Communications of the IIMA

Volume 16 | Issue 3 Article 3

2018

What Do Publicly-available Soccer Match Data Actually Tell Us?

Chester S. Labedz Jr.
Central CT State University, clabedz@gmail.com

Robert Schumaker
University of Texas at Tyler, rob.schumaker@gmail.com

What Do Publicly-available Soccer Match Data Actually Tell Us?

Cover Page Footnote

We thank Dr. George Stalker, Dr. A. Tomasz Jarmoszko and David Freeman for their early work with us on the shooting events database. We recognize the contribution of research funds arranged by Dr. Patty Root, interim business dean at Central Connecticut State University, in purchasing a database from Opta Inc. We thank the three anonymous reviewers who provided very helpful comments during the editorial process.

What do publicly-available soccer match data actually tell us?

Chester S. Labedz, Jr. (corresponding author)

Associate Professor of Management and Organization

School of Business

Central Connecticut State University

1615 Stanley Street

New Britain, Connecticut, USA

(401) 524-7711

clabedz@gmail.com

Robert P. Schumaker

Associate Professor of Computer Science

Department of Computer Science

University of Texas at Tyler

Tyler, Texas, USA

ABSTRACT

Media analyses of soccer statistics and game play has accelerated in recent years. This is evident in visual displays of ball and player tracking, average player locations and distances they run. These media depictions aim to be attractive, entertaining and informative for viewers. But are such statistics predictive of goal scoring and match outcome? To answer this question we review the sixty-four matches of the 2014 World Cup and examine nine common match statistics, and others, to evaluate their predictive value for goal production and match outcome.

Keywords: Soccer, World Cup, soccer metrics, goal prediction, match prediction

INTRODUCTION

Within the last 50 years, the study of soccer statistics has increased dramatically. These analyses can help to develop and test insights to improve player development, identify strategic weaknesses in opponents and promote organizational effectiveness through operations management practices. Because professional sports often involve substantial financial expenditures, those who manage competing teams have devoted increased attention to the analysis of match data, seeking a competitive advantage.

The statistical analysis of play in soccer began in England during the 1950s with study of the effect of passing tactics on goal scoring (Reep and Benjamin, 1966). It has expanded in recent years with increasing interest in computer-aided tracking of match developments (Sumpter, 2016). Media distribution companies have extended these analyses using visually attractive graphics and an array of descriptive statistics to entertain and inform consumers. However, the ultimate objective of soccer play is usually to outscore one's opponent. Some of these media creations may link to goal production, but for others the connection may be less obvious. The ultimate question becomes, what statistics can predict goal production and overall match success (Anderson and Sally, 2013).

We believe that current media graphics are not aligned with predicting goals and match success. Film and other data often are selectively highlighted after the fact to explain why a particular shot succeeded or failed. But do such data serve to forecast as well? We believe valid, testable theories predicting goal scoring, goal prevention and match outcomes await further study. The intent of this paper is to test the explanatory nature of current soccer statistics as they relate to goal production. By understanding the statistics used by media and the degree of their explanatory effectiveness, their contribution to understanding goal production should become apparent.

The structure of this paper is as follows. Section 1 is the Introduction. Section 2 contains the Literature Review and overviews of World Cup soccer, existing soccer analytics and prior research. Section 3 identifies Research Gaps. Section 4 introduces our Research Questions. Section 5 is our Model Construction. Section 6 contains Analysis. Section 7 offers our Conclusions and Section 8 Limitations and Opportunity for Future Research.

LITERATURE REVIEW

Soccer is the world's most popular sport, attracting fans and participants from more than 200 nations. Professional footballers play for approximately ten months a year and successful players can earn tens of millions of dollars annually.

Soccer is called *The Simplest Game* (Gardner, 1996) because to play it basically requires just a spherical (leather) ball and up to 22 footballers who "kick it around." A match is at least ninety minutes long and each team aims to score goals greater or equal in number to the goals it concedes. Goals are scored by propelling the ball, mostly by players' feet and heads, into a rectangular goal measuring 8 feet high by 8 yards wide and centered at their opponent's end of the field. Because skilled players can combine their activities in powerful and elegant feats, soccer is also called The

Beautiful Game. Early versions of the sport (a.k.a. association football) trace to England during medieval times with modern versions spread globally along British trade routes (Gardner, 1996).

