VarPmin+RIL.

Table 1: Calibration-Refinement decomposition of Brier's score pertaining to Home Win, Draw and Away Win under the Poisson regression model.

RESULT	BRS	SKI (%)	B-TEST	UNC	MET	REL	RES
					INT	0.0035 (1.4)	0.0644 (26.2)
HWIN	0.1849	24.8	1.035	0.2458	QUA	0.0030 (1.2)	0.0639 (26.0)
			[0.309]		ISO	**0.0116 (4.7)**	**0.0725 (29.5)**
					INT	0.0010 (0.5)	0.0036 (1.9)
DRAW	0.1849	1.4	3.995	0.1875	QUA	0.0031 (1.7)	0.0058 (3.1)
			[0.045]		ISO	**0.0099 (5.3)**	**0.0125 (6.7)**
					INT	0.0047 (2.2)	0.0505 (23.4)
AWIN	0.1700	21.2	0.001	0.2158	QUA	0.0029 (1.3)	0.0487 (22.5)
			[0.975]		ISO	**0.0078 (3.6)**	**0.0537 (24.9)**

BRS=REL-RES+UNC with REL: Reliability, RES: Resolution, UNC: Uncertainty according to different binning procedures (INT: Interval; QUA: Quantile: ISO-Regression).and expressed both in absolute value and p.100 of uncertainty (UNC).

Skill (SKI) defined as SKI=(BRSref -BRS)/ BRSref where BRSref=UNC so that SKI=(RES-REL)/UNC

B-TEST: Brier-Score Test for departure of its expectation from that induced by the null hypothesis of perfect forecast calibration expressed with its corresponding statistic and P-value within brackets.

2.2 Estimation from data

Practically, verification takes place from a data sample made of pairs $\{(p_i, x_i), i=1,...,N\}$ of ex ante probabilistic forecasts p_i and ex post binary outcomes x_i. Most quantities introduced previously can be estimated by their regular moment estimators. The expected quadratic score $\mathbb{E}[S(P,X)]$ is traditionally estimated by the empirical score:

$$\bar{S}(\mathbf{p}) = N^{-1}\sum_{i=1}^{N} S(p_i, x_i). \tag{5}$$

For the CR decomposition, REL and RES require estimations of $\mathbb{E}(X \mid P)$. If the forecasts take K distinct values $\{p_k, k=1,...,K\}$ with n_k occurrences of binary outcomes X, then $\hat{\mathbb{E}}(X \mid P = p_k) = \bar{X}_k = X_{k+} / n_k$ with $X_{k+} = \sum_{i=1}^{N} I(p_i = p_k) X_i$ and $\bar{X} = \left(\sum_{i=1}^{N} X_i\right) / N$. In such cases, the Murphy (1973) decomposition is fully applicable without restrictions:

$$REL = N^{-1}\sum_{k=1}^{K} n_k (\bar{x}_k - p_k)^2,\ RES = N^{-1}\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2,\ UNC = \bar{x}(1-\bar{x}). \tag{6}$$

In fact, in many applications as in forecasting Football match results, we stay in-between discrete and continuous distributions, facing many distinct forecast values. In such cases, forecasts have to be distributed into intervals named bins $B_1,..B_d,..,B_D$ and averaged within bins i.e., letting $I_d = \{i : p_i \in B_d\}$, pairs $\{(p_d, x_d), d=1,...,D\}$ are computed as $p_d = n_d^{-1}\sum_{i \in I_d} p_i$, $\hat{x}_d = x_{d+} / n_d$ where $n_d = \# I_d$, $\hat{x}_{d+} = n_d^{-1}\sum_{i \in I_d} x_i$. To avoid inconsistencies in the CR decomposition, two extra components of within bin variance and covariance must be added to those in (6). We skip such complications by adopting a simple procedure as advocated by Siegert (2017). Letting as in (5) $\bar{S}(\hat{\mathbf{x}}) = N^{-1}\sum_{i=1}^{N} S(f_i, \hat{x}_i)$ and $\bar{S}(\bar{x}\mathbf{1}_N) = N^{-1}\sum_{i=1}^{N} S(f_i, \bar{x})$ with $\mathbf{1}_N$, the unit vector of size N, the components of the mean score $\bar{S}(\mathbf{p})$ reduce to

$$REL = \bar{S}(\mathbf{p}) - \bar{S}(\hat{\mathbf{x}}),\ RES = \bar{S}(\bar{x}\mathbf{1}_N) - \bar{S}(\hat{\mathbf{x}}),\ UNC = \bar{S}(\bar{x}\mathbf{1}_N). \tag{7}$$

