FINAL REPORT

A Systematic Approach to Strategic Football Team Insights & Player Recruitment Analysis

Supervisor:
Dr. Anthony BROOMS

Student:
Christian GILSON

Supervisor:
Dr. Swati CHANDNA

Word Count:
13,468

September 1, 2021

Declaration

This project is submitted under University of London regulations as part of the examination requirements for the MSc degree in Applied Statistics and Computational Data Analytics. Any quotation or excerpt from the published or unpublished work of other persons is explicitly indicated and in each such instance a full reference of the source of such work is given. I have read and understood the requirements of the Birkbeck College Examinations Instructions to Candidates, including the relevant University of London regulations on Examination Tests and in accordance with those requirements submit this work as my own.

All data preparation and manipulation was completed using standard libraries in Python such as pandas and numpy. Most plotting was carried out using standard R and Python plotting libraries (ggplot and matplotlib, respectively) as well as utilising specialist football pitch plotting library mplsoccer. Proportional area radars utilise code with permission from the original author, Giacomo Marchesi, a fellow collaborator from Uppsala University's Mathematical Modelling of Football module from Uppsala's MSc Machine Learning program.

All code written to produce the work in this project has been made publicly available on my GitHub page at github.com/christiangilson, along with a separate package intended for use by practitioners to utilise the techniques developed throughout the work called xGils.

Abstract

Finding a competitive edge off the pitch is proving to be an increasingly existential challenge for professional football clubs. The same level of rigour from quantitative finance is now being applied to quantitative football to inform recruitment and matchday strategy using a systematic, data-driven approach to analyse structured data describing the blow by blow of on-field actions. The work presented here utilises every pass, cross, dribble, and shot attempted in the English Premier League from 2017/18 to 2020/21 to provide analyses, software, and applications in three critical problem areas faced by football clubs: (1) which data to use? (2) how to model event data to value individual player actions in a production environment and (3) how to apply action values to systematically identify appropriate recruitment targets and quantify opponent strengths and weaknesses.

The work first introduces a comprehensive data quality analysis that finds Opta to be by far the superior provider of on-the-ball events data over Wyscout. Secondly, the modelling software to value player actions by treating possession sequences as Markov chains has been made publicly available through the xGils Python package. xGils enables industry practitioners to easily adopt the introduced Bayesian Expected Threat value action framework with three essential features for production-ready use: (1) a vectorised Beta-Binomial updating mechanism to keep value surfaces up-to-date as football evolves and support ongoing data feeds (2) a seamless method to integrate event data from any data provider and (3) the ability to encode domain expertise via synthetic data as a Bayesian prior. Five applications result that neatly build on one another to produce radars illustrating intuitive key performance indicators for cross-player comparison for recruitment purposes, as well as high-dimensional representations that characterise a team's playing style. Finally, a significant innovation allows the indirect measurement of a team's defensive capabilities, solving for on-the-ball event data's contextual blindspot, enabling our systematic scouting approach to scale beyond what is humanly possible in both attack and defence.

I. INTRODUCTION

In 2002 the Oakland A's baseball team famously applied data science to player action data to produce player KPIs in a way that would forever change sports team recruitment [13]. One particular metric, wins above replacement, allowed the A's to quantitatively arbitrage low-cost, traditionally unconventional talent to unearth position-specific diamonds in the rough.

Applying "Moneyball" strategies to football recruitment has proven a much more complex challenge. Barcelona's net spend of €450m since 2014 — bringing one of the world's biggest clubs to the brink of bankruptcy and without a single recruit to truly emerge as a first team leader — is three times that of 2018/19 Champions League and 2019/20 Premier League winners Liverpool, highlighting the existential challenge of finding a competitive edge off the pitch. What makes football a different animal is that it's an invasion game featuring complex inter-player dynamics with valuable, game-changing contributions happening off the ball. As a result, it's tricky to tease out an individual's impact on team performance and even trickier to know how a new recruit will integrate with their new club's playing style.

Manchester City's recent off-the-pitch hire of Laurie Shaw to head up their data science and analytics initiatives made the front pages of Bloomberg and BBC News [5, 19], with the headline of the latter: "*Data experts are becoming football's best signings*". This highlights the importance of a world-class football club's long-term strategy to invest in data science expertise, infrastructure, and decision-making to provide a competitive edge off the pitch and enable future success on it.

Rather than attempting to come up with one-model-to-rule-them-all that'd solve all of Barcelona's recruitment woes, this project built three core models with which to calculate complementary player metrics that could also be aggregated at the team level to explore football club strengths and weaknesses in both attack and defence. The functional aims of the three central models were to:

1. Quantify the value, in units of goals, of individual player actions that move the ball around the pitch during a possession sequence — like passes, crosses, and dribbles — independent of whether or not the team scores at the end of the possession. This is the output of the Expected Threat value action model (**xT**, [24, 33]).
2. Explain shot chance quality via the probability of a typical player scoring a shot conditioned on the situation that player finds themselves in at the point of shooting. This is the output of the Expected Goals model (**xG**, [1, 23, 26, 28]).
3. Produce points-based ratings that can be ranked in the probable order of a player's relative strength for actions where two players contest a duel, like an aerial header, or an attacking dribbler trying to beat an opposing defender, where there is a definitive winner and loser. This is the Elo ranking system [6, 10].

Metrics from the three models were combined to produce a suite of intuitive, differentiated, and rankable player KPIs that could form the basis of systematic recruitment initiatives, such as:

- Identifying high and poor performing players within a squad;
- Identifying transfer targets in offensive and defensive positions to enable a club to mitigate key asset losses and strengthen positions of vulnerability.

In short, the project sought to use new data and apply novel modelling methods in a practical way to solve existing, high-value problems at football clubs, providing intuitive insights.

II. STRUCTURE OF THE WORK

We begin with a detailed summary of the data in section III: evaluating the availability, features, quality, and limitations of the data being analysed. Section IV introduces value action frameworks, covering the historical development and current state of the art in value action modelling. The construction of the three core models at the heart of the project is then described in detail. Section V provides five applications exploring concrete industry use cases relevant to recruitment and matchday strategy to ground and focus the techniques developed. Section VI provides a brief closing summary of the project's value and potential future work.

To enable reproducible research, the modelling code used throughout the project has been wrapped up in a Python library called **xGils** that's publicly accessible on GitHub via: github.com/christiangilson. In addition, all analyses and plotting code (written in Python and R notebooks) can be found in a second publicly accessible GitHub repository called **MSc-Applied-Statistics-Project-Code**.

The work is intended to be accessible to non-technical practitioners and enthusiasts with a primary interest in the final applications; semi-technical analysts with a working knowledge of data fundamentals but a preference to use tools and software to produce their analyses; and end-to-end modellers who may wish to understand the finer details of how the models were built and possibly code them from scratch. Non-technical readers may wish to skim section III before reading section V, whereas analysts would benefit from reading section III and section V in full, and skimming section IV. The tone when exploring the applications in section V is purposefully less formal to be more engaging for practitioners without missing any essential detail required by modellers to reproduce the work.

III. DATA

This project used on-the-ball football events data. Events data provide a blow-by-blow account of on-the-ball actions throughout a football match collected by analysts watching video footage, using software to classify actions into an events taxonomy and attribute contextual metadata to each action: including player and team identifiers, the timestamp, start and end positional coordinates, as well as action-specific features.

There are a handful of data vendors in this space, namely Opta, Statsbomb, and Wyscout [21, 27, 35], each with their own proprietary data collection software, methodology, and event taxonomy structure. The industry standard is to have two analysts per match, one focusing on manual data collection for each team. After the manual data collection stage, the data will go through an automated quality assurance process to sanity check and systematically fix the manually collected data (e.g. searching for impossible combinations of events and adjusting duel events featuring players from both teams to have consistent positional and timestamp data). There's often a final manual set of sanity checks, where potential issues are highlighted algorithmically, that require human judgement as to what fix, if any, is required [4].

Until recently, events data was prohibitively expensive for academic research. In November 2019 [22], Wyscout released a publicly available dataset containing the 1,826 matches played during the 2017/18 season across five European domestic leagues: English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue 1. Wyscout also released event data for

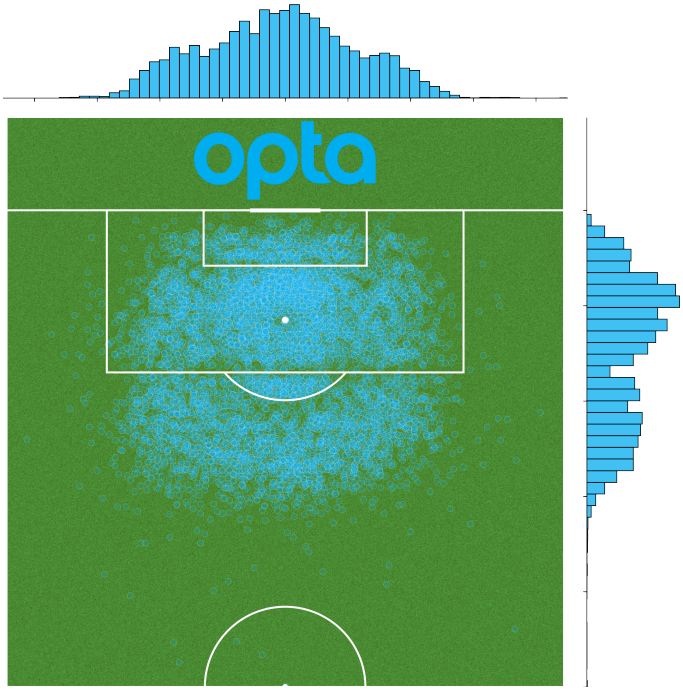

FIG. 1: A joint plot of *Opta shots* and their positional distributions for the English Premier League season 2017/18, whereby Opta data contains 8% more shots than Wyscout for this overlapping season.

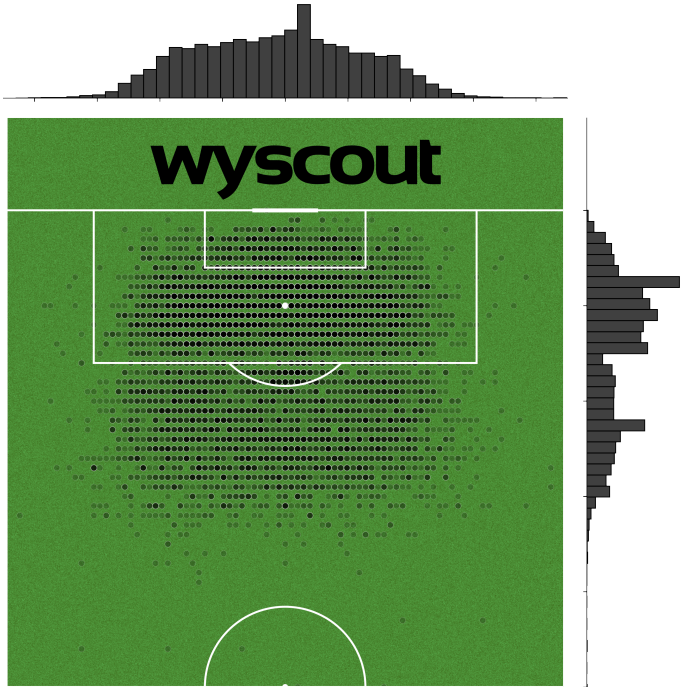

FIG. 2: A joint plot of *Wyscout shots* and their positional distributions for the English Premier League season 2017/18, the same season as in FIG. 1.

the 2018 FIFA World Cup (64 international matches) and the 2016 UEFA European Football Championship (51 international matches).

An alternative method of acquiring football events data for the purpose of quantitative research is through the legal scraping of web-based football analysis applications (in full accordance with the User-agent conditions within the robots.txt file of the application).

This project utilised the publicly available Wyscout dataset, offering geographical breadth for a single season, in combination with a webscraped (by the author) Opta dataset of 1,520 matches from the English Premier League for the 2017/18, 2018/19, 2019/20, and 2020/21 seasons as well as 250 matches from the 2017/18 and 2018/19 Champions League competitions, offering temporal breadth across different seasons for the same domestic league (enabling promising recruitment targets to be tracked through time) and a mixing of club teams from different nations in the Champions League (providing a direct window to compare relative league strength).

Opta Vs Wyscout Data Quality

FIG. 1 and 2 show the positions and and distributions of shots taken in the 2017/18 Premier League season from Opta and Wyscout, respectively. This was our one season of overlapping data between the two data vendors.

It was immediately apparent that Wyscout positional data is collected with less precision than Opta, representing the pitch as a 101x101 integer grid (providing 10,201 unique positions on the pitch), whereas Opta offers an additional order of magnitude of positional granularity in both x and y directions (providing 1,002,001 unique positions on the pitch). Opta's x-direction histogram is smoother and more bell-shaped around the penalty spot than Wyscout's flatter distribution as a result, with Wyscout's also producing a questionable discontinuity to the right of the penalty spot. It's interesting to observe a dip in the y-direction distribution at the edge of the penalty box in *both* the Opta and Wyscout data, highlighting a potential bias in human analysts, anchored by the pitch markings, to scatter actions either inside or outside of the box and avoid a visual boundary.

Opta's data contained 8% more shots than Wyscout's for our season of overlap (9,113 Vs 8,450), and 12% more actions over all event types (667,150 Vs 595,119) highlighting significant differences in the data collection *methodologies* between the two datasets. The consensus opinion held by industry practitioners — like Luton Town's Head of Recruitment Analysis, Jay Socik [25] — is that Opta provides substantially higher quality and more complete data than Wyscout, suggesting there are also material differences in the *quality assurance* processes, that follow downstream from collection, between the two providers.

A powerful way to test both the integrity and self-consistency of identifiers and taxonomies (player IDs, team IDs, and event categories in this project) *and* the accuracy and completeness for each of the two datasets in isolation is to *map* the datasets together at a granular level and compare using like-with-like summary statistics and sanity checks. This is the standard practice implemented at quantitative hedge funds when evaluating new datasets, and here we brought the rigour from quant finance to quant football. The point is to identify systematic differences between the two datasets (rather than just finding one-off errors that are present in virtually every dataset), understanding more and more about dataset features and idiosyncrasies as differences are manually investigated, to inform the decision of how best to use one or both datasets for the project being undertaken.

The mapping procedure to link the Opta and Wyscout datasets together consisted of the following three main stages:

1. Mapping the 20 2017/18 Premier League Opta team identifiers to Wyscout team identifiers *by hand*. E.g. Manchester

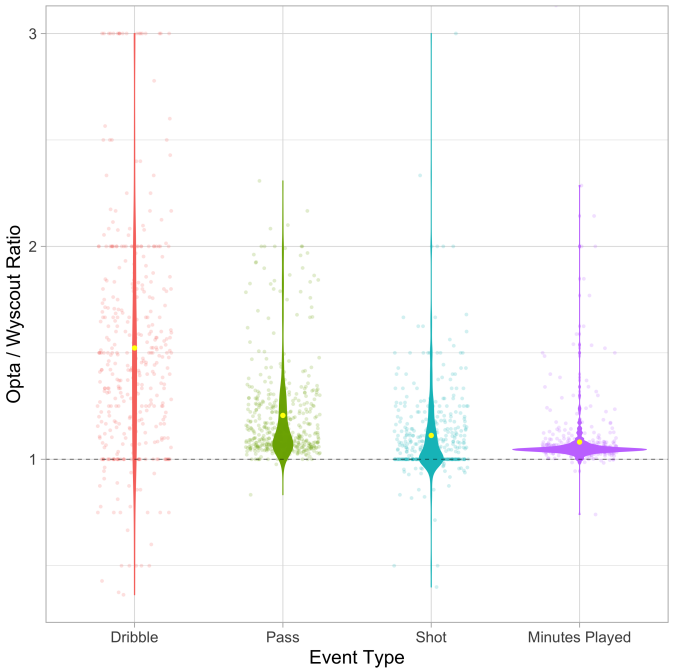

FIG. 3: A violin plot displaying distributions of Opta:Wyscout ratios for metric counts — dribbles, passes, shots, and minutes played — per player. Ratios over 1 illustrate Opta's data having greater counts for a given metric than Wyscout, for a particular player. Yellow dots represent ratio means.

United is TeamID 1 in Opta and TeamID 1611 in Wyscout.

2. Mapping 500 player names from Opta to Wyscout within the same team *computationally* using the Jaro Winkler string match algorithm [7, 11, 12, 34].
3. Manually normalising the Opta and Wyscout event taxonomies for event types pertinent to the project — dribbles, passes (including crosses), and shots — enabling counts of those actions to be compared for a particular player between the two datasets.

The computational Jaro Winkler player mapping step produced excellent results. 96% of the Opta player reference data was mapped accurately, requiring manual corrections for only a handful of mapping errors. The likelihood of an error was reduced by pre-filtering the possible Wyscout player match candidates for a given Opta name using the team mappings, therefore reducing the candidate space from 500 possible name matches to roughly 20, the size of a club's squad. TABLE I highlights some of the trickiest to map players, and illustrates the suitability of the Jaro Winkler algorithm in this problem setting.