FIFA and the 2014 World Cup

The Switzerland-headquartered Fédération Internationale de Football Association (FIFA) governs international soccer competition, and it conducted the quadrennial men's World Cup tournament across Brazil during summer 2014. Thirty-two nations qualified through regional matches to compete in the 64-match tournament. In total, the 2014 World Cup matches encompassed more than 6,000 minutes of play, with 112 different individuals scoring a total of 171 goals, an average of 2.67 goals per match.

FIFA-provided Data and Statistics

Throughout the 2014 tournament, FIFA provided an array of descriptive statistics (FIFA, 2014). Its attacking statistics focused on shot directionality, frequency and assists, while the passing statistics focused on pass length, frequency and success, offside (illegal) deliveries and success in crossing passes. FIFA provided defense statistics on attempted and successful tackles, clearances, opponents' shots saved or blocked and intercepted passes. It also reported player disciplinary statistics, fouls committed and any disciplinary cards incurred.

FIFA provided ten official documents for each match. These offer individual tracking-based statistics (distances covered, sprints and speed), pass completions (from-to, and by length categories), rosters and tactical line-ups (displaying starting and actual positional maps) and a summary match report. These offer a wealth of information for media and fan consumption.

FIFA's documents included nine sets of "heat maps" for each game. These cloud formation drawings depict players' actual locations on the pitch on a per second basis. Heat maps are provided for six 15-minute intervals of the match. At match end actual formation maps summarize the heat maps' data into corresponding average positions of each starting player and any substitutes who contribute meaningful minutes of action. Those summary actual locations appear as jersey-numbered circles, as do players' stated formations within their team's initial tactical line-up. Figure 1 provides examples (Germany, in match 61) of player heat maps, as well as the tactical line-up and a summary of actual positions of each player.

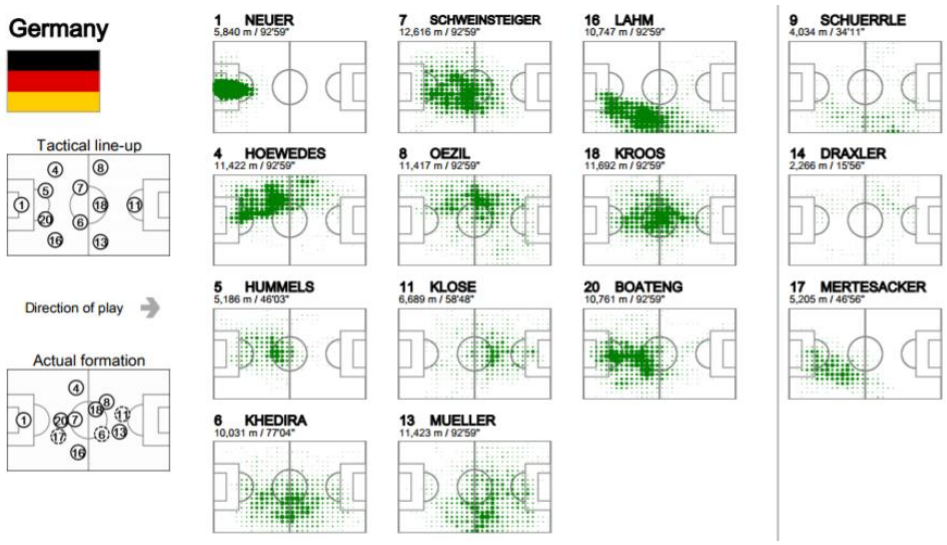

Figure 1. Examples of player heat maps, tactical line-up and actual formation maps.

Media-provided Statistics and Proprietary Data Sources

In addition to FIFA-provided data, pundits and scholars usually have additional sources available. Media companies that present live coverage of matches will supplement FIFA's statistics, often in real time. American viewers, for example, received additional data insight from broadcasters at ESPN in 2014 and Fox Sports in 2018. Third-party sport data firms also furnish proprietary information to league, team and other media clients. From such figures, fans may explore both player and team-based statistics for any of the World Cup matches. However, these data feeds conceal information needed to analyze causal relationships governing goal scoring or match success.

Prior Research

Scholars and media analysts have adopted a wide range of lenses and techniques in attempting to explain soccer behavior, performance and outcomes. We highlight a number of potentially predictive variables that these analyses suggest and permit us to test in our 2014 World Cup data.