This decomposition automatically satisfies the equality $\overline{S}(\mathbf{p}) = REL - RES + UNC$, and is equivalent to the original Murphy decomposition in the case of distinct discrete forecasts. It also ensures that i) resolution is nil when $\mathbf{p}$ is perfectly calibrated ($\mathbf{p} = \hat{\mathbf{x}}$), and ii) the constant climatological forecast $\mathbf{p} = \overline{x}\mathbf{1}_N$ is the only forecast satisfying RES=REL=0. Finally, it is potentially applicable to other proper probabilistic scoring rules, as the ignorance score $L(P,X) = -X\log(P) - (1-X)\log(1-P)$ (Dawid, 1986; Bröcker, 2012). Moreover, the statistic $2N\left[\overline{L}(\mathbf{p}) - \overline{L}(\hat{q}_p)\right]$ is an analog of the log-likelihood ratio statistic for a perfectly reliable forecast having an asymptotic Chi-square distribution with degrees of freedom equal to the number of parameters specifying the model for $q_p = \Pr(X = 1 \mid P = p)$.

Different binning techniques are available such as fixed threshold intervals and fixed quantile intervals with potential optimization of their number (Bröcker, 2012; Gweon and Yu, 2019). A promising one relies on the non-parametric isotonic regression implemented via the pool-adjacent-violators (PAV) algorithm with optimality properties (Dimitriadis et al., 2021).

It also provides a reliability diagram featuring graphically the CR decomposition of the joint distribution $[P,X] = [P][X \mid P]$ of binned data by the marginal (refinement) distribution of forecasts and plots of (re) calibrated probabilities $\hat{x}_d$ against automatically binned forecasts p_d.

Table 2: Calibration analysis via fitting a logistic model of the probability of Homewin, Draw and Awaywin (AWIN) on the logit of its probabilistic forecast under a Poisson regression model (POI)

Category	Criterion	Estimation	SE	T-Statistics	DF	P-value
Homewin	intercept	-0.259	0.119	4.700	1	0.030
	slope	1.113	0.129	0.765	1	0.382
	D0 vs D1	423.085 vs 417.489		5.596	2	0.061
Draw	intercept	0.153	0.466	0.108	1	0.742
	slope	0.932	0.346	0.039	1	0.843
	D0 vs D1	426.981 vs 422.981		4.000	2	0.135
Awaywin	intercept	0.076	0.149	0.261	1	0.610
	slope	1.053	0.134	0.156	1	0.693
	D0 vs D1	389.458 vs 389.176		0.282	2	0.870

Intercept (α) and slope (β) of the logit regression model with their estimation and standard error (SE). Deviance D(k)=-2L(k) where L(k) is the loglikelihood of the null model (0: α =0; β=1) vs the unspecified parameter model (1: $\alpha \neq 0$; $\beta \neq 0$); T-statistics: Wald for intercept=0 and slope=1; Deviance differences ΔD=D0-D1 and their corresponding degrees of freedom (DF) and P-values

To that respect, another way to assess Reliability via the conditional distribution $[X \mid P]$ of outcomes X given P is through a regression model, but in the framework of logistic instead of linear regression chosen by Reade et al. (2021). Following Cox (1958), the model relating X to P is written via a logit linear predictor: $\operatorname{logit}\left[\Pr(X_i = 1)\right] = \alpha + \beta \operatorname{logit}(p_i)$ with $\alpha = 0$ and $\beta = 1$ for perfect reliability and typical patterns of reliability diagrams with i) $(\alpha > 0, \beta = 1)$ for concave under-forecasting profiles, ii) $(\alpha < 0, \beta = 1)$ for convex over-forecasting profiles as well as iii) $(\alpha = 0, \beta > 1)$ for sigmoid, and iv) $(\alpha = 0, \beta < 1)$ for inverse-sigmoid profiles. Statistical tests are available (Wald and likelihood ratio tests) for challenging the different hypotheses about such patterns.