It's common for Latin American players such as Liverpool's Adrián to identify themselves (and the backs of their shirts) by their first name and Jaro Winkler's scoring complements this behaviour: weighting string matches more strongly towards the start of the string and attributing less weight to common suffixes like "del Castillo", which is also shared by Manchester City's Sergio Agüero. This meant the algorithm coped well in instances where Opta's naming conventions were more closely aligned to that of how the player identifies themselves on the pitch, whilst Wyscout often referenced the complete name.

After mapping team and player identifiers and normalising and aggregating event counts for passes, dribbles, and shots, the Opta and Wyscout datasets were linked together and ratios of Opta counts to Wyscout counts calculated per player, as well as the ratio of the number of minutes played by that player, for the 2017/18 Premier League season. The number of minutes played is an essential quantity to measure accurately, since it's used as the denominator to produce player metrics on a per 90 minutes played basis, normalising for injuries, mid-season transfers, a manager's choice of the starting XI and strategic use of substitutions, as well as the seemingly random amount of extra time allotted by the referee at the end of each half (if we exclude the Sir Alex Ferguson era) and myriad other reasons. Virtually no two players play the same number of minutes per season. FIG. 3 shows a violin plot displaying the distributions of those Opta:Wyscout ratios, where each point represents a ratio for a single player for the given metric. For all four metrics, the Opta counts are higher than Wyscout counts on average (with the distribution means represented by yellow dots). Dribble ratios provide the largest variance, stemming from a poorly defined and poorly designed Wyscout "duel" taxonomy branch. A user of Wyscout data is required to combine multiple "tags" to define an event as a dribble, where the two main configurations either produce far too many or too few dribbles, as the former includes player actions who take a few touches of the ball before passing or shooting, and the latter includes a subset of dribbles where the attacking player is specifically trying to beat an opposing player. This data quality analysis chose the lesser of two evils when synthesising Wyscout dribbles, favouring precision over recall and picking the stricter definition. True dribble actions within the Wyscout dataset will have undoubtedly been lost, causing the large size and spread of Opta:Wyscout dribble count ratios. Critically however for this project, where the quality of the analysis is dependant on the accuracy of counts of events that move the ball around the pitch, Opta models dribbles cleanly within it's "Attack" event taxonomy, shown in TABLE II, as successful or failed, and their definition of what a dribble is has undergone far more scrutiny than that of Wyscout's, as Opta's dribble statistics are used by the likes of Sky Sports and broadcast to billions.

Distributions for pass and shot event types in FIG. 3 tell a similar story: that of Wyscout systematically missing key actions within a match. One of the main systematic causes of Wyscout failing to record shots, after manually reviewing video footage where Opta had recorded a shot and Wyscout had not, are cases where the shot is almost immediately blocked by a defender. This means Wyscout uses forward-looking information to define something that happened in the past, a data collection practice in severe tension with a central philosophy to this work. *History should not depend on the future.* The data being analysed and models being built in this project, to describe the past and probabilistically predict the future, will not depend on a deterministic kick of the ball that happens in that future.

Lastly, the minutes played ratio distribution is sharply peaked with a mean ratio of 1.08, meaning players in Opta's dataset play 8% more minutes on average than a Wyscout player. We immediately see the reason for this in FIG. 4 in two histograms, one per vendor, for the number of minutes played per player per match. One expects a quad-modal distribution: a symmetric pair of modes, roughly mirrored at the 45 minute mark, representing players being substituted in and out throughout the match, a peaked mode around 45 minutes representing half-time substitutions, as well as a more prominent mode after 90 minutes, representing the majority of players that play a full match including injury time (and as of the 2019/20 season, VAR time [31]). This is precisely what we see with the Opta data, however, with the Wyscout data, the minutes played

Opta Player Name	Wyscout Player Name	Jaro Winkler Similarity
Dele Alli	Bamidele Alli	0.79
Ramiro Funes Mori	José Ramiro Funes Mori	0.79
David Silva	David Josué Jiménez Silva	0.81
Pedro	Pedro Eliezer Rodríguez Ledesma	0.83
Sergio Agüero	Sergio Leonel Agüero del Castillo	0.83
Adrián	Adrián San Miguel del Castillo	0.84
Emerson	Emerson Palmieri dos Santos	0.85
Gareth Barry	Gareth Barry	1.00

TABLE I: Selection of the lowest confidence Jaro Winkler similarity player name match results (above a set threshold) used to map players between Opta and Wyscout datasets. (Gareth Barry example is used to illustrate the Jaro Winkler similarity score of a perfect match.)

Attack Event	% of Attack Events	% of All Events
Pass	66.99	43.72
Failed Pass	18.66	12.18
Bad Touch	2.50	1.63
Fouled	1.76	1.15
Lost Possession	1.71	1.11
Lost Aerial Duel	1.69	1.10
Dribble	1.66	1.08
Failed Dribble	1.41	0.92
Chance Created	1.37	0.90
Aerial Duel	1.36	0.88
Offside Pass	0.33	0.22
Cross	0.29	0.19
Assist	0.16	0.10
Error	0.07	0.05
2nd Assist	0.02	0.01
Error	0.01	0.01
Foul Throw	0.01	0.00
Total	100	65.25

TABLE II: Opta **attack** event taxonomy proportions for four Barclays Premier League seasons 2017/18-2020/21.

attribute is winsorised at 90 minutes, manifesting as a single peak exactly at 90 minutes rather than a distribution beyond. This means any "per 90" normalised metric using Wyscout minutes played data will overestimate the metric by an unknown amount, where players at different teams will have their metrics biased differently.

In summary, Opta's event taxonomy model is more crisply defined and more intuitive to work with than Wyscout's tag-based model, whereby there is strong evidence that Wyscout's data collection methodology systemically introduces forward-looking information to populate their point-in-time events. Systematic sanity checks of the two datasets show Opta to be both more complete for event types critical to this study and more granular with respect to positional fidelity, with the main statistic used as a normalising denominator — the total number of minutes played per player — being biased low by an unknown amount in the Wyscout dataset. Through the lens of the data, it's as though the two datasets describe two different sports, with Opta's being closer to footballing reality. For these reasons, Opta data was almost exclusively used throughout the project for modelling and analysis, except for when explicitly specified.

Shot Event	% of Shot Events	% of All Events
Miss	36.49	0.51
Shot Blocked	27.31	0.38
Shot Saved	23.29	0.33
Goal	10.81	0.15
Hit Woodwork	1.90	0.03
Penalty Saved	0.17	0.00
Missed Penalty	0.02	0.00
Total	100	1.4

TABLE III: Opta **shot** event taxonomy proportions for four Barclays Premier League seasons 2017/18-2020/21.

Opta On-The-Ball Events Data

Opta's entire event taxonomy, covering attack, shot, defence, and pressure events can be seen in TABLE II, III, IV, and V, respectively. One simple takeaway from these tables is that only 0.15% of all on-the-ball events that happen in Premier League matches are goal scoring actions, and only a further 0.1% of all events are assists (not every goal is assisted by another player). Traditionally, goals and assists were the only statistics used to quantitatively judge attacking player output. *Our xT framework will incorporate all passes, crosses, shots, and dribbles, accounting for 60% of all on-the-ball events.* Our Elo system will provide complementary event coverage to the xT framework, utilising events from the defence and pressure taxonomies for pairwise comparisons in attacker Vs defender duels.

Limitations of the Data & Synthetic Shots

There are several limitations with on-the-ball event data, first and foremost the data being blind to off-the-ball context such as the strategic formation and positioning of the defending team. As legendary A.C. Milan defender Paolo Maldini once said, "If I have to make a tackle then I have already made a mistake." Optical tracking data serves as part of the toolkit to provide this fuller picture, tracking the ball and all on-field players at 25 frames per second, however this data is prohibitively expensive and unavailable for this study. We implement a novel application to *indirectly* measure a team's defensive strengths and frailties in section V.

A second, more subtle, limitation of on-the-ball event data is the absence of attempts of actions in scenarios with a vanishingly slim chance of success because players don't attempt them, owing to

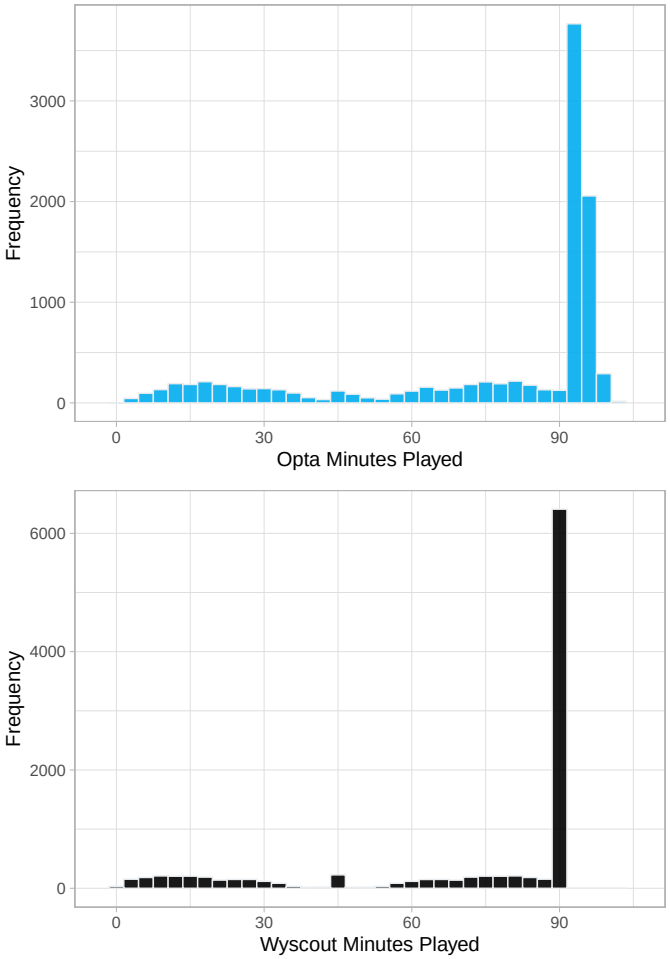

FIG. 4: Quad-modal histograms for Opta (upper panel) and Wyscout (lower panel) player minutes played per match. Both histograms exhibit the two symmetric modes roughly mirrored through the 45 minute mark, representing players being substituted in and out throughout the match and a peaked mode around 45 minutes representing half-time substitutions. Wyscout's fourth mode, the prominent peak at 90 minutes, illustrates the capping of minutes played to 90 in the Wyscout dataset, not including the expected 5-10 minutes of extra time applied to both halves that is included in the Opta data.

Defence Event	% of Defence Events	% of All Events
Ball Recovery	33.03	6.00
Clearance	13.99	2.54
Interception	6.89	1.25
Foul	6.54	1.19
Tackle	6.12	1.11
Aerial Duel	6.11	1.11
Dribbled Past	5.94	1.08
Lost Aerial Duel	4.87	0.88
Blocked Pass	4.70	0.85
Failed Tackle	3.98	0.72
Blocked Shot	2.02	0.37
Save	1.84	0.33
Offside Trap	1.20	0.22
Yellow Card	1.01	0.18
Conceded Goal	0.79	0.14
Catch	0.39	0.07
Punch	0.24	0.04
Ball Claim	0.09	0.02
Foul for Penalty	0.08	0.02
Conceded Penalty	0.07	0.01
Own Goal	0.03	0.00
Red Card	0.02	0.00
2nd Yellow Card	0.02	0.00
Saved Penalty	0.01	0.00
Total	100	18.13

TABLE IV: Opta **defence** event taxonomy proportions for four Barclays Premier League seasons 2017/18-2020/21.

Pressure Event	% of Pressure Events	% of All Events
Pressure on Pass	94.00	14.26
Offside Trap	4.32	0.66
Pressure on Shot	1.34	0.20
Shield Ball Out	0.34	0.05
Total	100	15.17

TABLE V: Opta **pressure** event taxonomy proportions for four Barclays Premier League seasons 2017/18-2020/21.

their domain expertise.

To encode shot data with such expertise [30], we created synthetic shots that can be optionally added to real shots as illustrated in the schematic in FIG. 5. The top pane of FIG. 5 shows real shot success percentages calculated in 2m x 2m bins on the pitch. The vast majority of shots are taken within the penalty area and from central areas just outside of the box around the "D". Shot success, as one might suspect, drops off as you shoot further from goal and as you shoot from tighter angles to goal, with a discontinuity at the penalty spot due to penalty kicks representing a different shot paradigm to that of open play (there are no defenders in the way!). These bins generally have hundreds of shots populating them from our four seasons of Opta data, producing representative shot success percentages for typical open play situations. As you move further from goal still, shots are rarely taken, and we draw attention to outliers by plotting footballs on the grid to represent goals, where less than 30 shots have been taken per bin, but the shot success percentage is over 10%. These outliers typically represent two situations:

1. "Shots" that weren't intended as shots by the player, like the goals being scored directly from corners in the top pane of FIG. 5. These actions are being classified as shots rather than crosses by virtue of their outcome: a historical event classification that's dependent on the future. The true denominator of shots from these positions is unknown (that knowledge is unrecorded, remaining between the ears of the player as to whether they intended to shoot, or the goal was a happy accident), but it is certain that the data grossly underestimates the total number of shot attempts required to score directly from a corner, and thus grossly overestimates the shot success probability (the Opta real shot data suggests you are guaranteed success!).
2. Shots that were taken at a position on the pitch that ordinarily would not result in a goal (and ordinarily would never have been taken in the first place), but due to a unique scenario that the player instantaneously and opportunistically calculates as a high likelihood chance, shoots and scores. An example of

this kind of scenario is when the defending goalkeeper has vacated their area (common when trailing in the dying minutes of a cup tie and their team has a corner kick), and the shooter is faced with an open goal. This unique context cannot be captured by on-the-ball events data. It would require (unavailable) optical tracking data to featurise the context to model the conditional probability of scoring a goal given the unique scenario.

The choice of treatment of such outliers is rarely agnostic to the specific research that follows. We chose to *add* synthetic shots, rather than manually *remove* outlier shots, since these shots and goals really happened — they're accurate data points — appearing as outliers due to domain expertise missing from the denominator: the number of shots *not taken* from a position on the pitch because the player had a choice, to shoot or to move the ball. Synthetic shots help to represent the shots not taken, to produce shot success probabilities closer to reality if players were forced to shoot from outlier locations. The middle pane of FIG. 5 shows the synthetically generated shots (added in equal measure to the total number of real shots), only produced for bins with small numbers of real shots, where it was hard-coded that a shot from a player's own half would not result in a goal, and shots from the opposing half would have a slim chance of scoring. The resulting shot success map, combining real and synthetic shots can be seen in the bottom pane of FIG. 5, representing shot success probabilities conditional on players having no choice other than to shoot from those locations.

The logical reason a player chooses not to shoot from such outlier locations is because a pass or dribble is, on average, a more valuable action. This insight motivates the method of construction of our Markov chain model to value such actions.

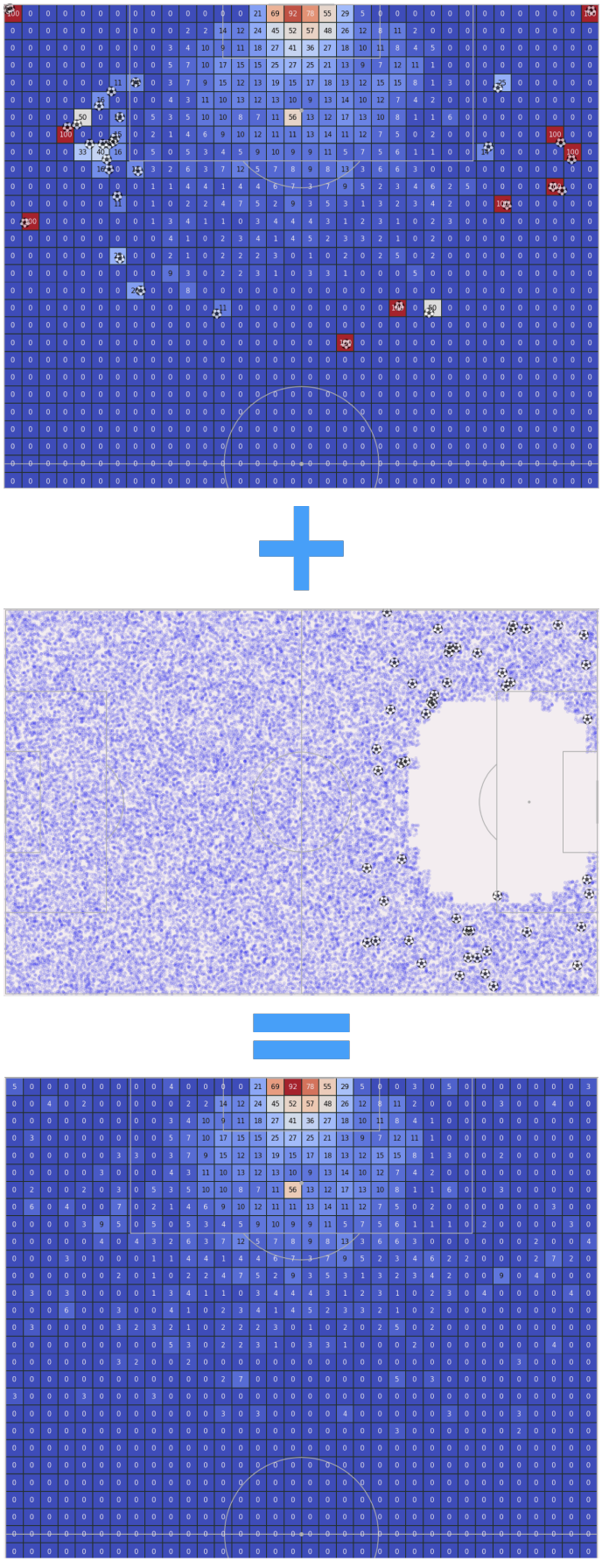

FIG. 5: Addition of synthetic shots (middle panel), representing the shots players choose *not to take* because success would ordinarily be impossible, to real Opta shots (top panel), to produce a more realistic shot success surface (lower panel).