Ball possession, Time of. Scholars have considered the effects of team possession on soccer match outcomes in more than a dozen cases (Collett, 2013). Collett found that ball hegemony, a team's preponderance of time in possession, did not consistently predict match success and urged reexamination of this metric's overall value. Perhaps a fresh look at this metric with respect to goal production might yield better results.

Pitch location (zone): shooter. Pundits also point to the importance of centrality and proximity to the goal as important factors in scoring. Bialkowski et. al. (2014) call attention to parts of the playing field from which players attempt lower probability shots. Lucey et al. (2015) focus on quantifying the value of shots taken, based on granular player tracking data recorded in 353 games in a 20-team season. They focus on the ten-second windows preceding shots including their launching points, direction of defenders from shooters, defenders' formations, attacking players'

pace, attacking context, and teams' relative motion. Jarmoszko, Labedz and Schumaker (2016) adopted the four field zones of shots (back, center, left and right) that are reported at media websites in assessing location effects of shot-taking on match outcomes.

Proximity: nearest defender. Soccer commentators regularly draw attention to the degree of open space that a scorer or passer enjoyed in the buildup of play to a goal, and the concomitant failure of a defender to mark his man more closely. Analyses of successful shots often explain that a defender has lost his mark; that is, he has not kept proximate enough to the opposing shooter (Lucey, Bialkowski, Monfort, Carr and Matthews, 2015).

Footedness of shooter and shot. Most players prefer to use a favored foot in kicking, especially shooting (Xavieur and Anjali, 2017; Haaland, 2002). Scholars and analysts point out scoring opportunities that disappear as a player attempts to shoot only with his dominant foot (Hoff and Haaland, 2002), while his defender works to force him to use his less-preferred one.

Nature of set pieces. Jarmoszko et. al. (2016) focused on nine variables that might be predictive of overall match outcomes (i.e., win, lose and draw). They reported that most lacked predictive value, but that set-piece attempts (in which play resumes after a stoppage whistled by the referee) did have significance values indicating statistical association with win, lose or draw match results in World Cup 2014.

Fastbreak Scoring or Longer Buildups. Jarmoszko et. al. (2016) also reported that fast break goals exhibited statistically significant association with teams' win, loss or draw results. In considering this finding, we note that it does not suggest causality. Whether fast break goals lead to wins, or whether a team's in-match winning position leads to increased successful fast break opportunities, was unanalyzed.

Using 1990 and 1994 World Cup data, Hughes and Franks (2005) observed that successful teams produced one-third more shots than did unsuccessful ones, and that about one in nine shots scored. From this, they concluded that teams would benefit from longer passing buildups that led to increased shot generation and total goals.

Shot Productivity and Efficiency. Pundits and scholars suggest that teams that attempt more shots on goal than their opponents will fare better. They also suggest that teams that are more selective in their shots on goal than their opponents will similarly fare better. Zambom-Ferraresi et. al. (2018) tested the effects of shot taking on teams' season-long point totals for matches won and drawn in European competition.

Player Position Variance. Bialkowski et al. propose a graphical representation method that updates players' relative roles from tracking data and use them to visualize formations. This approach deemphasizes attention to players' mean positioning, instead producing 5-minute smoothed depictions of role assignments.

Successful Pass Percentage. Zambom-Ferraresi et. al. (2018) explore the characteristics of the passing sequences that lead to goal scoring, including the final assist before a goal is scored. They examine the effects of total passes and successful ones (passing accuracy) in predicting teams' season-long league records.

Expected Goals. While the foregoing variables are examined for their role in predicting goal production, we note that sport enthusiasts recently have introduced xG as a possible candidate for predictive use. It has begun to appear in media discussions, usually without technical explanation. Rathke (2017) collects criticisms of various xG approaches, and explains that variations have not yet displayed reliable utility. Consequently we do not test the xG approach in this paper.

Other prior research has examined the use of Poisson models in soccer prediction (Rue and Salvesen, 2000; Koopman and Lit, 2015; Groll, Schauberger and Tutz, 2015). This stream of research is more rooted in academic statistics and is not commonly presented by FIFA or the media.