3 Application

The purpose of this illustration is to assess the performance of probability forecasts of outcomes of the UEFA champions league (the so called C1) matches played during the group stage (GS). Four seasons were considered from 2017 to 2020. Forecast is based on a simple log linear Poisson regression model applied to score lines with intercept, home effect and two time-dependent ELO team covariates. The models was fitted to ex ante data, namely score lines of all the matches played during the 3 previous seasons e.g., 2017, 2018 and 2019 as training sample to forecasts of the 96 GS matches of 2020; the same applies to forecasts of the 2019 GS based on 2016, 2017 and 2018 seasons and so on. Inference about parameters of the Poisson loglinear model is based on posterior distributions and probability forecasts are obtained as expectations of predictive distributions. Computations are carried out via the Win/OpenBUGS software. As top reference, we considered the classical Bookmaker Odds (ODD) as 3-Way Odds implied Probabilities with probabilities derived as $p_{m,j} = o_{m,j}^{-1} / \sum_{k=1}^{3} o_{m,k}^{-1}$ where $o_{m,j}$ is the betting odd for $j = 1,2,3$ (WDL) edited by OddsPortal (here an average of 10 to 12 odds from well-known betting companies).

Table 3: Characteristics of conditional distributions of probability forecasts given the outcomes under two Forecasting procedures: Poisson regression (POI) and Odds Probabilities (ODD)

Method		Home Win		Draw		Away Win	
		POI	ODD	POI	ODD	POI	ODD
Sample sizes		217-167		288-96		263-121	
Mean %	X=0	37.88	35.01	20.22	20.97	24.30	23.24
	X=1	63.08	62.73	22.34	23.89	44.84	48.62
	Dif 1-0	24.20	27.71	2.02	2.93	20.54	25.38
Wilcoxon	Z	9.93	10.76	3.59	3.55	9.09	10.13
	P-val	*<0.0001*	*<0.0001*	*0.0002*	*0.0002*	*<0.0001*	*<0.0001*
KS	D	*0.473*	*0.511*	*0.236*	*0.236*	*0.447*	*0.521*
	P-val	*<0.0001*	*<0.0001*	*0.0007*	*0.0007*	*<0.0001*	*<0.0001*
C-statistic	Estimation	0.795	0.820	0.622	0.624	0.789	0.820

Sample sizes of forecasts having X=0 vs X=1 respectively; Z: Normal approximation of the Wilcoxon-test with one sided P value; KS: Kolmogorov-Smirnoff two sample test on Max [F(X=0)-F(X=1)] C-statistic: Harrell's concordance index varying from 0.5 (no discrimination) to 1 (perfect discrimination) equal to AUC (area under the ROC curve)

3.1 CR decomposition of Brier's score

Due to the large number of unique probability profiles (377 among N=384 matches), forecasts were binned in three different ways: i) Fixed threshold intervals: D=10 from 0.0 to 1.0 for Home Win; D=5 with bounds at 0.10,0.15,0.20,0.25 and 0.35 for Draw and D=8 for Away Win with the first 7 bins equally spaced from 0 to 0.7 and the last one from 0.7; ii) Quantile thresholds intervals: deciles for Home Win and Away Win and quintiles for Draws iii) Bins automatically determined by the pool-adjacent-violators (PAV) algorithm used to set up the nonparametric isotonic regression deployed to estimate the conditional $q_p = \Pr\left(X = 1 \mid P = p\right)$ outcome probabilities by minimizing the regression MSE with respect to D:

$$\sum_{d=1}^{D}\sum_{i=1}^{N} I\left(p_i \in \left[b_d, b_{d+1}\right]\right)\left(q_d - p_i\right)^2 , \tag{8}$$

under the constraints of isotonicity (q_d estimation is a non-decreasing function of the original p_i 's). Results are displayed in Table 1, for Home Win, Draw and Away Win categories considered separately in terms of absolute values of Mean Brier score and its components (REL, RES, ACC) and Skill. Lack of reliability turns out to be small (lower than 5.5% of UNC) with miscalibration estimated a little bit higher under iso-regression. These values are supported by statistics and P values of Brier's score tests of departure from zero miscalibration (Spiegelhalter, 1986; Sellier-Moiseiwitch and David, 1993).