IV. MODEL & ALGORITHM DEVELOPMENT

A. Modelling Introduction

The major modelling focus of this work centres around the Expected Threat (xT) value action framework, which models possession sequences — consecutive actions (transient states) performed by one team — as Markov chains that reach absorbing states when either a goal is scored or possession is turned over to the opposing team. As such, Expected Threat is categorised as a **possession-based** model, one of three prominent approaches for valuing actions using on-the-ball event data in the literature [15, 18, 24, 29, 33, 36]. The other two approaches — that bookend the extremes of low-complexity and high-interpretability to high-complexity and low-interpretability — can be summarised as:

1. **Count-based**: Regression models, typically regressing goals scored or shots taken onto counts of different actions to estimate coefficients per action that act as the action values. A player's overall value score is calculated by the coefficient weighted sum of the number of times they perform each action [16, 17].
2. **Feature-based**: Machine learning models where the game state — the set of the latest 3 actions up to and including the current action — is represented by a more complex set of features than possession-based approaches (which purely use the location of the ball as the sole feature describing the game state for any particular action) [8].

The original implementation of xT by Karun Singh [24] provided richer insight into the value of individual player actions than count-

based approaches, with the added sophistication showing what every armchair pundit already knew: not every pass, cross and dribble should be valued equally. Being location-based, the output of the model is a *value surface*, providing visual intuition of where goal threat is generated on the pitch to technical and non-technical practitioners alike. This interpretability is lost with machine learning approaches like VAEP [8], where action values are derived from complex non-linear functions over large sets of features [33]. For our work to be of practical use in an industry setting, insights must be intuitive and interpretable rather than pulled from a black box, if they are to be accepted and integrated as part of the analyst tool kit.

The primary limitation of xT motivates the second core model used in the project, Expected Goals (xG). In a literature comparison between the original xT implementation and VAEP, xT rankings for the top 25 Premier League players in the 2018/19 season favoured creative players that completed key passes and dribbles whereas VAEP's top 25 ranking tended to favour goalscorers, due to VAEP assigning high action value scores for goals [33]. This is absolutely no surprise, as the xT model purely assigns values to transient actions, like passes, crosses, and dribbles. Any shot, successful or otherwise, will always receive a 0 score.

xG therefore complements xT, providing a shot-specific statistic that probabilistically explains shot chance quality. A conditional probability on a binary outcome, xG is the first statistic with a sophistication beyond normalised counts of goals and assists to have crossed over into mainstream football analysis (debuting on Match of the Day in November 2017), as well as providing professional football clubs with a systematic tool to inform shot-taking decision making. On Match of the Day, it's not uncommon for a pundit to subjectively comment "9 times out of 10, they should score that!" when a striker has missed a golden chance in front of goal. xG, in this setting, acts as a descriptive statistic, and if it were to agree with the pundit in this instance, then xG would equal 0.9. *Shot action values will thus be defined by excess xG: the signed difference between actual goals scored and xG.*

Motivating a Bayesian Approach

Team strategy and player behavior is changing on-the-pitch. The top plot of FIG. 6 shows a simple linear regression model applied to the mean distance from the line bisecting the pitch longways (cutting through both penalty spots) for clusters of 100 shots arranged chronologically through time for the seasons 2017/18 to 2020/21 (43,813 shots in total). The bottom plot shows the same analysis but for the mean distance to goal. In both linear regression fittings the p value associated with a negative gradient coefficient is statistically significant at the 0.1% level of significance, demonstrating that players are taking shots that are both *closer to goal* as well as *more central* to the middle of the pitch, reducing the angle to goal. They're "making the goal bigger" before they shoot.

Rather than producing a static xT value surface that may become stale as football strategy evolves or as rules and regulations change (like the introduction of VAR), we instead took a Bayesian updating approach that uses the previous season's data as a prior and updates monthly via a Beta-Binomial conjugate analysis. This means monthly xT value surfaces produced in this project have a memory of up to the previous two years of events, where the current surface used to value tomorrow's actions strictly do not contain forward-looking information. Our Bayesian framework also supports two key properties for use in industry:

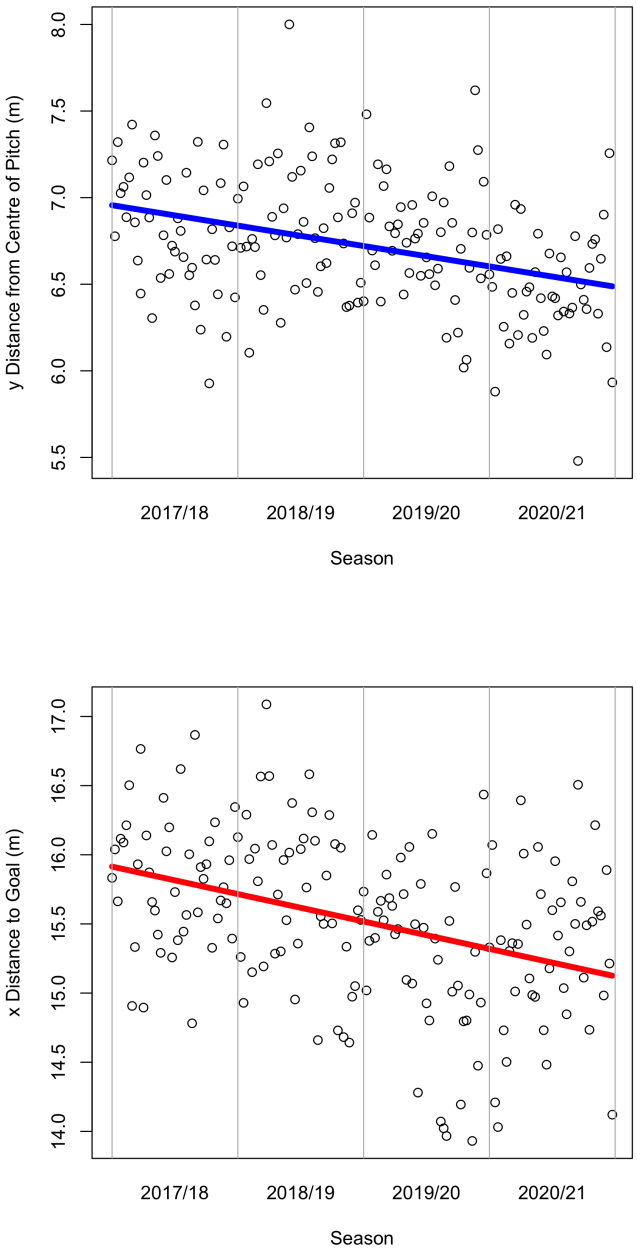

FIG. 6: Vertical distance from the centre of the pitch (upper panel) and horizontal distance to goal (lower panel) plotted against time for clusters of 100 shots chronologically ordered from 2017/18 to 2020/21. p values associated with the negative gradient coefficients for both plots are statistically significant at the 0.1% level of significance following a linear regression fitting.

1. It's trivial to integrate data encoded with domain expertise — such as our synthetic shots — as additional priors.
2. It's production-ready: the vectorised Bayesian xT methods within the xGils package are configured to process ongoing feeds of event data and can process the full four year Opta history in less than a minute on standard hardware.

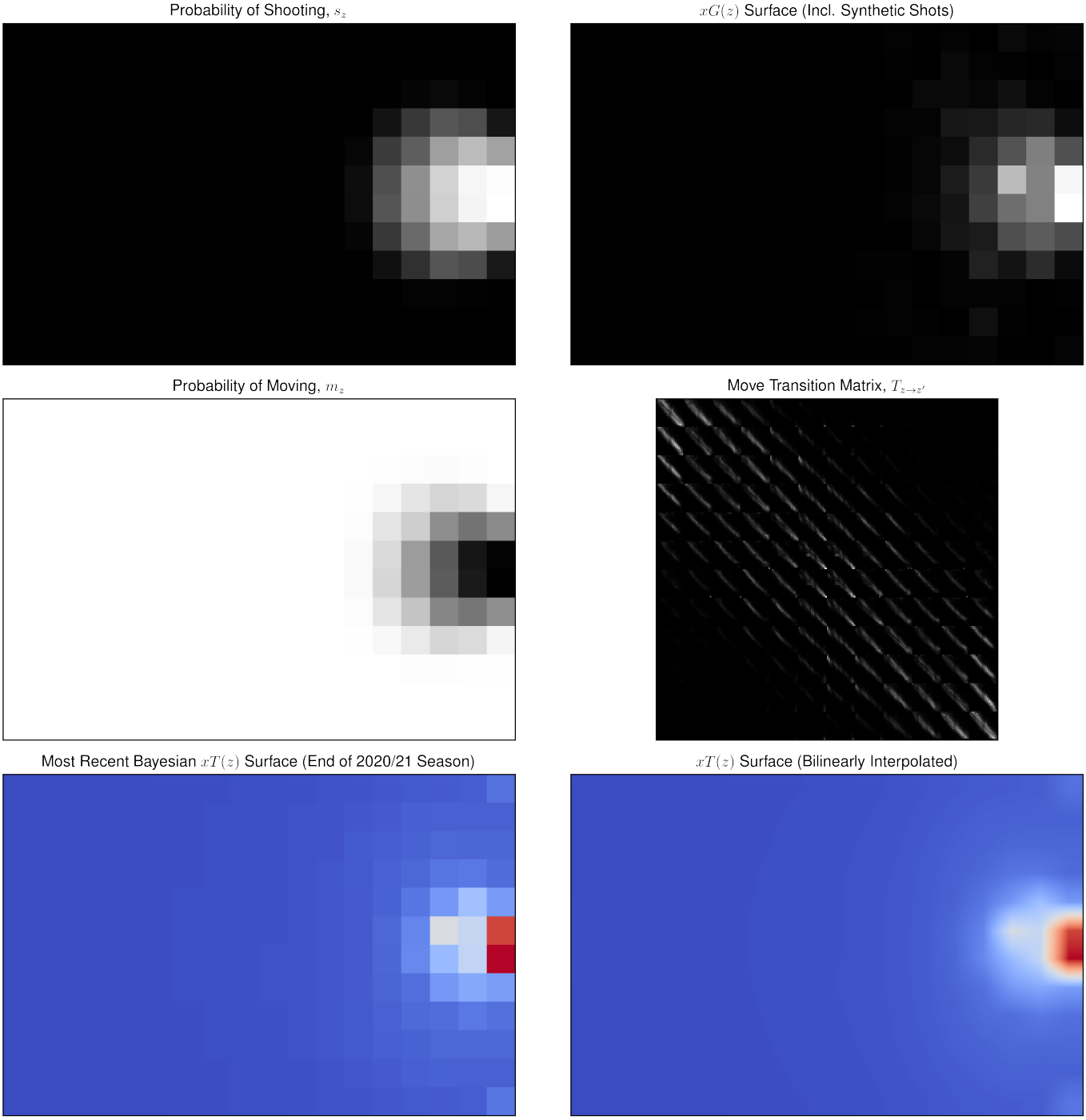

FIG. 7: Components of EQ. (1) followed by the application of bilinear interpolation to the output 12×18 xT grid surface to produce a smoother xT value surface with which to take advantage of Opta's excellent positional resolution for precise value action attribution.

B. Bayesian Expected Threat (xT)

With xT, actions are valued, in units of goals, by firstly producing a value surface on the pitch that reflects the probability of scoring a goal at a later state in the possession sequence given the current position (state) on the surface, and secondly calculating the surface value differences resulting from actions that move the ball.

xT overlays an $M \times N$ grid of zones to represent the pitch, where the Markov model can be written as:

$$xT(z) = s_z \cdot xG(z) + m_z \cdot \sum_{z'=1}^{M\times N} T_{z\to z'} \cdot xT(z'), \qquad (1)$$

where $xT(z)$ is the expected threat value for zone z, s_z is the probability that a player would choose to shoot from zone z and m_z is the complementary probability that a player would choose to move the ball (either by passing, crossing, or dribbling) from zone z. $xG(z)$ is the simplified, location-based Expected Goals (xG) probability at zone z — that is, the probability of scoring a goal conditional on a shot being taken from z — and T is the transition matrix that the characterises the Markov chain, defining the transition probabilities between zones. The value attributed to action a_i in a possession sequence by moving the ball from z to z' is calculated as:

$$V_{xT}(a_i) = xT(z') - xT(z). \qquad (2)$$

This method of calculating individual player action values provides desirable properties when compared to traditional statistics like goals and assists. Each action is assigned a score in isolation, decoupling the reward of successfully completing a valuable action from the end possession outcome. If a team's star trequartista plays a world-class through ball, putting a goal on a plate for their fellow striker, but the striker fluffs the chance, the playmaker still gets the credit. This is especially important in football compared to other sports, as it's such a low scoring game. xT also rewards moving the ball into more threatening locations — hence the name! — rather than simply moving the ball closer to goal.

The individual input components of EQ. 1, as well as the output xT value surface, are visualised in FIG. 7 for a 12×18 pitch. Every element of the four inputs — s_z, m_z, $xG(z)$, and T — represents a probability, hence the applicability for a Beta-Binomial conjugate system. For a 12×18 xT value surface, we update 47,304 posterior probabilities each month ($12 \times 18 \times 3$ for s_z, m_z, and $xG(z)$ and $(12 \times 18)^2$ for T), which in turn become the next month's priors. To initialise the very first season (2017/18) of Opta data, we use Wyscout's 2016 International European Cup events data as a prior, where the Bayesian xT modelling framework developed in this work makes constructing xT value surfaces for different underlying event datasets that use different event taxonomies trivial. All that is required is for the user to assign vendor-specific event types to one of six vendor-agnostic classification labels:

1. Successful passes;
2. Failed passes;
3. Successful dribbles;
4. Failed dribbles;
5. Successful shots;
6. Failed shots,

making the framework easy to use, regardless of the vendor data available to that user. The SPADL (Soccer Player Action Description Language) unified format to represent on-the-ball events data from a handful of proprietary data vendor formats shares this functional aim [8], however, it's implementation effectively reduces all datasets down to their lowest common denominators, inevitably reducing the richness of the original (each dataset having their own idiosyncrasies and differentiating features). This is highly undesirable, especially if the user only has access to data from a single vendor to begin with. By contrast, our implementation preserves the full richness of the underlying data, and takes less than ten minutes to integrate.

In addition to taking a Bayesian approach and engineering the framework to be easy to use for practitioners, there are two more extensions to our xT model beyond the original implementation by Singh:

1. **Negative consequences for failed movement actions**: if a player misplaces a pass or cross or is tackled when dribbling, a penalty term is calculated as the negative of the maximum of the opposition threat (as a result of the turnover) *plus* the starting threat for the team initially in possession, and zero. This impacts two player profiles who may otherwise accrue inflated xT statistics: a) players who frequently try high-risk, high-reward actions that often fail; and b) defenders who frequently give the ball away from positions of relatively low threat, to the opposition who would immediately be in possession of the ball in a dangerous area.
2. **Bilinear interpolation**, as shown the final pane of FIG. 7, a two-dimensional interpolation of the 12×18 gridded xT surface is produced to take advantage of Opta's superior positional granularity, where the enhanced precision enables us to attribute xT values to shorter passes and dribbles, that otherwise may have received a score of 0 as they did not move the ball into a different zone. This is critical in and around the penalty area as the value surface gradient sharply increases, highlighting that small differences in location at the business end of the pitch can dramatically increase the odds of scoring.

xT surface values are produced by iteratively solving EQ. (1) via dynamic programming until it converges for all zones z, which empirically takes ≈ 25 iterations, as shown in FIG. 8.

C. Expected Goals (xG)

The simple, location-based Expected Goals statistic calculated in the previous section is specific to the xT framework.

In this section, we build a logistic regression xG model to describe how the probability of shot success is shaped and to value shot actions via the excess xG metric. 20 years on from the first fitting of a logistic regression model to explain chance quality [23], it's still the model of choice for practitioners like American Soccer Analysis, a leading football analysis provider, owing to the model's interpretability [1].