RESEARCH GAPS

Research has not assessed such a range of predictions across a tournament like the FIFA World Cup. Further, in many cases dependent variables of extant research extend beyond single-match results. We will test many of the ball-possession, shooting and other variables suggested by researchers, for any statistically significant associations with goal scoring and single match success.

Additionally, we review the mapping of players' nominal, actual and average positions on the pitch for explanatory value. Cloudlike heat maps, and the average position charts derived from them, indicate where, not why, a player usually operates on the field. We suspect that average positions mask the salience of actual positions data that identify the locations from which goal scoring shots, and the passes that led directly to them, occur. While attacking players cannot control the proximity of a defender, the choice of tactics employed based on proximity can directly impact goal production. We look to explore that gap in knowledge, tying the immediacy of player positioning to goal scoring.

RESEARCH QUESTIONS

In response to these gaps we ask the following research questions:

1. *What soccer-related statistics are linked to goal production?*

Research to date suggests that the predictive ability of a range of soccer statistics for goal production is limited. We test six factors that seem to be under-examined.

2. *What soccer-related statistics are linked to match success?*

Research similarly suggests that the use of soccer statistics for prediction of match success is also of limited value. To address this we test three variables of ball possession time, fastbreak scoring and pass success percentage.

3. *What role does player position variance have on assists or scoring?*

Players take their positions on the pitch as a result of a number of factors. These include their offensive and defensive responsibilities as assigned by their managers, plus match context and ball location. These influences may cause players to be distant (over spans of match time like 15-minute intervals) from assigned or average positions. By analyzing the variance of such distances we plan to examine the effect of tactical line-ups and actual formations on assists and goal production.

MODEL CONSTRUCTION

Overall, for data sources, we used Opta and FIFA data for in-match player positioning, ESPN for shots at goal, and match video coded by domain experts for shot-related variables (e.g.: footedness, zone location of players, defender distance, etc.) if not captured by those media sources.

Of the 1,688 shot attempts, 651 missed the goal, 404 were saved by goalkeepers, 401 were blocked by opponents and 45 struck the goal's woodwork. Of the 171 successful goals scored during regulation time, 12 were penalty kicks, 12 resulted from corner or free kicks, 11 were scored on fastbreak counter-attacks, 5 were own goals and 131 were scored in normal run of play. There was some overlap in these events, so we counted a maximum of 1,666 actual scoring opportunities. By our definitions, shots on goal include non-penalty goals scored, goalkeeper saves made, and woodwork struck. Shots at goal are shots otherwise blocked and those that miss the goal. For average positional data, we focused on play occurring only within the ninety minutes of regular time plus stoppage time, because the FIFA-provided interval maps cover only the first six 15-minute segments.

We extend the testing by Jarmoszko et. al. to predictor variables that are measured at nominal, ordinal or ratio levels. A chi-square approach will be adequate in some cases. We use additional techniques (logistic and ordinal regression) when the nature of the tested variables warrants it.

To answer our first research question on goal production, we analyzed six common media statistics used in World Cup 2014, covering 1,688 shot opportunities in 64 matches:

1. Pitch location (zone): shooter
2. Proximity: nearest defender
3. Footedness of shooter and shot
4. Nature of set pieces
5. Shot Productivity
6. Shot Efficiency

Pitch location (zone): shooter

Current positional descriptions that merely locate a player within or outside of the 18-yard penalty area rectangle are insufficiently granular. Our dimensional analysis expands field zone of shot from the original four that FIFA reports to the sixteen shooting zones depicted in Figure 2. We feel that this representation can better capture shooting and scoring difficulty from certain areas of the pitch and add depth to our analysis.

Specifically, the figure's rectilinear lines mirror the goal and penalty areas, and midfield line, as actually drawn upon the pitch. The first five of the numbered segments (zones) subdivide the goal area. Rectangular area 3 measures 8 yards wide and 6 yards deep; the other four small zones each are isosceles right triangles of side six. Segments 6 through 10 are carved similarly from a rectangle measuring 18 by 44 yards, minus the combined area of zones 1 through 5. In one example, central zone 8 measures 8 by 12 yards. Similarly, segments 11 through 15 carve up one half of the entire pitch, which in the twelve stadia used in FIFA Brazil 2014 had mean dimensions of 58 yards from goal to midfield and 76 yards across, minus already-mentioned zones 1 through 10.