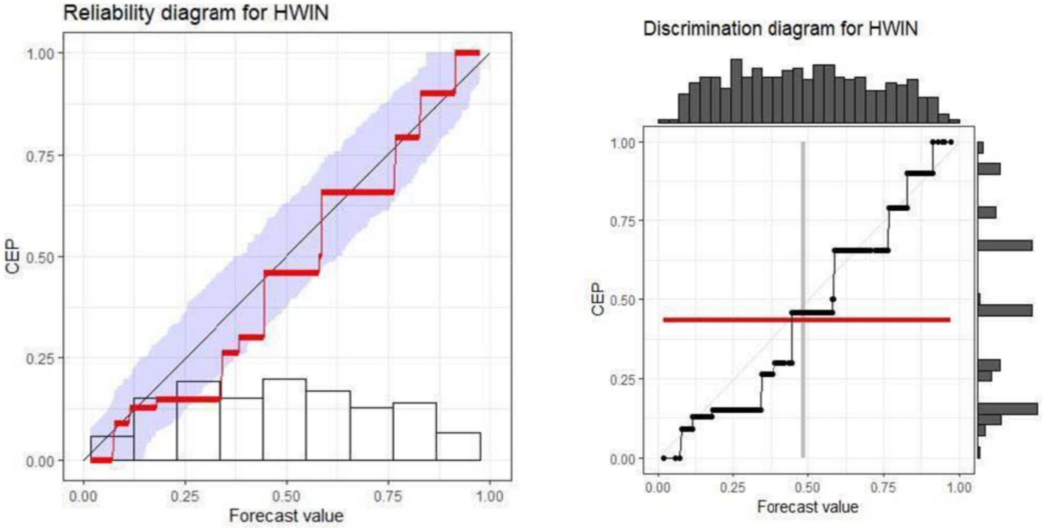


Figure 1: Reliability Diagrams for Home Win Probability Forecasts with plots of the Conditional Probability Events (CEP) against the Forecast Probability Values via Iso Regression. On the left : Reliability with Point 95% Consistency Bands. On the right: Discrimation diagrams with marginal distributions of forecasts and calibrated values.

Skill values are also appreciable for Home win (25 to 30%) and Away win (20 to 30%) with a little advantage of Odds vs Poisson of around 5. On the opposite, skill remains quite poor for Draw: 1.4 to 3.0% for Poisson and Odds respectively. Corresponding reliability diagrams were produced with an example shown for Home Win under Poisson and forecasts on Fig 1. There is some evidence of over-forecasting for Home Win practically all over the forecast range for Poisson, and at least, below p<0.60 for Odds (not shown) resulting in under-forecasted Draws. These conclusions are confirmed by the results (table 2) of the logistic regression with intercept and slope calibration coefficients showing clearly over-forecasting of Home Wins $(\hat{\alpha} = -0.26, \hat{\beta} = 1.11)$ with Pvalues equal to 0.03 and 0.38 respectively while Away Wins are very well calibrated $(\hat{\alpha} = 0.076, \hat{\beta} = 1.053)$.

3.2 LB and Yates' decompositions of Brier's score

Likelihood-base as well as Yates' decomposition of Brier Score rely largely on the concept of discrimination between the conditional distributions of forecasts with positive outcomes vs forecast with negative outcomes. Characteristics of these two distributions are given in Table 3 and Fig 2. Differences between the means of these two conditional distributions are much more marked for Home Win and Away Win than for Draw. Again, these differences are more pronounced with Odds than Poisson as also reported by the Harrell c-statistics around 0.8 for Home Win and Away Win, and only 0.6 for Draw. Graphically, the boxplots confirmed this situation showing a clear separation of the two distributions for Home Win and Away Win and a tiny one for Draw.

Detailed accounts of Yates' and LB decompositions are shown on Table 4. In short, what emerges from them lies in the large role and weight given to the covariance component: 41 to 48% for Home Win and Away Win under Poisson and 50 to 55% with Odds with nevertheless a non-negligible part devoted to "noise" variance of forecasts (15 to 18%). The same picture applies to Draw but with much more tiny components, especially discrimination and covariance.