Since xG models express chance quality as a probability of scoring a goal, a fundamental aim will be to produce a well-calibrated model. When a model is globally well-calibrated, the percentage of *actual* shots that are successful, when those actual shots are grouped into a bucket defined by a range of predicted xG values, the *actual* shot success percentage should be close to the central xG value of the bucket. We could not fit different xG models for kicked shots and headers separately to explore conditional calibration, as our Opta dataset does not contain a flag to distinguish between kicked and headed shots.

To quantitatively measure how good our probability forecasts are, we focused on two metrics:

1. Log loss: measuring the ability of the model to predict independent test data by calculating the mean difference (loss) between real binary shot outcomes and our predictions via the log loss function, which is specifically well-suited for predictions that are probabilistic. **We want this to be as small as possible.**
2. AUC (area under the curve): representing the chance that the model will assign a higher xG probability to a goal rather than a missed shot, if a single successful shot and a single unsuccessful shot are picked at random from an independent test sample. **We want this to be as large as possible.**

Expected Goals Feature Engineering

It was straightforward to transform the initial x, y Opta shot coordinates to produce two *basic* monotone location features:

- Distance to goal in the x direction in metres, D_x (where the x direction is the longest pitch axis).
- Distance from the line bisecting the pitch longways (cutting through both penalty spots) along the y axis in metres, D_y.

These *basic* features were used to fit a baseline model to compare log loss and AUC measures against more sophisticated models constructed using additional feature sets, as well as provide an easy to interpret jumping-off point (helped by the monotone property of both features).

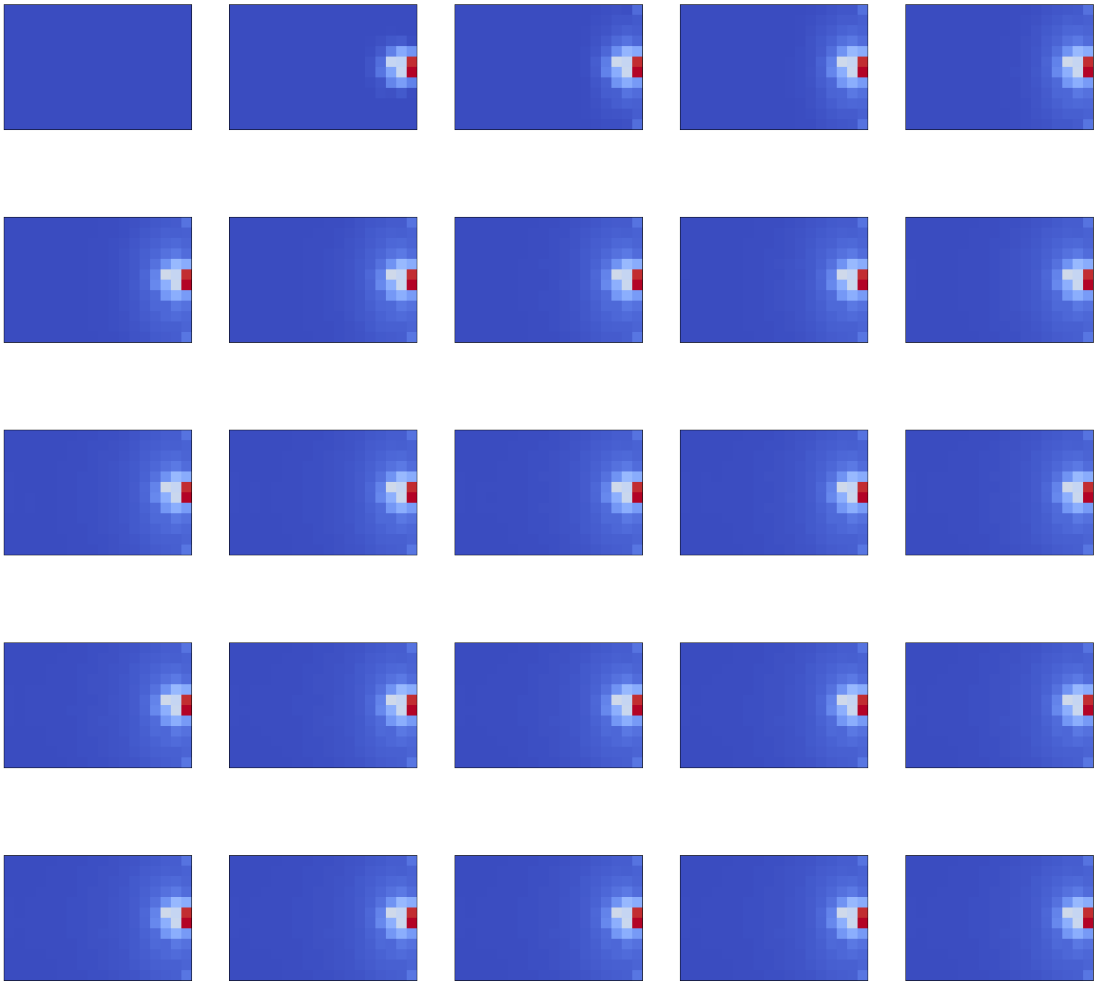

FIG. 8: Calculating the xT value surface by iteratively solving EQ. (1) where $M = 12$ and $N = 18$, using all relevant pass, cross, shot, and dribble actions. The first iteration, top-left panel, is initialised with all zones having zero threat. Convergence is achieved, bottom-right panel, when the delta between iterations is negligible *for all zones.*

We then produced derived geometric features:

- Shooting angle with respect to the mid-point of the goal using basic trigonometry.
- Shooting distance with respect to the mid-point of the goal, $D = \sqrt{D_x^2 + D_y^2}$.
- Visible angle with respect to both goal posts [28].

The visible angle, θ, is calculated via:

$$\theta = tan^{-1} \frac{7.32 \times D_x}{D_x^2 + D_y^2 - (7.32/2)^2}, \tag{3}$$

where 7.32 represents the width between the two goal posts in metres. We refer to the combination of derived geometric features and *basic* location features as *added* features.

We then started to engineer more situational features. We wanted to be able to differentiate between match situations that make scoring easier and harder. If it's more difficult for a player to score when their team is losing 4-0 and down to 10 men, than a similar chance when cruising to victory in a game where the opposition suffered an early red card, then those features should be considered. We however did not include features that we wished to summarise with in the resulting analysis, such as the player's position.

To try to capture off-the-ball context to measure the intensity of

defensive pressure on the shot, we use the "Pressure on Shot" action from Opta's pressure taxonomy and take a count of the number of defenders applying pressure, and also calculate the duration of the shooter's possession to determine how quickly the player took to shoot after receiving the ball. We also sum the cumulative team possession duration as a proxy for counter attacking chances where the defending team may not have had time to settle into their defensive shape at the time of the shot. The list of situational and contextual features (not including interaction terms or additional powers of features) beyond the *added* features produce our *advanced* feature set as follows:

- Game state: the point-in-time difference in goals between the two sides;
- Red card count: the point-in-time number of red cards a team has received;
- Player possession time;
- Team possession time;
- Shot pressure count: the number of defenders providing defensive pressure on the shot.

Expected Goals Model Fitting

Models were trained using the three feature sets on a stratified sample of 75% of the Opta shot dataset (43,813 shots in total to split into training and test sets), leaving 25% left for testing purposes. Penalties were removed prior to model fitting and assigned xG values equal to the the mean penalty success rate. Model performance measures and calibration curves were calculated on the same test set for all models.

Coefficient estimates and their associated p values for models fit with the *basic* and *added* feature sets are shown in FIG. VI and VII, respectively. The *basic* feature coefficients (both statistically significant at the 0.01% level) support what the heatmap in FIG. 5 suggests and many armchair pundits would agree on: the closer you get to goal, the better your odds of scoring. Indeed the *basic* model predicts that you can more than double your chances of scoring if you shoot from the penalty spot (with an estimated probability of scoring of $xG = 0.20$), compared to if you shoot from outside the area, from around the "D" (where the extra 10 yards reduce the probability of scoring to $xG = 0.09$). These conditional probabilities are calculated from the inverse logit of the linear predictor formed by the weighted sum of the feature values and their coefficient estimates, plus the intercept. The *basic*, baseline model also sported a seemingly impressive accuracy measure of 90%, despite not predicting a single goal (all predicted xG values on test shots were lower than 0.5). Shots have a highly skewed outcome distribution. They only go in, on average, 10% of the time. This highlights accuracy as a poor metric to use to measure xG model performance, as even the most crude model — predicting failure every time — will score high marks for accuracy.

The model fit with the *added* feature set provides a better fit to the data with a lower log loss and higher AUC, as shown in TABLE VIII. Of the two angle features, visible angle has the most significant p value and produces the largest contribution to the log loss decrease and AUC increase (as measured by fitting two more models, each without one of the angles). This feature significance again should feel intuitive for players and fans alike: by choosing to move into a shooting position that makes the goal bigger, the chance of scoring increases. Out of all features used to fit models, visible angle and shooting angle are the only pair of features that have a correlation with a magnitude of over 0.5, producing a correlation of -0.73. Both angles were, however, kept in the *added* model owing to their statistical significance in the presence of the other, and the *added* model having the optimal log loss and AUC measures when compared to supplementary models fit without one of the angles.

Feature	Coefficient Estimate	p value
Intercept	-0.244	$<.001$
D_x	-0.103	$<.001$
D_y	-0.100	$<.001$

TABLE VI: Logistic regression fitting summary for model fit using *basic* feature set.

Feature	Coefficient Estimate	p value
Intercept	2.724	$<.001$
D	-0.164	$<.001$
D_y	-0.184	$<.001$
Shooting Angle	0.961	.02
Visible Angle	1.862	$<.001$

TABLE VII: Logistic regression fitting summary for model fit with *added* feature set.

When fitting our *advanced* model using the *advanced* feature set, we also included the squared distance, D^2, as well as interaction terms between angles and distances. Finally, we augmented the training data of real shots with our synthetic shot data, to encode more domain expertise into the training of the model. *Advanced* features for synthetic shots were generated as follows:

- Sampling from an exponential distribution for the *player possession* and *team possession* lifetimes.
- Sampling from a Poisson distribution for the *pressure count on shot* and *red card count* features.
- Sampling from a normal distribution for the *game state* feature.

The *advanced* model produced a 12% decrease in log loss and a 12% increase in AUC with respect to the added model: a substantial improvement that demonstrates the importance of situational and contextual features, the interactions between features, and the value of baking domain expertise into training data. TABLE IX displays the model summary for the *advanced* model, as well as an additional column that marks which features *would no longer be statistically significant at the 5% level if the training data had not augmented with the synthetic shots*. It's notable that our synthetic data enabled

Model	Log Loss	AUC
Basic	0.293	0.740
Added	0.288	0.746
Advanced	**0.253**	**0.835**
Eggels [9]	—	0.785
Van den Hoek [32]	—	0.796
Noordman [20]	0.279	0.802

TABLE VIII: Model performance measures for the *basic*, *added* and *advanced* model on our test dataset. The *advanced* model was trained on the same test data as the *basic* and *added* models, plus the synthetic shot data. Lower rows display recent results from the literature that use Wyscout data.

Feature	Coefficient Estimate	p value	Sig. Without Syn.
Intercept	-12.330	<.001	
D	0.501	<.001	×
D^2	-0.005	<.001	
Shooting Angle	2.630	<.001	
Visible Angle	17.230	.003	×
Visible Angle $\times D$	-0.898	.01	
Visible Angle $\times \mathrm{D}^2$	0.012	.03	×
Shooting Angle $\times D$	-0.169	<.001	×
Game State	0.717	<.001	
Red Card Count	-0.314	.01	
Shot Pressure Count	-1.017	<.001	
Player Possession Time	0.006	.02	×
Team Possession Time	-0.002	.004	

TABLE IX: Logistic regression fitting summary for *advanced* model trained using synthetic shots. *Sig. Without Syn.* column marks features that would no longer be statistically significant at the 5% level if the training data had not been augmented with synthetic shots.

us to incorporate more statistically significant feature interactions into the model. The (signs of the) situational and contextual feature coefficients can interpreted as:

- The *larger* the difference in the current number of goals scored Vs the opposing team, the *better* your chances are of scoring, compared to a shot taken from the same shooting position, but a worse game state. This could represent a proxy for shooter confidence.
- The *more* red cards your team has received, and thus the less players you have on the pitch, the *lower* your odds are of scoring a goal.
- The *more* defenders applying pressure to the shot — likely increasing the odds of the shot being blocked — the *worse* your chances are of scoring.
- The *longer* the shooter takes in possession before shooting — perhaps to take an extra touch to get the ball under control — the *better* the chance of scoring.
- The *faster* the team can create a chance to shoot from a possession — possibly by counter attacking — the *better* the chance of a positive outcome.

The reduction in log loss error as the sophistication of training data composition and features engineered increases, from the *basic* baseline model to the *advanced* model, can be seen in greater detail in FIG. 9. Here, log loss is measured per zone from an 8×12 grid of zones to represent the pitch. We see the largest test errors in the baseline model occurring close to goal at tight angles from the left hand touch line (with the goal on the right from the shooter's perspective). It's clear as angle features are introduced, and subsequently angle-distance interactions are modelled, that the primary sources of log loss error are substantially reduced.

Following the iterative model fitting process, calibration curves were plotted to support the final assessment as to whether our model is fit for purpose to forecast shot outcome probabilities. These curves can be seen in FIG. 10, where calibration curves that lie closer to the "perfectly calibrated" curve can be directly interpreted as a confidence level in those predictions. We can immediately see why the baseline model failed to predict a single goal as the calibration curve fails to span xG probabilities above 0.5. Curves for the

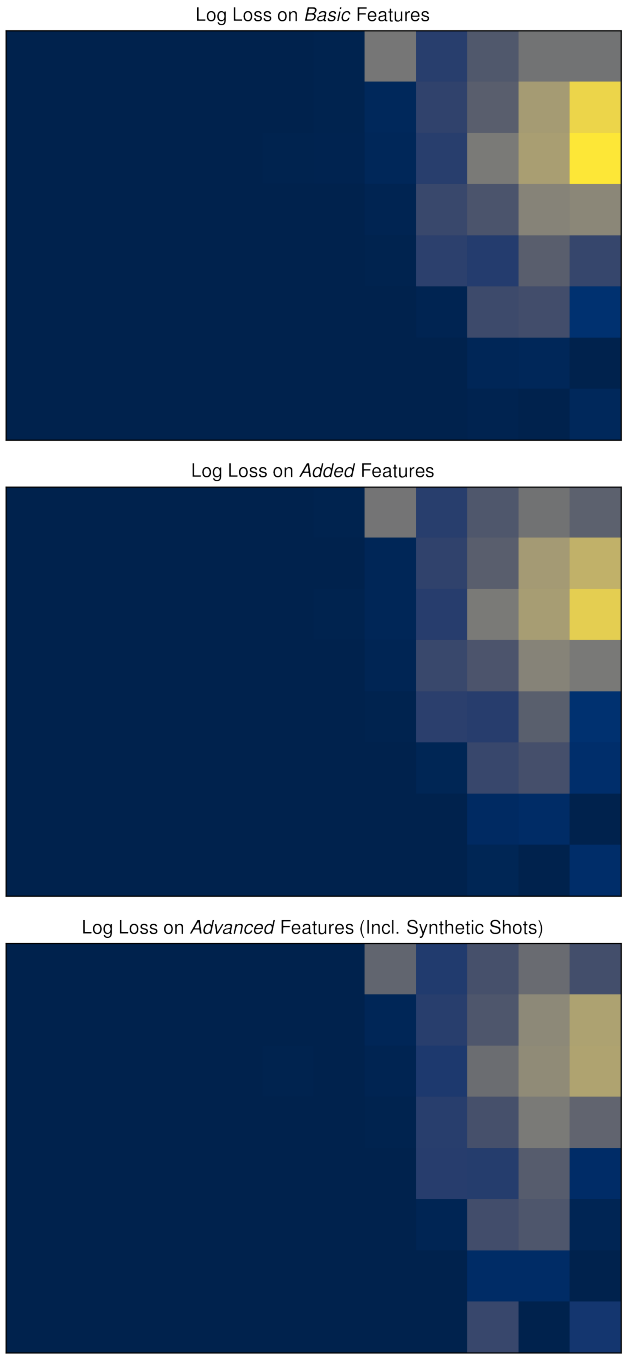

FIG. 9: Log loss heatmaps illustrating the reduction in log loss error as angle features are introduced in the *added* feature set, and further reduction as angle-distance interactions are modelled in the *advanced* feature set, alongside the inclusion of synthetic shots in the model training data.

added and *advanced* models show similarly good calibration for xG probabilities below 0.5, the calibration region that accounts for the vast majority of shots as shown in the histogram below the curves. It's the calibration region above xG values of 0.5 that the *added* and *advanced* curves start to diverge, where the *added* model can be seen to underestimate the probability of shot success for chances that are more likely to be scored than missed. To visually illustrate the effectiveness of augmenting the training data with synthetic shots, we plot two *advanced* curves, representing *advanced* models trained with and without the synthetic data (all curves are produced on the same test data). Every plotted point shows the *advanced*

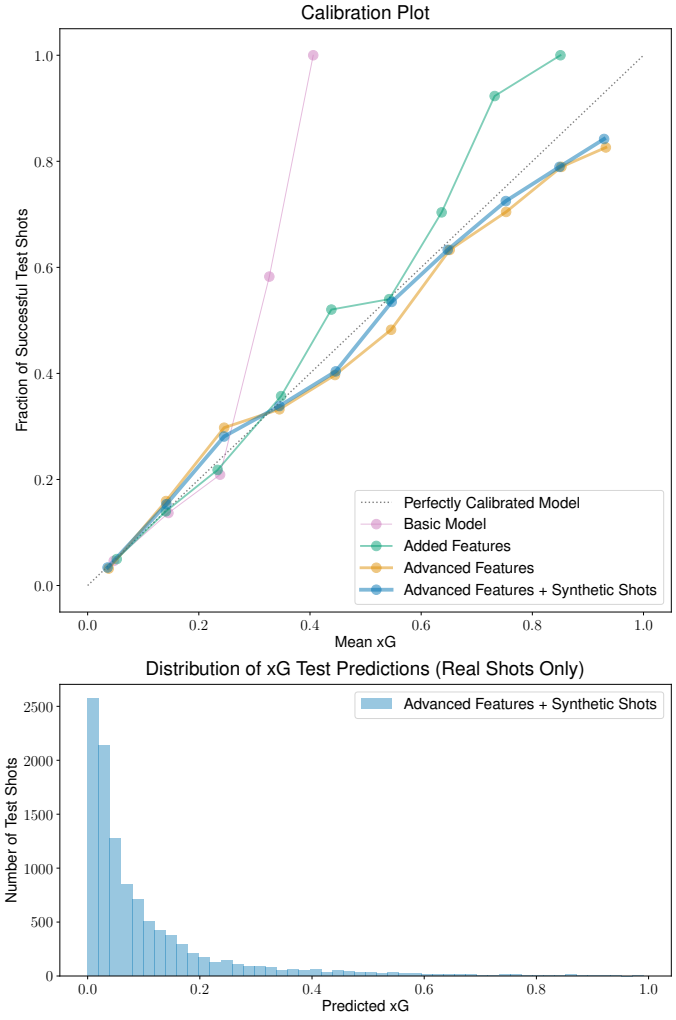

FIG. 10: Calibration curves for the three models fit using the *basic*, *added*, and *advanced* feature sets. Two variants of the *advanced* curves are plotted: one with and one without model training using synthetic data.

model trained using synthetic data to be better calibrated.