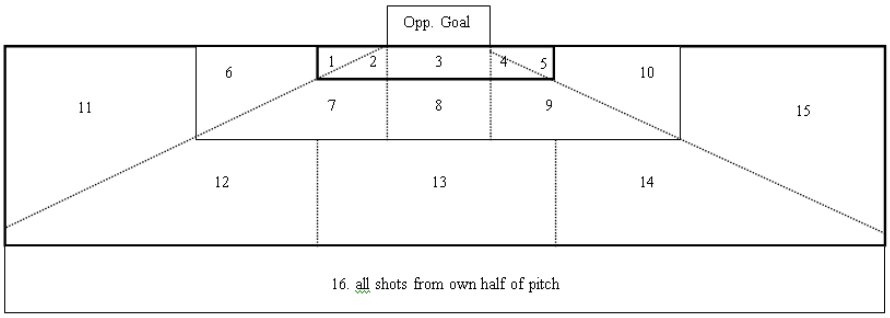

Figure 2. Shooting zones.

Proximity: nearest defender

The distance between the defender and shooter can have a definite impact on shot quality and success. A defender at a substantial distance may allow a shooter to improve shot quality whereas a defender in close contact can increase pressure and lead to errant shots. On the other hand, shooters often jump into or on top of defending players, especially to gain leverage in heading shots near to goal. In those cases their shots may prove disproportionately effective while their distances from nearest defenders will, of their own making, appear minimal (i.e., within 1 meter). For the purposes of our study we categorized nearest defender distance from the shooter as an ordinal variable; within 1 meter, 1 to 2 meters, and over 2 meters.

Footedness of shooter and shot

Shooters usually have a preference for which foot they like to use for goal production. If a defender is able to force the shooter to use their non-preferred foot, will this statistically impact goal production?

Nature of set pieces

Jarmoszko et.al. (2016) identified a significant association between "attack mode" (fast break, regular play, set piece) and match result. They reported that set-piece attempts (in which play resumes with a kick after a stoppage whistled by the referee) did have significant statistical association with win, lose or draw match results in World Cup 2014. We re-assess those data for their relationships to shot (rather than match) results.

Shot Productivity

Shots taken in the direction of ("at") goal reflect a base tier of team offensive production in soccer, shots "on" goal a second level, and actual goals scored the ultimate aim of attacking activity.

For shot productivity we analyzed shots on goal, non-penalty goals scored, goalkeeper saves, woodwork struck, blocked shots and other misses. We examine whether increased shot taking meaningfully impacts a teams' likelihood to launch threatening shots on goal.

Shot Efficiency
For shot efficiency, we analyzed the percentage of goals scored from attempts made.
To answer our second research question on predicting match success, we analyzed three common media statistics for their association:

1. Ball possession, time of
2. Fastbreak scoring
3. Pass success percentage

Ball possession, time of
Does the amount of time a team controls the ball correlate to match success? Excluding drawn matches and those decided by extra penalty kicks, we test whether possession differentials translate into goal differences.

Fastbreak scoring
Soccer is mostly a low scoring game. In the 2014 World Cup, opposing teams combined in only fourteen percent of matches for more than 4 goals. Goals often develop during short periods of possession. Media point to fastbreaks as such scoring opportunities. How does 2014 World Cup data support this emphasis?

Pass success percentages
Ball control is touted as an important component of team success. We examine the relationship between pass success and match results.

To answer our third research question on the predictive value of average positioning during matches, we compared average positions reported by third parties with players' actual positions, measured at the moments the goal-scorer released his shot and when the final assist was played to him, leading to that shot.

ANALYSIS

To answer our first research question of *what soccer-related statistics are linked to goal production*, we analyze the following variables.

Pitch location (zone): shooter and Proximity: nearest defender. Table 1 presents summarized and raw data on shooters' pitch locations (rows) and nearest defenders' proximity to shooters (columns) on goal scoring success. Zone locations correspond to Figure 2. (We exclude own goals and penalty kicks.) The Table displays two distinct biases in scoring success: shots launched nearer-to-goal, or in the center relative to goal, entered the goal in higher percentages. Thus declining proximity-to-goal of shooting led to declining likelihood of scoring: 31% (38/122) in adding zones 1-5, 13% from zones 6-10, and 2% from zones 11-15.