Table 4: Yates's and LB decompositions of Brier's score pertaining to Home Win, Draw and Away Win under the Poisson regression model (POI)

Factors	Home Win		Draw		Away Win		All	
	value	%	value	%	value	%	value	%
UNC	0.2458	100.0	0.1875	100.0	0.2158	100.0	0.6490	100.0
(-2)COV	-0.1190	-48.4	-0.0076	-4.0	-0.0886	-41.1	-0.2150	-29.5
VPB	0.0144	5.8	0.0001	0.0	0.0091	4.2	0.0236	3.6
VPW	0.0413	16.8	0.0031	1.7	0.0336	15.6	0.0780	12.0
RIL	0.0024	1.0	0.0017	0.9	0.0000	0.0	0.0042	0.6
REF	0.0556	22.6	0.0032	1.7	0.0427	19.8	0.1016	15.6
-DIS	-0.0144	-5.8	-0.0001	-0.0	-0.0091	-4.2	-0.0236	-3.6
CB2	0.1436	58.4	0.1817	96.9	0.1363	63.2	0.5396	83.
BRS	0.1849	75.2	0.1848	98.6	0.1700	78.8	0.5397	83.1

Yates's decomposition into 5 components as follows BRS=UNC-2COV+VPB+VPW+RIL with UNC: Uncertainty, COV: Covariance between forecast and outcome, VPB: Variance among means of probability forecasts with outcome=1 and outcome=0, VPW : Average of Within groups variance and RIL marginal bias squared between the two groups, according to forecasting procedures (POI and ODD models) . Likelihood base decomposition into 3 components BRS=REF-DIS+CB2 with REF: Refinement or Sharpness of forecast variance, DIS=Discrimination same as VPB and CB2: Type 2 bias equal to VPW-2COV+RIL.

4 Discussion

This presentation was deliberately restricted to the most popular (strictly) proper scoring rules as this properness property is a cornerstone of decision theory based on minimizing expected loss (or maximizing utility) (Bernardo, 1979; Gneiting and Raftery, 2007). They provide an incentive for ex ante honesty and reward ex post accuracy. Little was said about verifying probability forecasts for multiple categories taken simultaneously. Probability scoring rules are extended easily to that situation as shown in Table 4 for Yates' decomposition. Unhappily, such an extension is not straightforward for the CR decomposition. One reason for that lies on how to define bins. Procedures have been proposed to that respect by Broecker (2012) based on some functions of the probability vector for the J multiple categories of interest. A simple way to handle the multiclass setting is by treating the problem as J one-versus-all binary events via e.g., a logistic-type regression with standard normalization of outcome probabilities. More generally, such forecasting verification methods already gained much attention in other fields especially in machine learning especially due to miscalibration of neural networks and its applications to health and medicine (Guo et al, 2017).

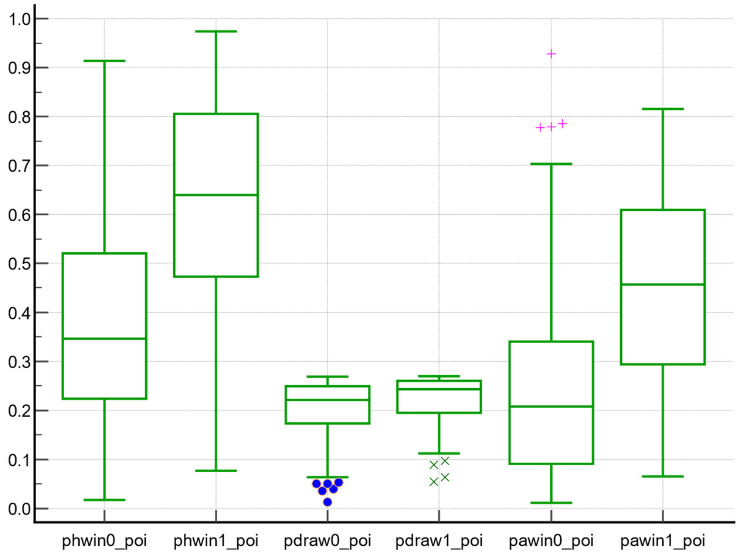


Figure 2: Box-plots of the conditional distributions of probabilistic forecasts of binary events (Home Win, Draw and Away Win) given the observed outcomes (X=0) and (X=1) and according to the Poisson loglinear regression (poi)

Although much of the theory and applications originated from meteorological literature (Winkler and Murphy, 1979; Jolliffe and Stephenson, 2003; Wilks, 2011), there had been a few attempts to apply some of these analytical procedures to football match results, especially in the EPL (Forrest and Simmons, 2000; Selliti Rangel, 2018; Wheatcroft, 2019; Read et al, 2021) but not enough. This area would benefit from a more systematic utilization. Regarding our application to the UEFA Champions League, it turns out that forecasts of Away Wins is both well calibrated and refined with good discrimination properties. The same trend can be seen on Home Wins, but with a trade-off between that over-forecasted category and under-forecasted draws with little discrimination of draw forecasts. It looks as if this category stands apart from the two others.