We have strong evidence that our *advanced* model is fit for purpose to produce well-calibrated xG forecasts. When compared against recent literature results in TABLE VIII, our model produces the lowest log loss and highest AUC, though it's likely that the quality of our Opta data will contribute towards our superior model performance, given the literature modelling was predominantly carried out on Wyscout data, and our data quality analysis in figure 3 showed Opta data to be superior. Often, the best way to improve model performance is to simply (but not always cheaply!) use better data. Our *advanced* model was used to apply xG values to the full Opta dataset, and subsequently calculate our metric to value shot action values, *excess xG*, defined as the signed difference between actual goals scored and xG.

As a final modelling step, we performed a dominance analysis to gain an understanding of relative feature importance in terms of how much each feature contributes towards the xG prediction [2, 3, 14]. A dominance analysis fits multiple models using every feature combination for a given model (i.e. every combination of our *advanced* feature set used to fit our *advanced* model). Models are then compared via pairwise comparisons using McFadden's R^2 as a measure of the variance explained by each fitted model (the logistic regression analogue of the R^2 statistic used in a linear regression setting). FIG. 11 shows the results of a dominance analysis of our *advanced* model. Game state is shown to be the most important individual factor in our model, having a much larger impact on xG predictions than shot pressure count, player and team possession times, and the effect of red cards, which is fascinating as it's the feature that best serves as a proxy to shooter confidence — that is, a feature providing context of what's going on between the ears of the player. The importance of angles and distances is somewhat difficult to untangle as the interaction terms cause dominance to be diluted, however a simple observation is that angles appear to have greater importance on xG predictions than distances.

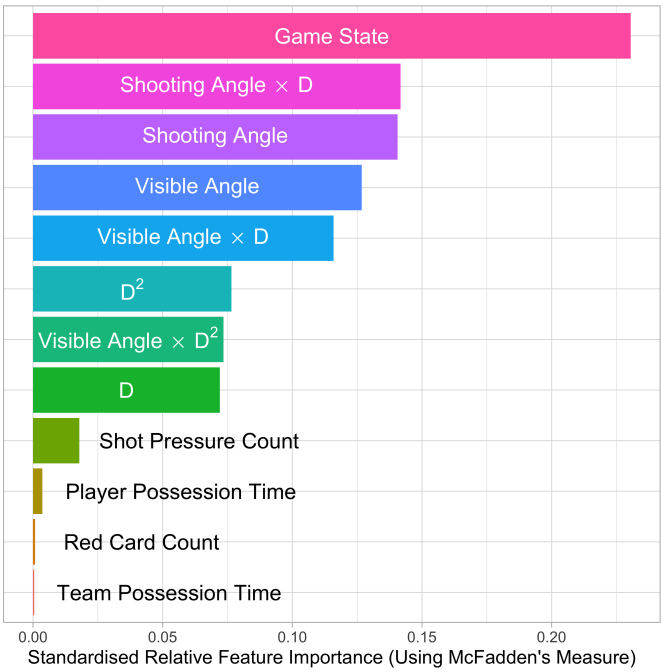

FIG. 11: Dominance analysis of our *advanced* model [2, 3, 14].

D. Mean Elo (mElo)

To complement the xT value action framework that rewards player movement into more threatening locations on the pitch, we have also implemented a custom mean Elo ranking system, originally conceived to rank chess players, to rank the dribbling and aerial duel abilities of players in one-on-one duels. With the Elo system, winners of duels are rewarded more handsomely the better the opponent they beat and they're penalised when they suffer a loss.

In order to implement the Elo system, we first needed to match the attacking side of a duel with the defensive side (neither Wyscout nor Opta provides this as part of their data product). When the Opta data is being collected, one data collector focuses on the home team, whilst a second, independent data collector focuses on the away team. Therefore the timestamps between the action of an attacking duelist and resulting reaction of the defensive duelist may not be perfectly synced. Opta applies computational post-processing to the manually tagged data to identify and automatically correct inconsistencies within the data [4]. Automatic corrections are applied to player positions, but not the action timestamps. Using this information, we created the following algorithm to map attacking

actions and defending reactions together, to enable the application of Elo rankings:

Algorithm 1 Mapping attack and defence duel actions

$df_A \rightarrow$ attacking dataframe
$df_D \rightarrow$ defending dataframe
$df_C \rightarrow$ mapping candidate dataframe
for each attack in df_A **do**
  $df_C = df_D[$
    $df_D.match = df_A.match$
    $df_D.team \neq df_A.team$
    $df_D.period = df_A.period$
    $df_D.timestamp \geq df_A.timestamp - 5$ seconds
    $df_D.timestamp \leq df_A.timestamp + 5$ seconds
    $df_D.x = df_A.x$
    $df_D.y = df_A.y$
  $]$
$mapping = df_C.sort(\textit{closest timestamp})$
return $mapping$
end for

Usually, the two sides of an Elo competition are trying to achieve the same objective: to win a game. In a football attack Vs defence duel, like dribbling, there are two objectives:

1. The dribbler objective to beat the opposition player;
2. The opponent objective to stop the dribbler,

thus we want to reward dribblers with points and penalise opponents if the first objective is achieved, and vice versa if the second is achieved, hence our custom Elo system implementation produced two dribble scores: one for attacking, and one for defending.

Each player had an Elo rating initialised at 100 points for three different Elo-based metrics:

- Dribble attack;
- Dribble defence;
- Aerial duels.

The expectation (a probability) of player A beating player B, E_A, is defined as:

$$E_A = \frac{1}{1 + 10^{(R_A - R_B)/400}}, \tag{4}$$

where R_A is the Elo rating of player A and R_B is the Elo rating for player B. As a general rule, if player A has an Elo rating of 100 points higher than player B — i.e. $R_A - R_B = 100$ — then player A has a 64% chance of beating B. Elo ratings were updated following each duel via:

$$R_A \coloneqq R_A + k \times (d - E_A), \tag{5}$$

where $d = 1$ if player A wins the duel (and is otherwise 0 if they lose), and k is a coefficient that controls the sensitivity of how quickly ratings will change to reflect changes in player performance. We set the k-factor to be 20 in our analysis, following the conventional wisdom of how Elo scoring is applied in chess [10].

To help produce more robust ratings, our implementation of the Elo system produced mean Elo ratings (mElo, [6]), taking the mean Elo scores from 10,000 iterations of the Elo algorithm, randomising the duel order each time.

V. APPLICATIONS

We focus our xT and xG models and mElo system on the four seasons of Opta data — 2017/18 to 2020/21 — to produce five applications with relevance to player recruitment and team strategy use cases as motivated in section I. Each application successively builds on the previous to produce insights of increasing sophistication.

Season	Rank	Player	xT per 90	Value (£m)
2017/18	1	Philippe Coutinho (Liverpool)	0.399	121.5
2017/18	2	Chris Brunt (West Brom)	0.378	3.1
2017/18	3	Cesc Fàbregas (Chelsea)	0.335	31.5
2017/18	4	Robbie Brady (Burnley)	0.311	9.0
2017/18	5	Kevin De Bruyne (Man City)	0.304	135.0
2018/19	1	Trent Alexander-Arnold (Liverpool)	0.350	72.0
2018/19	2	James Milner (Liverpool)	0.346	13.5
2018/19	3	Ryan Fraser (Bournemouth)	0.338	27.0
2018/19	4	Pascal Groß (Brighton)	0.326	9.0
2018/19	5	James Maddison (Leicester)	0.322	36.0
2019/20	1	Kevin De Bruyne (Man City)	0.380	108.0
2019/20	2	Trent Alexander-Arnold (Liverpool)	0.376	99.0
2019/20	3	Robert Snodgrass (West Ham)	0.310	4.3
2019/20	4	Ashley Young (Man United)	0.305	2.9
2019/20	5	Pascal Groß (Brighton)	0.302	8.6
2020/21	1	Trent Alexander-Arnold (Liverpool)	0.312	67.5
2020/21	2	Luke Shaw (Man United)	0.300	37.8
2020/21	3	Kevin De Bruyne (Man City)	0.276	90.0
2020/21	4	Raphinha (Leeds)	0.273	27.0
2020/21	5	Matt Ritchie (Newcastle)	0.266	2.7

TABLE X: Top 5 xT rankings per 90 minutes per Premier League season. Market values taken from TransferMarkt and are taken as of the end of each season.

The applications also serve a dual purpose: to enable the evaluation of the value action framework itself. Players that top xT and excess xG ranking tables should not be a surprise, and agreement between our Bayesian xT framework and an expert committee like the Professional Footballers Association (the PFA) serves to validate the model. The true value of the Bayesian xT framework is to find undiscovered talent before conventional wisdom assigns a prohibitive price tag to the player, pricing out all but the world's elite clubs with the largest chequebooks. After all, the original Moneyball strategy was not to find the best baseball players but to arbitrage undervalued talent.

1. Player xT Rankings

Our first application simply aggregates xT values per player per season and normalises per 90 minutes played. TABLE X shows Manchester City's Kevin de Bruyne (creative midfielder) and Liverpool's Trent Alexander-Arnold (right-back) to be perennial producers of the most goal threat over the past four seasons. Both have won Premier League titles with their respective clubs and accumulated personal accolades, such as de Bruyne's 2019/20 PFA Player of the Year award and Alexander-Arnold's 2019/20 PFA Young Player of the Year award.

Philippe Coutinho, who topped the chart for the 2017/18 season, would likely have gone on to win personal honours, too, had it not been for the ill-fated mid-season transfer from Liverpool to Barcelona for £122m. This spelled the start of Barcelona's recruitment nightmare: buying Europe's most promising players for exorbitant fees and having them flop at the Camp Nou as they struggled to integrate into their new team. These quantitative results highlight

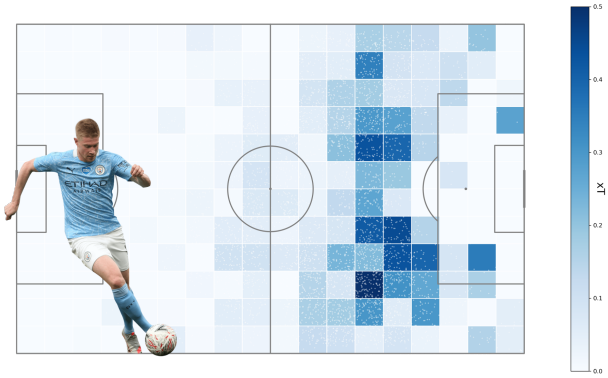

FIG. 12: xT profile for Manchester City's Kevin de Bruyne for the 2017/18-2020/21 Barclays Premier League seasons and winner of the 2019/20 PFA Player of the Year.

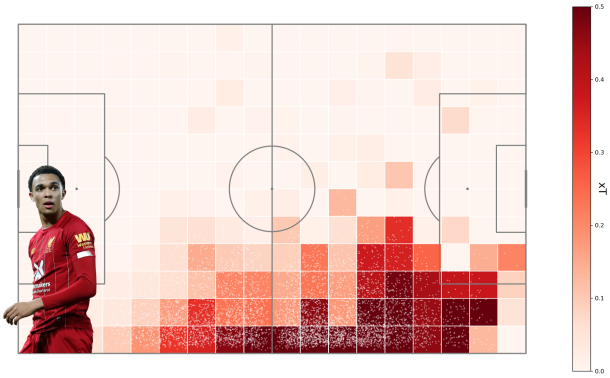

FIG. 13: xT profile for Liverpool's Trent Alexander-Arnold for the 2017/18-2020/21 Barclays Premier League seasons and winner of the 2019/20 PFA Young Player of the Year.

that our xT value action model successfully surfaces the best attacking players the Premier League has to offer (and that our Bayesian approach enables us to value actions in our very first season of data, rather than losing seasons of expensive data to model training like the VAEP approach [8]); and the Barcelona reality shows that there's more to recruitment than just picking players with the best stats. Liverpool, on the other hand, spent the £122m wisely, buying Southampton's Virgil van Dijk (centre-back) for £76m and Roma's Alisson (goalkeeper) for £56m — both of whom would be instrumental in Liverpool winning Champions League *and* Premier League silverware in successive seasons.

We observe that whilst de Bruyne and Alexander-Arnold generate similarly impressive quantities of threat, they do so in different ways and different areas of the pitch. FIGs. 12 and 13 illustrate player xT pitch profiles for de Bruyne and Alexander-Arnold, respectively, showing how de Bruyne operates centrally, generating threat by exploiting the space in-between the opposing centre-halves and full-backs, whereas Alexander-Arnold sticks to the right flank before fanning threat towards the opposing penalty as he approaches the final third of the pitch. We analyse threat at a more granular level via threat components — xT from passes, crosses, or dribbles — and combine our additional Elo-based metrics and excess xG to produce the radar in FIG. 14. Both players are exceptional at threatening passes — scoring in the 99th percentile of xT via passing per 90 minutes. Both are also excellent crossers of the ball (with Alexander-Arnold just edging de Bruyne), where

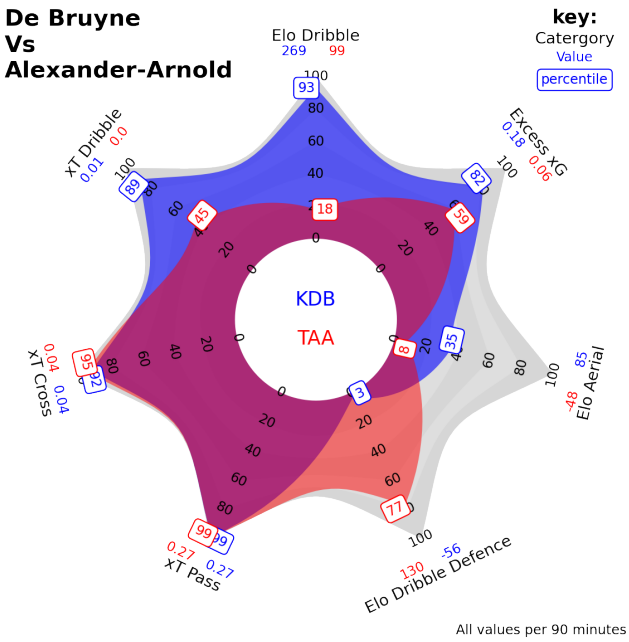

FIG. 14: Proportional area radar chart comparing Manchester City's Kevin de Bruyne to Liverpool's Trent Alexander Arnold. Statistic values and percentiles are aggregated over all four seasons of Opta data.

de Bruyne favours shallower crosses from the right, just outside the box towards the back post, whilst Alexander-Arnold prefers to cross from closer to the touchline. Indeed, it was this skill of Alexander-Arnold's that stunned Barcelona in the 2018/19 Champions League knock-out stages, where his pinpoint, low-driven corner kick found Divock Origi for a tap-in winner before the leaderless Barcelona defence could get into position.