Per Figure 2, the zones may be characterized as left, center or right of the goal, not just more or less proximate to it. Relative to face of goal, shots launched from center zones 3, 8 and 13 entered goal nearly 11% (36/373) of the time, while left- (zones 1, 2, 6, 7, 11, 12) or right-side (zones 4, 5, 9, 10, 15, 15) shots scored 9.6% and 5.2% respectively.

The percentages we state in the text for varying likelihood of scoring associated with declining proximity of shooting (31, 13 and 2 percent) or with left-to-right angle (11, 9.6 and 5.2 percent) all can be derived by addition of values from clusters of table cells, followed by taking percentages.

At the end of Table 1, we explicitly state the composite goal scoring success percentages (9.5, 9.2 and 9.0 percent) relative to nearest defender proximity for all non-PK shots. To evaluate the performance of nearest defenders who were more- or less-proximate to shooters at time of shot, we used an independent samples median test. The null hypothesis was that the median values of goals was the same across the three categories of defender proximity stated atop Table 1. The resulting test significance was 0.881, so we were unable to reject H_0. Overall, nearest defender proximity did not matter in shooters' success. This result may be considered somewhat surprising.

Because shooting is a risk-reward decision by the shooter there may be game characteristics or game situations where there is in fact a difference due to defender proximity. The independent samples median test ignores lurking variables and only has good power for a statistically significant difference that is only determined by the three defender proximity zones defined in the previous paragraph. Therefore, failing to reject H_0 may be because there is no difference or because the test has low power due to lurking variables.

			Proximity:	[0,1]		(1,2]		> 2 yds.	
	TOTALS								
Zone	**Goals**	**Shots**	**Goals Pct.**	Goals	Shots	Goals	Shots	Goals	Shots
1	2	7	**29**	1	3	0	2	1	2
2	3	17	**18**	2	12	0	3	1	2
3	31	75	**41**	13	46	9	16	9	13
4	1	12	**8**	0	6	1	2	0	4
5	1	11	**9**	0	5	0	4	1	2
6	4	49	**8**	2	19	1	19	1	11
7	26	176	**15**	15	79	6	66	5	31
8 non-PKs	51	319	**16**	19	208	16	77	16	34
9	12	148	**8**	2	66	4	47	6	35
10	3	45	**7**	1	19	0	12	2	14
11	0	4	**0**	0	2	0	0	0	2
12	2	132	**2**	0	20	0	38	2	74
13	15	521	**3**	4	119	5	145	6	257
14	1	120	**1**	0	16	1	35	0	69
15	0	8	**0**	0	4	0	2	0	2
16	0	2	**0**	0	0	0	0	0	2
SUMS	**152**	**1646**	**9**	59	624	43	468	50	554

Proximity percentages		9.5%	9.2%	9.0%

Table 1. Shooting Success Rates by Zones and Defender Proximities.

Footedness of shooter and shot. Among players who took shots at goal in World Cup 2014, 153 shot exclusively with their right feet and 79 with their left feet. (We exclude headed shots and shots taken with other body parts.) 131 took at least two shots with each foot. Across 1,392 footed shots, the results (goals from shots) were: right (47/663), left (30/362), two-footed with right (14/180) and two-footed with left (29/187). More than seven in ten shots were taken by "one-footed" players, and over 60% of all kicked shots were right-footed. However, left-footed shots by "two-footed" players produced the greatest success rate (goals from 15.5 percent of shots attempted).

Nature of set pieces. We tested the relationships between attack mode shooting situations and shot result as the dependent variable. Table 2 presents the Chi-square test findings. With two (Set Piece and Field Zone) of these predictor variables, we can reject the null hypotheses that they are independent of Shot Result.

Element	Goals-to-Shots Percentages				d.f.	Critical	Max	p-value
	Direct	Corner	Fastbreak	None				
Set piece	4.6	12.5	26.0	8.6	3	26.372	7.815	0.000
	Left leg	Head	Right leg	Other				
Body part	10.7	11.7	7.3	---	3	7.459	7.815	0.59
	Left	Center	Right	Distant				
Field zone	14.1	19.5	7.5	2.4	3	106.59	7.815	0.000

Table 2. Relationships of Shot Results and Previously presented Predictive Variables.