References

Bradley, A. A., T. Hashino, and S. S. Schwartz. (2003), ”Distributions oriented verification of probability forecasts for small data samples”, *Weather Forecasting* ,18, 903–917.

Brier, G. (1950), “Verification of forecasts expressed in terms of probability”, *Monthly Weather Review*, 78, 1-3.

Bröcker, J. (2012) Probability forecasts. In *Forecast Verification: A Practitionner's Guide in Atmospheric Science,* Second Edition. Edited by l. T. Jolliffe, and D.B. Stephenson. 2012 John Wiley & Sons, pp 119-139.

Cox, D.R. (1958), "Two Further Applications of a Model for Binary Regression", *Biometrika*, 45, 562-565.

Dawid, A. P. (1986), "Probability Forecasting", *Encyclopedia of Statistical Science*, 7, 210-218.

Dimitriadis, T. Gneiting, T. and Jordan, A.I. (2021) *Stable reliability diagrams for probabilistic classifiers*. Proceedings of the National Academy of Sciences of the United States of America **118** (8) DOI: 10.1073/pnas.2016191118

Forrest, D., and R. Simmons. (2000), "Forecasting Sport: The Behaviour and Performance of Football Tipsters", *International Journal of Forecasting*, 16, 317-331.

Gneiting, T., and A.E. Raftery. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", *Journal of the American Statistical Association*, 102, 359–378.

Guo, C., Pleiss, G., and K.Q., Weinberger.(2017), "On Calibration of Modern Neural Networks", arXiv: 1706.04599v2

Gweon, H., and H. Yu. (2019), "How reliable is your reliability diagram? " *Pattern Recognition Letters*, 125, 687-693.

Jolliffe, I., and D. Stephenson. (2003), "Forecast Verification: A Practitioner's Guide in Atmospheric Science", John Wiley-Blackwell

Murphy, A.H. (1973), "A new vector partition of the probability score", *Journal of Applied Meteorology,* 12, 595–600.

Murphy, A. H., and R.L., Winkler. (1987), "A general framework for forecast verification". *Monthly Weather Review* ,155, 1330-1338.

Reade, J. J., Singleton C., and A. Brown. A. (2020) Evaluating Strange Forecasts: The Curious Case of Football Match Scorelines. *Scottish Journal of Political Economy*, 68, 261-285.

Scarf, P., and J. Selliti Rangel Jr. (2016), "Models for outcomes of soccer matches". In *Handbook of Statistical Methods and Analyses in Sports*, Eds by J. Albert, M.E., Glickman, T. B, Swartz ,and R.H. Koning, 341-354. CRC Press, Chapman pp. 341-354.

Seillier-Moiseiwitsch, F., and Dawid, A.P. (1993), "On Testing the Validity of Sequential Probability Forecasts", *Journal of the American Statistical Association*, 88, 55-359.

Selliti Rangel Jr, J. (2018), "Estimation and Forecasting Team Strength Dynamics in Football: Investigation into Structural Breaks", PhD thesis, University of Sallford, UK.

Siegert, S. (2017), "Simplifying and generalising Murphy's Brier score decomposition." *Quarterly Journal of the Royal Meteorological Society*, 143, 1178–1183.

Spiegelhalter, D. J. (1986), "Probabilistic prediction in patient management and clinical trials", *Statistics in Medicine* **5**, 421–433.

Wheatcroft, E. (2019), "Interpreting the skill score form of forecast performance metrics", *International Journal of Forecasting*, 35, 573-579.

Wilks, D. S. (2011), "Statistical Methods in the Atmospheric Sciences", 3rd Edition Academic Press, Oxford.

Winkler, R.L., and A.H. Murphy. (1979), "The Use of Probabilities in Forecasts of Maximum and Minimum Temperatures", *The Meteorological Magazine*, 108, 317-329.

Yates, J.F. (1982), "External correspondence: Decompositions of the mean probability score", *Organizational Behavior and Human Performance,* 30, 132–156.