Where the two players start to differ is threat via dribble. Alexander-Arnold is a below-average dribbler, both in his ability to strategically move into space of increasing threat (expressed via the dribble xT statistic) and his ability to beat an opponent one-on-one to get into that position of increased threat (expressed via the Elo dribble statistic). In contrast, de Bruyne ranks in the top 10% of Premier League players for his technical ability to beat an opposing defender on the dribble *and* his tactical ability to identify positions to move into. As one might expect when comparing an elite full-back and elite attacking midfielder, de Bruyne is the superior finisher (as expressed by the excess xG statistic), and Alexander-Arnold is the superior defender on the ground (as expressed by the Elo dribble defence metric). Neither player is strong in the air, as shown by both player's below-average aerial Elo ratings.

This initial application showcases our ability to rank players based on the offensive threat they generate via our Bayesian xT framework and subsequently perform an intuitive deep-dive to compare players in different positions, surfacing the primary locations and components of threat as well as our complementary excess xG and Elo metrics.

2. *Paired xT Concentration*

We build on the initial application by calculating *paired xT concentration* to look for key partnerships between pairs of players, be-

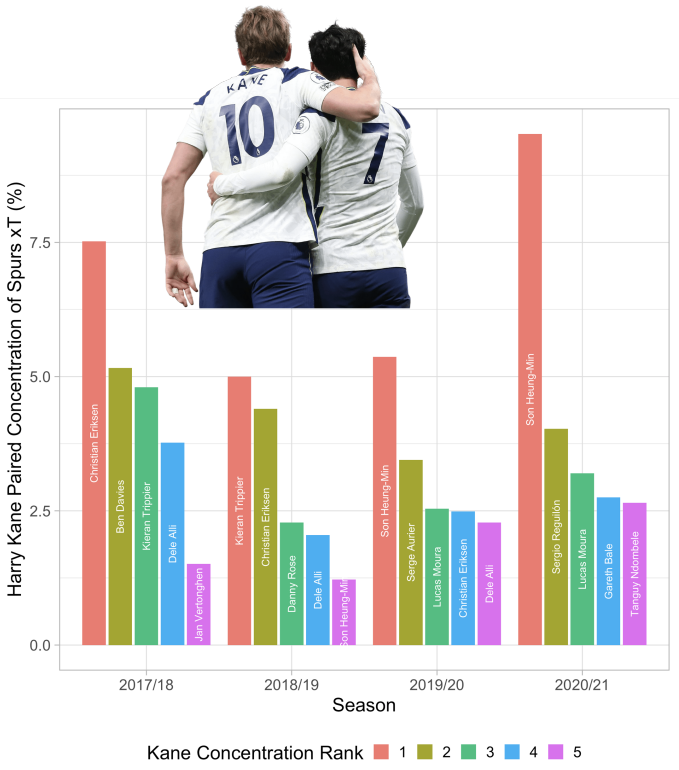

FIG. 15: Paired xT concentration as a percentage of Spurs' total xT generated per season between Harry Kane and his teammates over the past four seasons, showing an increasing reliance on Kane's partnership with Son Heung-min as a result of Christian Eriksen leaving the club for Inter Milan.

Season	Rank	Player	ExG per 90	Value (£m)
2017/18	1	P.E. Aubameyang (Arsenal)	0.658	67.5
2017/18	2	Mohamed Salah (Liverpool)	0.541	135.0
2017/18	3	Oumar Niasse (Everton)	0.438	7.2
2017/18	4	Sergio Agüero (Man City)	0.408	72.0
2017/18	5	Gabriel Jesus (Man City)	0.378	72.0
2018/19	1	Sergio Agüero (Man City)	0.695	58.5
2018/19	2	Sadio Mané (Liverpool)	0.621	108.0
2018/19	3	P.E. Aubameyang (Arsenal)	0.570	63.0
2018/19	4	Raheem Sterling (Man City)	0.532	126.0
2018/19	5	Mohamed Salah (Liverpool)	0.525	135.0
2019/20	1	Sergio Agüero (Man City)	0.853	46.8
2019/20	2	Olivier Giroud (Chelsea)	0.706	6.3
2019/20	3	Mason Greenwood (Man United)	0.636	45.0
2019/20	4	Danny Ings (Southampton)	0.630	18.0
2019/20	5	Raheem Sterling (Man City)	0.627	115.2
2020/21	1	Gareth Bale (Spurs)	1.013	16.2
2020/21	2	Diogo Jota (Liverpool)	0.695	40.5
2020/21	3	Kelechi Iheanacho (Leicester)	0.656	18.0
2020/21	4	Edinson Cavani (Man United)	0.611	5.4
2020/21	5	Harry Kane (Spurs)	0.554	108.0

TABLE XI: Top 5 excess xG (ExG) rankings per 90 minutes per Premier League season. Market values taken from TransferMarkt and are taken as of the end of each season.

tween which significant fractions of the overall team threat is generated. This is, therefore, an application that speaks to both player recruitment and team strategy use cases.

Using the Opta data, it is straightforward to determine the recipient of an action that moves the ball from one player to another and thus aggregate threat values over pairs of players. We then calculate paired xT concentration by normalising the threat generated per pair by the overall threat generated by the team per season. Teams with high concentration pairs will be more fragile for two reasons:

- Something may happen to one of the players within the pair, whether that be a long-term injury, a bid of one of the players from another club, or a prolonged period of poor form;
- Opposing teams may strategically target the partnership by man-marking one or both players or looking to cut the supply line between the pair.

Harry Kane, and his partnerships with Christian Eriksen and Son Heung-min at Tottenham Hotspur, are prominent examples of such key-person risk and team fragility. Kane features in the top five of the highest concentrated pairs for multiple seasons from 2017/18 to 2020/21 and FIG. 15 displays Kane's most concentrated xT partnerships over the four seasons.

In 2017/18, 7.5% of Spurs' entire offensive threat was generated between Eriksen and Kane, producing the largest absolute paired xT of any pair of players in the Premier League. Over the next three seasons, Eriksen's contribution diminished until his eventual sale to Inter Milan after the 2019/20 season, and Son's contribution dramatically increased to the point that the Kane-Son partnership generated nearly 10% of all team threat, contributing more than twice the concentration than any other Spurs pair. It's no wonder that Son was rewarded with a new contract after the 2020/21 season and that Spurs chairman Daniel Levy failed to pick up the phone when Manchester City came calling for Harry Kane with a £150m bid in the 2021 summer transfer window. Whilst the contractual moat has bought Spurs at least one more season of this essential pair, it does not mitigate the fragility. Opposing teams should increasingly look to cut the supply between the two players (perhaps easier said than done), and a long-term injury to either player would be devastating.

To show how our suite of metrics could form the basis of a recruitment strategy should Kane look to leave Spurs in a future transfer window — or simply provide increased strength in depth to cope with the hectic intensity of a Premier League club's fixture calendar — we first look for Kane replacement candidates by ranking excess xG in TABLE XI, as a replacement would need to be an equally lethal finisher for them to fulfil the same centre forward role in the team.

Certain candidates from TABLE XI can be ruled out immediately due to being practically infeasible — Arsenal would never sell star striker Pierre-Emerick Aubameyang to their North London rivals, and Manchester United would never sell local Mancunian wonderkid Mason Greenwood after his breakout season in 2019/20. Potential players of plausible interest include Manchester City's legendary poacher, Sergio Agüero, who had announced that he would be leaving City for a new challenge at the end of the 2020/21 season, and Leicester City's 24-year-old Kelechi Iheanacho, who was even more lethal in front of goal than Harry Kane, boasting an excess xG score of 0.695 compared to Kane's 0.554 per 90 minutes in 2020/21. Iheanacho's crowd-sourced TransferMarkt market value of £18m is a fraction of Kane's, too. The radar to compare Kane and the two replacement candidates using aggregated metrics for all four seasons of data is shown in FIG. 16. All three strikers boast clinical 99th percentile excess xG finishing scores and, as one might expect from goalscorers wired to selfishly monopolise the team's chances, rank

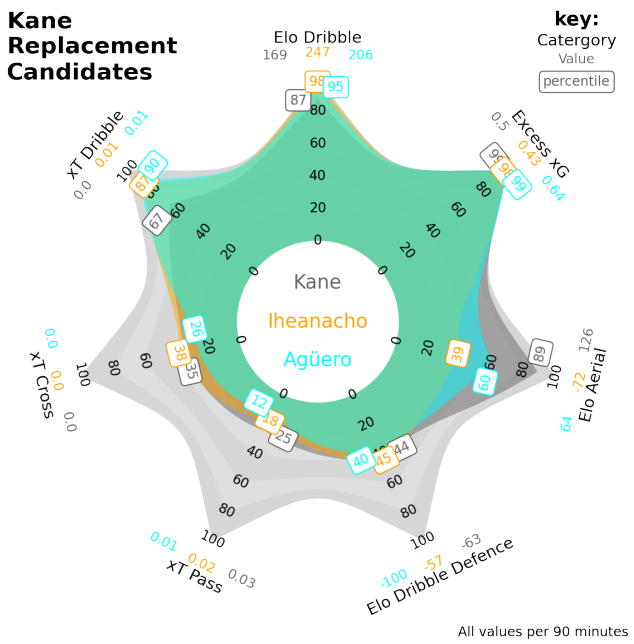

FIG. 16: Proportional area radar comparing Spurs' Harry Kane with systematically identified replacement candidates Sergio Agüero (Man City) and Kelechi Iheanacho (Leicester City).

lowly for crossing, passing, and ground defence metrics (a very different set of profiles to those in FIG. 14). The strikers principally differ with respect to their dribbling and aerial abilities. Kane is the best header of the ball by some margin, with Iheanacho the worst. Iheanacho, however, possesses the strongest ability to beat his man one-on-one in the dribble and is better than Kane tactically, with his dribbles moving him into positions of higher threat than Kane's. At 24 years old, Iheanacho still has time to develop his aerial ability and could be a genuine Spurs prospect to add strength in depth at a reasonable price.

This application showcases the ability of this work to systematically identify team fragility, either for a team looking to mitigate key-pair risk through recruitment diversification or for opposing teams to find partnerships to target strategically.

3. Team xT

The third application aggregates xT values *per team* per season to produce heatmaps of where teams focus their attacking threat over the pitch, highlighting their offensive strengths and weaknesses. Team xT pitch profiles for the "Big Six" Premier League clubs — Manchester United, Manchester City, Liverpool, Chelsea, Arsenal, and Tottenham Hotspur — are shown in FIG. 17. These heatmaps tell several stories: Arsenal and Spurs' fall from grace from perennial "top four" contenders to mid-table obscurity as their attacking threat has become increasingly blunted; Chelsea re-balancing their threat over both wings after a disappointing season of predictable attacks down the right flank in 2017/18 which saw them finish outside the Champions League places; and Liverpool and Manchester City shifting the focus of their threat from central areas in 2017/18 to the wings in recent years, to take advantage of their world-class winger talent in Mo Salah and Sadio Mané (Liverpool) and Riyad Mahrez and Raheem Sterling (Man City).

The focus of this application, however, centres around Manchester United, a team that has struggled to rediscover its identity for the best part of a decade following the departure of Sir Alex Ferguson — the most successful manager in British football history — after 27 years at the helm. As football formation strategy evolved from the classic 4-4-2 to a more dynamic 4-3-3, Ferguson evolved with it. He replaced the role of the wide midfielder (ordered not to deviate too far from the touchline and to return to the dressing room after 90 minutes with chalk on their boots), with more general forwards with a greater licence to tactically interchange positions as they feel out weaknesses in the opponent's defence during the opening phase of a game. Indeed, Sir Alex's forward trio of Cristiano Ronaldo, Carlos Tevez, and Wayne Rooney that won the Champions League in 2008/09 is still regarded as one of the most devastating attacks in modern football.

FIG. 17 shows Manchester United's attack becoming increasingly unbalanced and predictable, with offensive threat primarily being generated down the left wing through forward Marcus Rashford and left-back Luke Shaw in the 2020/21 season (note, Shaw was the #2 ranked source of threat behind Alexander-Arnold in TABLE X). The maps in FIG. 17 may therefore be used by an opposing manager to quantitatively assess the origin of the main threats they will face such that they can more confidently allocate defensive resources when not in possession of the ball. This application also has strong relevance to player recruitment: to strengthen in positions of systematically identified weakness and add depth to the squad if only one player can perform to the required standard in a given position. Typically, a Premier League club will look to have two high-performing players per position. In 2020/21, Marcus Rashford played almost every game (playing large parts through injury), whilst game time was more evenly split on the right wing between Mason Greenwood and Daniel James. FIG. 18 shows a radar comparing the three wingers. Rashford is superior in the air with by far the best Elo aerial metric, with none of the three players in the same category of clinical finishers as Kane, Iheanacho, and Agüero (Greenwood's finishing reverted to the mean after his breakout season in 2019/20). The striking outlier is James' poor dribbling ability, an essential skill for an elite winger. James' Elo dribble score puts him in the 46th percentile, reflecting a below-average ability to beat an opposing defender one-on-one. Worse still is his tactical ability to recognise which spaces he should dribble into to increase the threat of scoring a goal. *Remarkably, James generates a net negative xT score via the dribble, meaning that on average, he dribbled into less threatening locations on the pitch than where he started.* The radar thus answers the question of why Manchester United appear unbalanced in how they generated threat throughout the 2020/21 season. In games where Greenwood lined up alongside Rashford, United threatened over both flanks in more equal measure than when James was on the field. It is perhaps no surprise that Manchester United's primary transfer target in the summer 2021 transfer window was Borussia Dortmund's star right-winger Jadon Sancho. After a two-year-long transfer saga, United successfully signed Sancho for £73m and shortly after sold James to Leeds for £24m.

With this application, team xT pitch profiles have been shown to objectively identify offensive strengths and weaknesses of teams at a high level, providing insight to opposing teams as to where to focus defensive efforts, as well as posing questions for strategic recruitment that can be systematically answered through the inspection of player radars in those problem areas. By extension, the application also provides the tools to measure the impact of new players

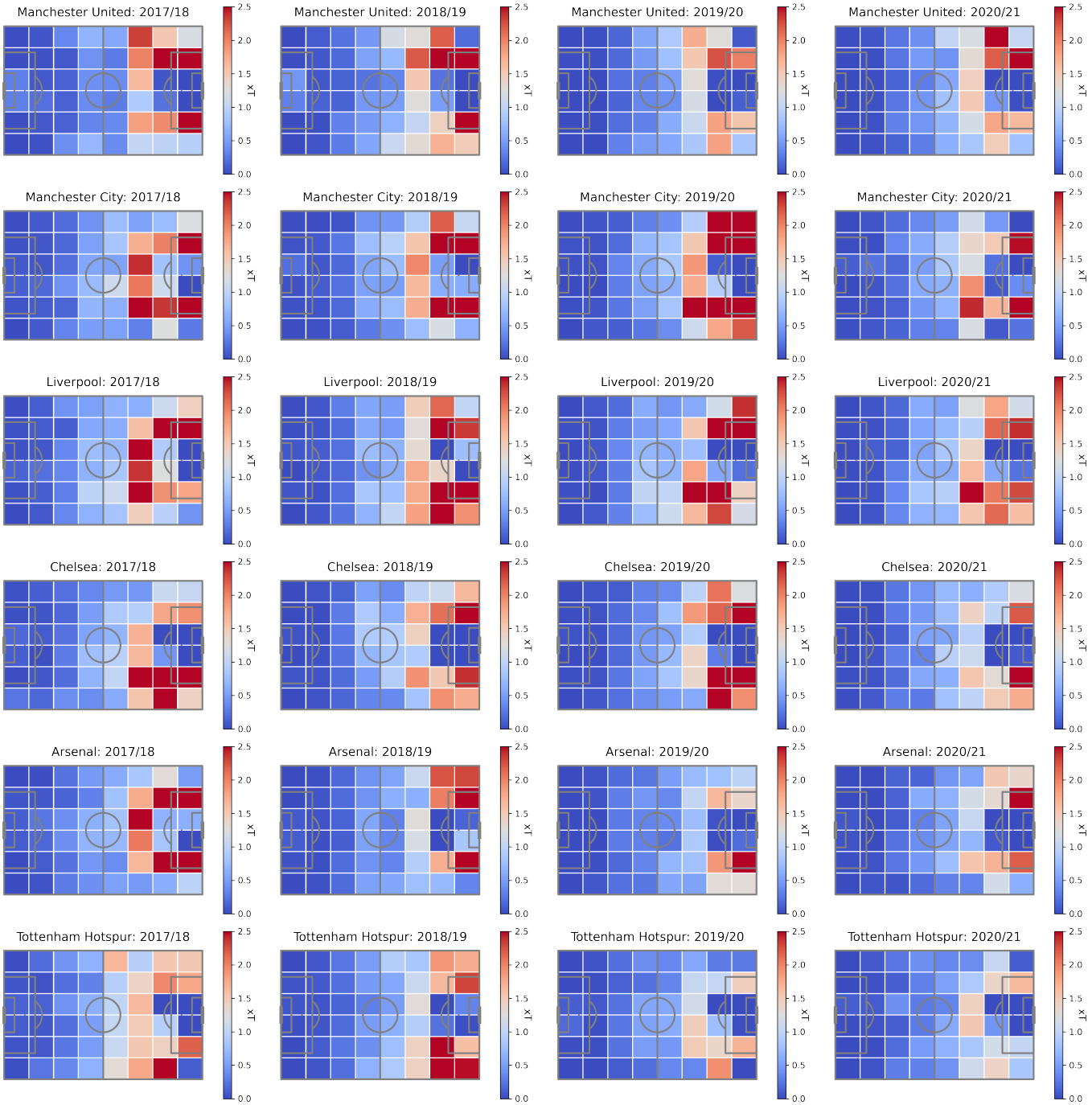

FIG. 17: Aggregated xT per team per season showcasing different play styles and threat strategies of the Premier League's "Big Six" clubs. Attacks flow from left-to-right.

and objectively assess recruitment success.

4. Delta xT

The primary limitation of on-the-ball events data is that it's blind to off-the-ball context. This mainly impacts our ability to quantitatively measure and analyse the defensive side of the game beyond counting tackles, interceptions, and clearances (and perhaps normalising for possession). Optical tracking data to directly capture player and ball positions at 25 frames per second remains prohibitively expensive to all but the world's wealthiest clubs, and even then, only provides data for a fraction of all games played (teams do not share this data with their rivals, and thus is only available for a team's home games).

Defensive metrics must be included as part of a comprehensive quantitative toolkit to assess team strengths and weaknesses. As Sir Alex once said, "*attack wins you games, defence wins you titles*". We therefore engineered an entirely novel application to *indirectly* measure a team's defensive strengths and weaknesses. This application is called *Delta xT*, where the following methodology produces a Delta xT heatmap — the defensive analogue to the offensive threat heatmaps produced in the previous application — for a given team:

- Calculate the mean opponent threat per $M \times N$ zone *versus* the team of interest (i.e. the home and away tie);
- Calculate the mean opponent threat per $M \times N$ zone *exclud-*

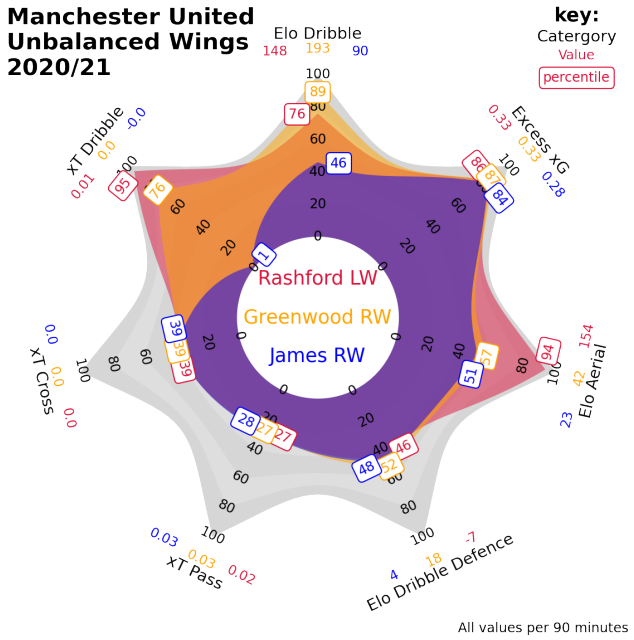

FIG. 18: Proportional area radar comparing Manchester United's wingers in the 2020/21 season: Marcus Rashford (predominantly played down the left wing), and Mason Greenwood and Daniel James (predominantly played on the right wing).

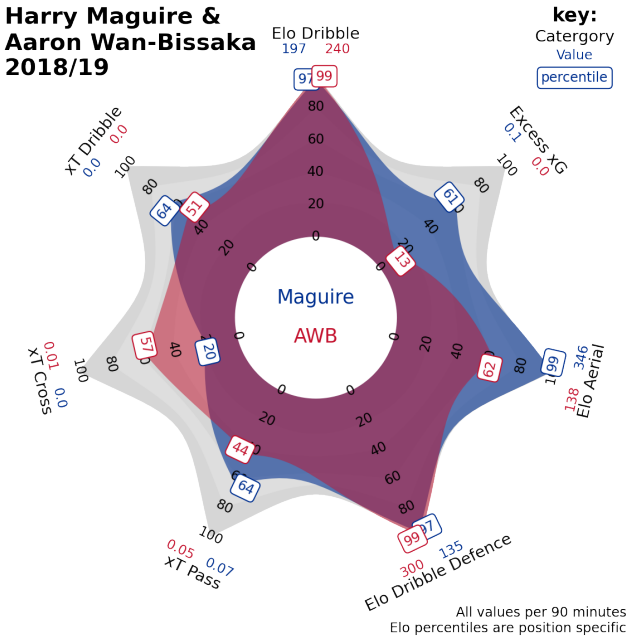

FIG. 19: Proportional area radar showing Manchester United's 2019 summer signings following the 2018/19 season: Leicester City's Harry Maguire (£80m, centre-back) and Crystal Palace's Aaron Wan-Bissaka (£50m, right-back).

ing the team of interest (i.e. the *other* 36 games the opponent plays a season);

- Calculate the mean of the threat deltas (*versus* team of interest − *excluding* team of interest) for all 19 opponents in the league.

Hence, Delta xT provides an indirect-but-intuitive method to quantitatively identify where opponent's deviate from their usual strategy when specifically playing against the team of interest, however subtle, normalising for the different play styles of different teams. The application is data-hungry: each team's Delta xT heatmap requires every single attacking action from all opponent teams within the league for that season. It therefore utilises 19 times more data as an offensive heatmap. Delta xT heatmaps for the "Big Six" are shown in FIG. 20, where the heatmaps are orientated such that the team of interest attacks from left to right, and thus the left-hand half contains the goal being defended.

Sir Alex's words appear to ring true when inspecting Manchester City's Delta xT heatmaps in FIG. 20. Manager Pep Guardiola has led City to three Premier League titles in the past four seasons, coming second only to Liverpool in 2019/20. Guardiola has spent nearly £500m on defenders and goalkeepers since he arrived at the club in 2017, most of which has been spent on central defenders. The biggest spend came in the summer prior to kicking off the 2017/18 season, buying 23-year-old goalkeeper Ederson from Benfica for £36m, 27-year-old right-back Kyle Walker from Spurs for £47.5m, 23-year-old left-back Benjamin Mendy from Monaco for £52m, and 23-year-old centre-back Aymeric Laporte from Athletic Bilbao for £58.5m. This brand new defence conceded the fewest goals in the Premier League in 2017/18, with the minor blip in that year's Delta xT heatmap around the left-back position highlighting the long-term injury to Mendy, forcing Guardiola to deploy defensive midfielder Fabian Delph out of position at left-back for most of the season. With an average age of 24, one might have expected this newly-formed defensive block to play together for the best part of a decade, and yet following the 2019/20 season, after losing the title to Jürgen Klopp's Liverpool, Guardiola purchased two more centre-backs: 23-year-old Ruben Dias from Benfica for £65m and 25-year-old Nathan Ake from Bournemouth for £40m. City's Delta xT for 2019/20 shows the emergence of defensive frailty at the heart of the defence, with opposing teams altering their usual attacking patterns to focus on City's centre-backs. Despite scoring far more goals than any other team in 2019/20, City lost their title to Liverpool by nearly 20 points, with Liverpool's defence, led by left centre-back Virgil van Dijk, conceding the fewest goals in the league.

The impact of Dias' arrival at City in 2020/21 — which would see him win the PFA Player of the Year award — and van Dijk's season-ending injury after only five games were the two key events that many pundits believed shaped the course of the 2020/21 season and handed the title back to City. City's 2020/21 Delta xT map objectively illustrates Dias' impact at the right side of central defence. The previous season's weakness is transformed into a strength as opponents strategically avoid Dias' lane of influence. In contrast, Liverpool's Delta xT profile for 2020/21 is almost unrecognisable to the previous three. In previous seasons, with van Dijk positioned on the left side of central defence, the primary chink in Liverpool's armour can be seen at right-back where opponents target the space left behind by Alexander-Arnold's forward runs as shown in FIG. 13. In 2020/21, however, Liverpool's opponents specifically targeted the absent sphere of van Dijk's influence, both at left-back and the left side of central defence as evidenced in FIG. 20. A significant advantage of the Delta xT application is that it enables an objective introspection about where things are going wrong, rather than pandering to public perception. Liverpool would concede 27% more goals in 2020/21, with many fans and pundits pointing the blame at Alexander-Arnold, which subsequently resulted in England manager Gareth Southgate dropping the player from England's

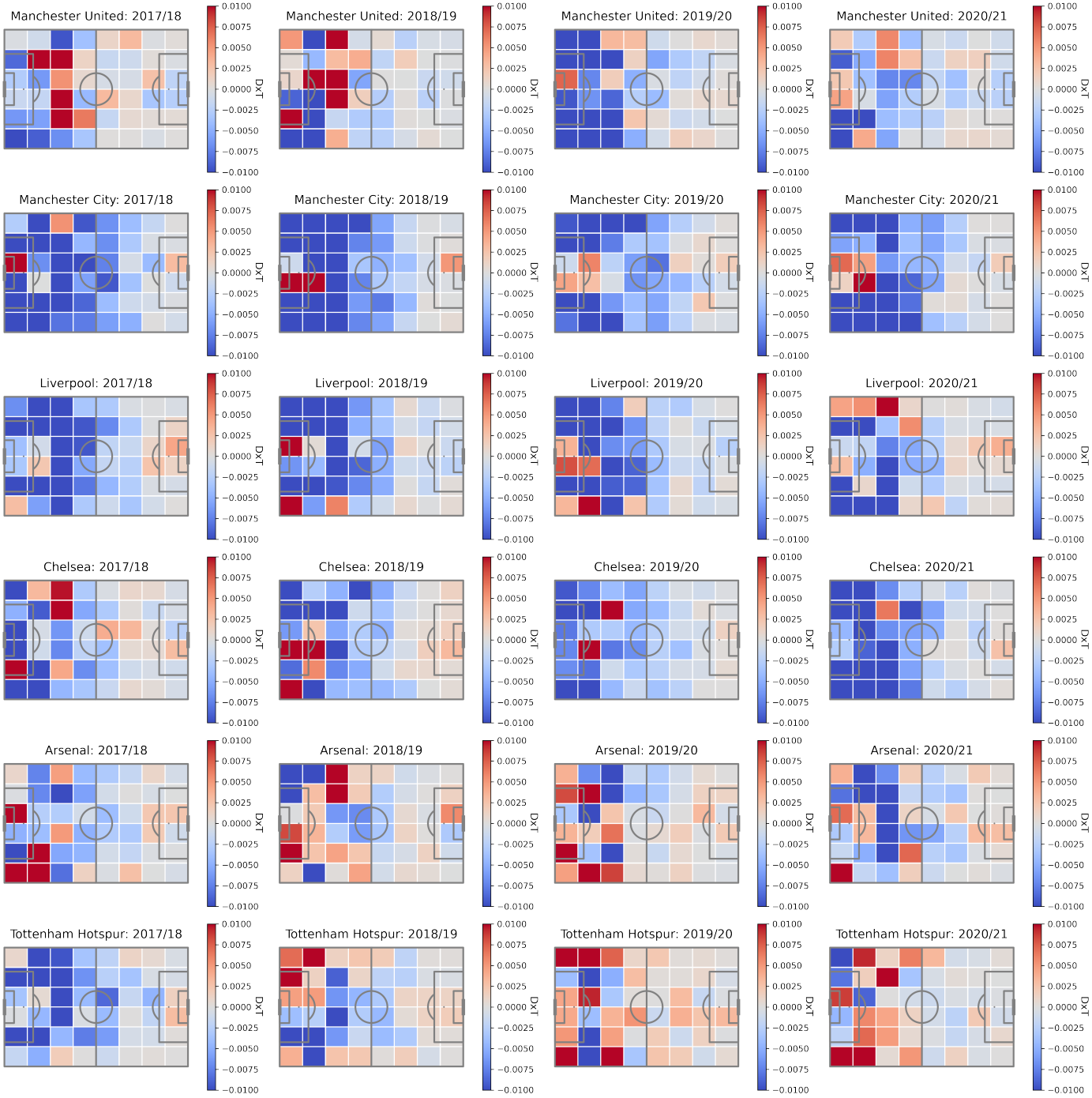

FIG. 20: *Delta xT* strategies carried out by the opponents of the "Big Six" Premier League clubs for the four seasons from 2017/18 to 2020/21. Opponents attack the "Big Six" from right-to-left.

European Championship squad. As our systematic analysis shows, however, Alexander-Arnold was the source of the greatest offensive threat in 2020/21 as reported in TABLE X and was likely not the cause of Liverpool's defensive frailty as displayed in FIG. 20.

Chelsea and Manchester United's defences have also undergone substantial fortification, with United's key transfer business being done prior to the 2019/20 season, purchasing Leicester City's 26-year-old Harry Maguire at centre-back for £80m — a world record fee for a defender — in addition to 21-year-old right-back Aaron Wan-Bissaka from Crystal Palace for £50m. Maguire and Wan-Bissaka's radar from their 2018/19 campaigns, shown in FIG. 19, highlight Maguire as being the *best* aerial duelist in the league, and Wan-Bissaka to be by far the *best* ground dribble duelist. The two defenders also exhibit complementary playmaking attributes. Maguire possesses an above-average ability to dribble the ball out from the back and pass the ball forward to teammates in more threatening positions, whilst Wan-Bissaka possesses an above-average ability to cross the ball from right-back. In short, United bought well-rounded defenders who excelled at stopping opposing attacks in the air and on the ground, who were also capable of turning defence into attack, and in Maguire's case, could contribute to the team's goal tally with headed goals from set-pieces. The effects can be seen in FIG. 20 as United's broad defensive frailties are transformed into a backline with no apparent weakness for opponents to

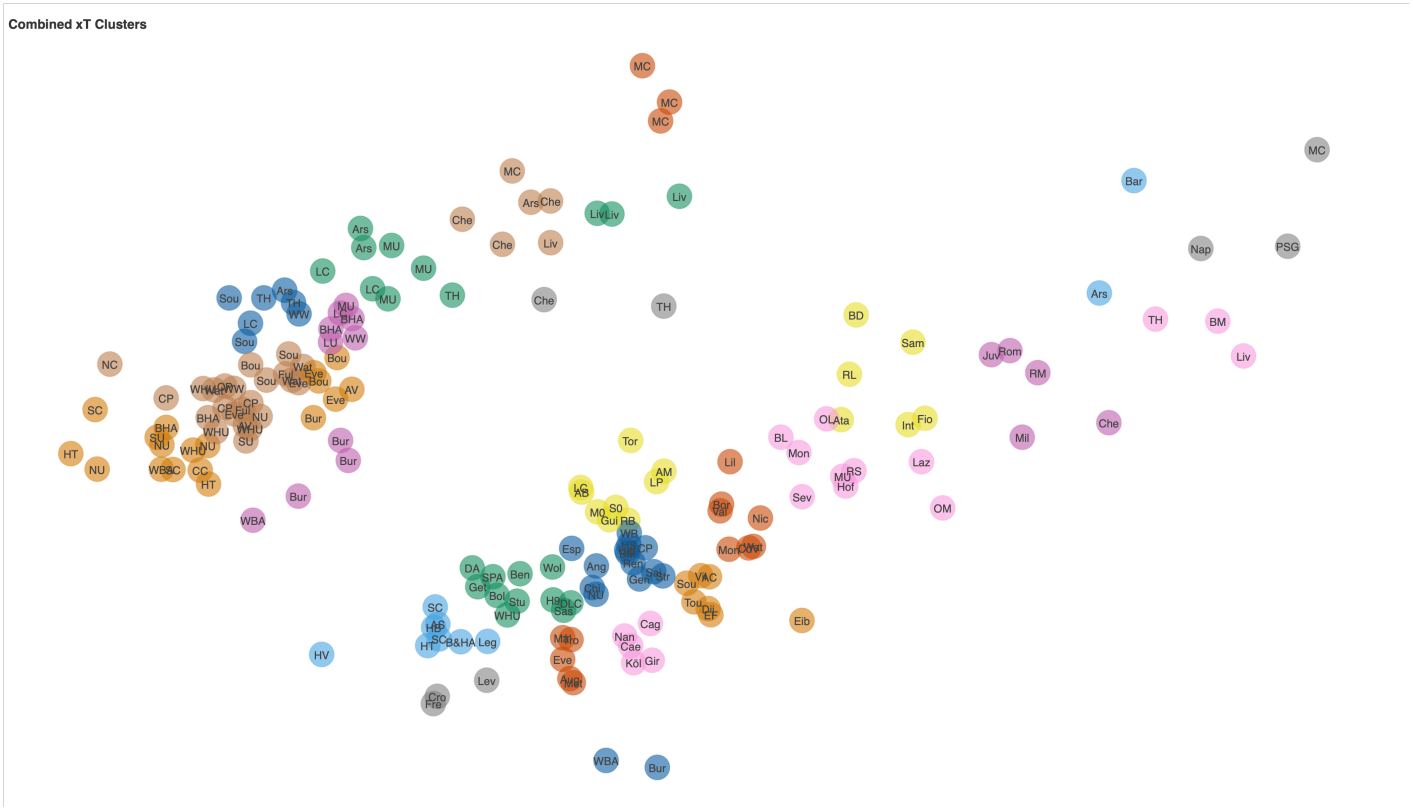

FIG. 21: Team-season clusters produced by K-means clustering following principal component analysis in two dimensions. Opta supercluster resides along the top-left diagonal, whilst the Wyscout supercluster resides along the bottom-right diagonal. Additional context for Opta clusters can be found in TABLE XII.

exploit in 2019/20.

Delta xT presents a novel application to indirectly measure and intuitively analyse the blindspot of modern quantitative football analysis, enabling teams to not only systematically identify points of weakness in opposition defences but also identify positions of vulnerability within one's own team and subsequently measure the impact and success of recruitment, as explored in detail for the two Manchester clubs.

5. Team Similarity Clustering

Each offensive xT and defensive Delta xT heatmap provides an $M \times N$ vector characterising a team's offense and defence per season. End-users can configure the $M \times N$ resolution in the xGils package, where we also make the xT and Delta xT vectors for both Opta and Wyscout datasets publicly available at high resolution (12×18) and plotting resolution (6×8) to enable reproducibility of our team-based applications without the need to acquire expensive Opta data. *At high resolution, each team may therefore be characterised by combining xT and Delta xT vectors to produce a high-dimensional,* 432*-length (*$12 \times 18 \times 2$*) vector per season.*

As highlighted with Barcelona's ill-fated transfer of Philippe Coutinho, identifying and acquiring talent is not enough to guarantee future on-the-pitch success. Integration of the new player into their new team is essential. A player asked to play a similar role in a similar team is more likely to repeat their previous high-performance levels than if asked to perform a different job in an unfamiliar system. As an extreme example of this point, a goalkeeper would never be transferred to a new club and asked to play centre forward. In our final application, we utilise our characterising vectors to cluster similar teams to offer an additional layer of rigour to our systematic recruitment analysis.

To explore team similarity at a more interpretable dimension, we apply principal component analysis to produce two orthogonal eigenvectors that maximise the variation from the high-dimensional characterising vectors to enable a projection onto two dimensions for clustering and visualisation. In two dimensions, the curse of dimensionality that would have degraded the performance of K-means clustering in 432 dimensions is not an issue. We proceeded with K-means clustering on both Opta and Wyscout datasets to produce $K = 28$ clusters for a combined 178 team-seasons: 80 from Opta — four seasons for each of the Premier League's 20 clubs — plus 98 from Wyscout, representing 98 teams spread over five European leagues for 2017/18. Clusters are visualised in FIG. 21 with complementary cluster context for Opta clusters provided in TABLE XII. The purpose of including Wyscout data here was to produce a "proof in the pudding" validation of the data quality assessment in section III that Wyscout's data is of such poor quality that it essentially represents a different sport to the one being more faithfully captured by Opta. The choice of $K = 28$ was motivated to produce average cluster sizes of 6 or 7, thus making it possible for a cluster to contain five seasons (four from Opta, one from Wyscout) for the same Premier League team with room to spare. Rather than the 20 2017/18 English Premier League clubs represented by Wyscout residing in the same clusters as their Opta counterparts, *there wasn't a single cluster that contained a mixture of teams from both Opta and Wyscout*. In fact, the two datasets produced two completely orphaned superclusters, with the Opta supercluster residing along the top-left diagonal of FIG. 21, and the Wyscout supercluster re-

siding along the bottom-right diagonal. Given both datasets were prepared and analysed in precisely the same way, this result shows the primary source of variation of the combined Opta and Wyscout data manifests from fundamental differences in the data collection methodologies between the two data vendors, rather than variation on the pitch between football teams.

Closer inspection of the Opta clusters in TABLE XII is fascinating. Whilst the underlying xT value surfaces are constructed using a combination of actual goals and synthetic goals to produce the location-based xG surface, the Bayesian xT value action framework ultimately only values actions that move the ball. Therefore shots and goals receive zero value, and thus the characterising vectors have no knowledge of how many goals a team has scored or conceded. And yet clusters 1,2,3 clearly represent title challenging teams, clusters 4,5,6 represent teams challenging for European places, clusters 7 and 8 loosely represent mid-table sides, and the remaining clusters represent teams often involved in a relegation dogfight. Manchester City and Liverpool, arguably the Premier League's two best teams over the past four seasons, each populate pure clusters from 2017/18 to 2019/20, illustrating their uniqueness prior to the 2020/21 season before they both join cluster 3. *To play against Manchester City or Liverpool from 2017/18 to 2019/20 would have been unlike playing against any other team in England*. We also observe the *progression* through time of Leicester City and Manchester United as both teams rise through the clusters into cluster 4, alongside the Spurs side that reached the Champions League final in 2018/19, and the Arsenal team that reached the Europa League final, also in 2018/19.

From the recruitment perspective, the similarity between Leicester City and Tottenham Hotspur across multiple clusters over different seasons could provide an additional layer of data-driven confidence in our Kane-Iheanacho example. Following the identification of Kelechi Iheanacho as a possible Harry Kane replacement at Spurs, there may be a better chance of Iheanacho integrating successfully at White Hart Lane given his experience at Leicester City rather than if a similar player was identified from, say, Huddersfield Town. However, this hypothesis, whilst plausible, requires testing.

This final application neatly builds on the four that came before to project high-dimensional representations that characterise how football teams attack and defend onto a 2-D page that's intuitive for a practitioner to interpret. It also serves as a final sanity check of our Bayesian xT framework: these clustered results really do make sense. For both recruitment and matchday strategy use cases, a valuable extension would be to add event data from additional leagues from a reliable source, like Opta. To quantitatively support the transfer of Harry Maguire from Leicester City to Manchester United based on his relevant skillset and experience, and subsequently measure his impact in Manchester United's defence and the success of his recruitment, is satisfying. But Maguire was no diamond in the rough. The true value of the techniques developed within this work is realised when the systematic scouting scales beyond what a human, or even a team of humans, can achieve. Increasing the geographical coverage of the data would also be of strategic benefit to teams playing in international competitions, like the Champions League and the Europa League. Applying the team similarity analysis would enable the comparison between unknown opponents drawn in the early stages of the competition to be likened to more familiar domestic teams, where there is already a body of institutional wisdom within the club of how to play against their familiar foes.

Cluster #	Opta Cluster Contents
1	Manchester City (**MC**) 2017/18, 2018/19, 2019/20
2	Liverpool (**Liv**) 2017/18, 2018/19, 2019/20
3	Arsenal (**Ars**) 2017/18
3	Liverpool (**Liv**) 2020/21
3	Manchester City (**MC**) 2020/21
3	Chelsea (**Che**) 2018/19, 2019/20, 2020/21
4	Chelsea (**Che**) 2017/18
4	Tottenham Hotspur (**TH**) 2017/18
5	Tottenham Hotspur (**TH**) 2018/19
5	Arsenal (**Ars**) 2018/19, 2020/21
5	Leicester City (**LC**) 2019/20, 2020/21
5	Manchester United (**MU**) 2018/19, 2019/20, 2020/21
6	Manchester United (**MU**) 2017/18
6	Leicester City (**LC**) 2018/19
6	Brighton and Hove Albion (**BHA**) 2019/20, 2020/21
6	Leeds United (**LU**) 2020/21
6	Wolverhampton Wanderers (**WW**) 2019/20
7	Southampton (**Sou**) 2019/20, 2020/21
7	Tottenham Hotspur (**TH**) 2019/20, 2020/21
7	Arsenal (**Ars**) 2019/20
7	Leicester City (**LC**) 2017/18
7	Wolverhampton Wanderers (**WW**) 2018/19
8	Aston Villa (**AV**) 2020/21
8	Bournemouth (**Bou**) 2017/18, 2018/19
8	Everton (**Eve**) 2018/19, 2019/20, 2020/21
8	Burnley (**Bur**) 2019/20
9	Burnley (**Bur**) 2017/18, 2018/19, 2020/21
9	West Bromwich Albion (**WBA**) 2017/18
10	Bournemouth (**Bou**) 2019/20
10	Fulham (**Ful**) 2020/21
10	Watford (**Wat**) 2017/18, 2018/19
10	Southampton (**Sou**) 2017/18, 2018/19
11	Huddersfield Town (**HT**) 2017/18, 2018/19
11	Newcastle United (**NU**) 2018/19, 2019/20, 2020/21
11	Cardiff City (**CC**) 2018/19
11	Brighton and Hove Albion (**BHA**) 2017/18
11	Stoke City (**SC**) 2017/18
11	Swansea City (**SC**) 2017/18
11	Sheffield United (**SU**) 2020/21
11	West Ham United (**WHU**) 2017/18
12	Aston Villa (**AV**) 2019/20
12	Crystal Palace (**CP**) 2017/18, 2018/19, 2019/20, 2020/21
12	Everton (**Eve**) 2017/18
12	Fulham (**Ful**) 2018/19
12	Norwich City (**NC**) 2019/20
12	Sheffield United (**SU**) 2019/20
12	Newcastle United (**NU**) 2017/18
12	Watford (**Wat**) 2019/20
12	West Ham United (**WHU**) 2018/19, 2019/20, 2020/21
12	Wolverhampton Wanderers (**WW**) 2020/21

TABLE XII: Opta K-means cluster mappings for FIG. 21.

VI. SUMMARY

The opportunistic acquisition of Opta data before commencing analytical work proved instrumental in the project's success. Most quantitative football research in an academic setting uses Wyscout data out of necessity because it's far more affordable than Opta. But, as in most walks of life, there is no such thing as a free lunch for data. We found Opta's data to be more accurate, complete, spatially precise and significantly more straightforward to work with than Wyscout data. Furthermore, Wyscout's data contained two detrimental sources of bias: look-ahead bias stemming from deflected shots being disproportionately omitted from the dataset and the minutes played attribute being winsorised at 90 minutes. The former bias would have caused xG probabilities assigned to shots to be artificially high, and the latter would have caused many of our derived metrics, normalised per 90 minutes, also to be biased high.

Our novel Bayesian approach enabled domain expertise to be encoded within the xT value action framework via a synthetic shot prior, representing the shots players choose not to take because success would ordinarily be impossible. Additionally, the Beta-Binomial updating mechanism provided a way of keeping xT value surfaces up-to-date as football evolves, making our model fit for purpose within the industry.

The Premier League's most threatening players, such as Manchester City's Kevin de Bruyne and Liverpool's Trent Alexander-Arnold, were identified by aggregating the values of transient actions that move the ball, with action values calculated via our smooth Bayesian xT value surfaces. Recruitment candidates to reduce Tottenham Hotspur's fragile reliance on striker Harry Kane were systematically scouted via excess xG, our metric that objectively values shots and thus finishing ability. Our xG logistic regression model used to calculate excess xG, trained using a combination of real and synthetic shots, produced impressive model performance measures compared to other xG models within the literature, with a log loss of 0.253 and an AUC of 0.835, owing to our engineered game state and interacting angle-distance features and better data. Once scouted, cross-player comparisons were visualised via proportional area radars comprising our intuitive suite of xT-, xG-, and Elo-based metrics. An extension of this work would be to produce a more expansive set of position-specific metrics, including goalkeepers, integrating additional data types like optical tracking and training performance data.

The innovative Delta xT application allowed the indirect measurement of a team's defensive capabilities, solving for on-the-ball event data's contextual blindspot, enabling our systematic scouting approach to scale beyond what is humanly possible in both attack and defence. The natural extension of the work here would be to expand the geographical coverage beyond the Premier League, to realise the "Moneyball" ambition of the techniques developed. The Elo system developed to rate and rank player duels could similarly be applied to normalise relative league strength to support the expansion.

Finally, offensive xT and defensive Delta xT heatmaps, measuring a team's strengths and weaknesses in attack and defence, were combined to produce high-dimensional vectors characterising a team's playing style. Clusters of similar teams were produced by projecting the high-dimensional characterising vectors in 2-D and clustering via K-means. Our cluster analysis showed that playing against Manchester City or Liverpool from 2017/18 to 2019/20 would have been unlike playing against any other team in England, and that Manchester United and Leicester City's play styles have considerably evolved over the past four seasons. Further work in this area would test the hypothesis that players transferred from similar teams are more likely to integrate successfully into their new clubs. If true, this would be a valuable addition to the systematic scouting tool kit to support critical recruitment decision-making.

At the heart of the work has been an intention for the metrics and analyses to be intuitive to understand by domain experts and decision-makers at football clubs, supported by visualisations that are easy to interpret and software that's easy to use.

[1] American Soccer Analytics. What are expected goals?, 2017. https://bit.ly/3sSUSAW.

[2] Razia Azen and David V. Budescu. Comparing Predictors in Multivariate Regression Models: An Extension of Dominance Analysis. *Journal of Educational and Behavioral Statistics*, 2006.

[3] Razia Azen and Nicole Traxel. Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. *Journal of Educational and Behavioral Statistics*, 2009.

[4] Carl Bialik. The People Tracking Every Touch, Pass And Tackle in the World Cup, 2014. https://53eig.ht/2VqnXqN.

[5] Bloomberg. Man City's Big Winter Signing Is a Former Hedge Fund Brain, 2021. https://bit.ly/3tZcSZK.

[6] Andrew Clark, Kate Howard, Andy Woods, Ian Penton-Voak, and Christof Neumann. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLOS ONE*, 2018.

[7] Cohen, William and Ravikumar, Pradeep and Fienberg, Stephen. A comparison of string metrics for matching names and records. *Proc of the KDD Workshop on Data Cleaning and Object Consolidation*, 2003.

[8] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions Speak Louder than Goals: Valuing Player Actions in Soccer. *arXiv*, 2019.

[9] H. Eggels. Expected goals in soccer: Explaining match results using predictive analytics. *Masters Thesis, Eindhoven University of Technology*, 2016.

[10] Arpad E. Elo. *The Rating of Chessplayers Past & Present*. 1978.

[11] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association 84:414-420*, 1989.

[12] M. A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine 14:491-498*, 1995.

[13] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. 2003.

[14] Wen Luo and Razia Azen. Determining Predictor Importance in Hierarchical Linear Models Using Dominance Analysis. *Journal of Educational and Behavioral Statistics*, 2013.

[15] Nils Mackay. Introducing Possession Value Framework, 2019. https://bit.ly/3iqW1vU.

[16] Ian G Mchale and Stephen Davies. Statistical analysis of the effectiveness of the fifa world rankings. *Statistical Thinking in Sport*, 2007.

[17] Ian G McHale, Philip A Scarf, and David E Folker. On the Development of a Soccer Player Performance Rating System for the English Premier League. *Interfaces*, 2012.

[18] John Muller and Matthias Kullowatz. Goals Added Deep Dive Methodology. https://bit.ly/3xvwUfI.

[19] BBC News. Data experts are becoming football's best signings, 2021. https://bbc.in/38oo6Ot.

[20] Rogier Noordman. Improving the estimation of outcome probabilities of football matches using in-game information. *Masters Thesis, Amsterdam School of Economics*, 2019.

[21] Opta. Available Online, 2021. https://www.statsperform.com/opta-football/.

[22] Luca Pappalardo, Paolo Cintia, Alessio Rossi, and Paolo Ferragina. A public data set of spatio-temporal match events in soccer competitions.

Nature Scientific Data, 2019.

[23] Richard Pollard and Charles Reep. Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 1997.

[24] Karun Singh. Introducing Expected Threat (xT), 2019. https://karun.in/blog/expected-threat.html.

[25] Jon Socik. Twitter thread: Opta Vs Wyscout data quality, 2019. https://bit.ly/3C2345T.

[26] John Stanton. Expected goals: What are we learning from new metric used on Match of the Day?, 2017. https://bbc.in/3mFCk5U.

[27] Statsbomb. Available Online, 2021. https://statsbomb.com/.

[28] David Sumpter. *Soccermatics: Mathematical Adventures in the Beautiful Game*. 2016.

[29] David Sumpter. Using Markov chains to evaluate football player's contributions. 2017. https://bit.ly/3Alr9mK.

[30] David Sumpter. Twitter thread: Fake shots, 2020. https://bit.ly/3s1O57x.

[31] Four Four Two. Why there could be five more hours of action in the Premier League in 2019/20, 2019. https://bit.ly/3jtKQSn.

[32] N. Van den Hoek. Improving expected-goals models: Towards more accurate values for individual shots by considering more detailed information. *Masters Thesis, Jheronimus Academy of Data Science*, 2019.

[33] Maaike Van Roy, Pieter Robberechts, Tom Decroos, and Jesse Davis. Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP. *AAAI-20 Workshop on AI in Team Sports*, 2019.

[34] W. E. Winkler. The state of record linkage and current research problems. *Statistics of Income Division, Internal Revenue Service Publication R99/04*, 1999.

[35] Wyscout. Available Online, 2021. https://wyscout.com/.

[36] Derrick Yam. Attacking Contributions: Markov Models for Football, 2019. https://bit.ly/2VqxHlj.