In each test, we tested 1,666 shots. Maximum upper-tail critical value was determined to test probability of exceeding the critical value when $\alpha = 0.05$ and degrees of freedom = 3. Goals-to-Shots Percentages divide goals scored by shots attempted, cell by cell.

Shot Productivity and Shot Efficiency. FIFA's website breaks out shot data by team across the 2014 tournament. From it, we studied two measures of shooting productivity. By our data definitions, 606 shots on goal include non-penalty goals scored (157), goalkeeper saves made (404), and woodwork struck (45). To these, shots at goal add shots otherwise blocked (401) and those that miss the goal (651).

We first learned that added shot taking at goal did not meaningfully increase teams' likelihood to launch threatening shots on goal. A regression line of best fit indicates that added shot taking at goal increased shots on goal by less than one percent. Second, we focused on goals scored. Because penalty kicks were almost certain (93%) to become goals, and defenders' own goals are exceptionally unlucky, we excluded the 17 of them. Plotting the remaining 154 goals against the teams' numbers of shots on goal, we learned that added shots on goal decreased goals scored by a similar small percentage. We present these team-by-team results in the two plots of Figure 3;

regression lines and three sample teams' measures of shooting productivity, both scales, are marked.

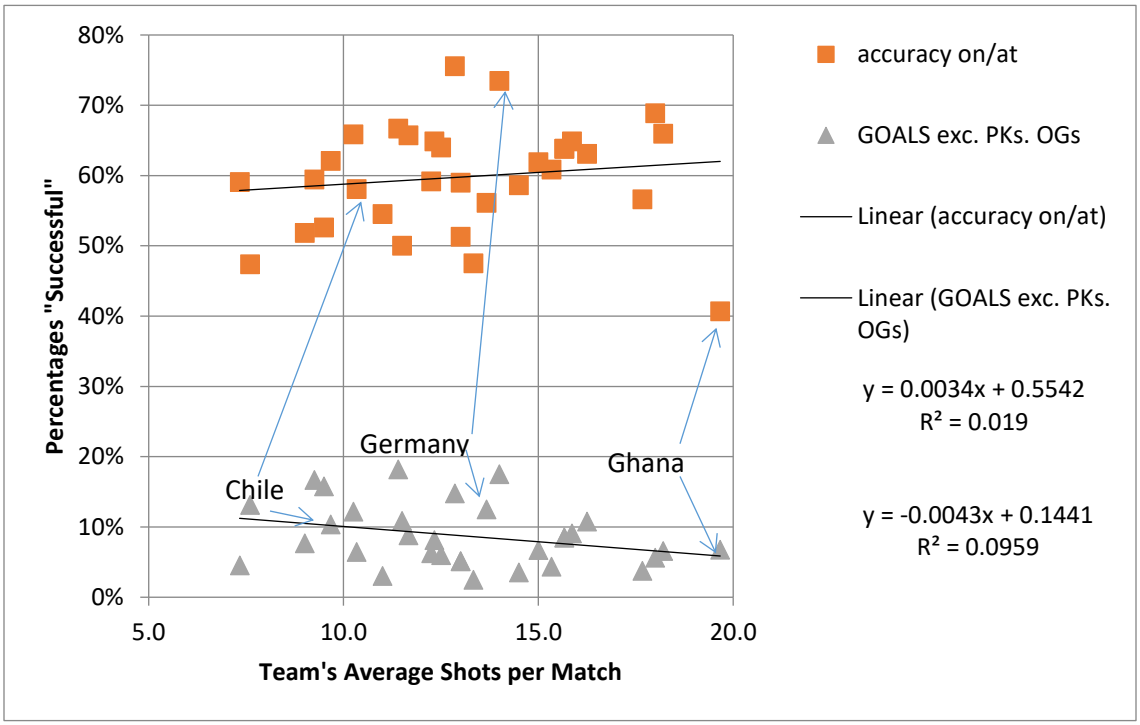

Figure 3. Tourney Measures of Shot Productivity, by Team.

These weak correlations do not suggest (at the tournament level) that taking more shots per match led to more shots actually on goal, nor that more shots on goal led to more goals scored. We next turn to comparisons at the match-level of shots taken by opponents in each of the tournament's 64 matches.

At the match-by-match level of analysis, do increasing levels of shots forecast greater goal scoring? Do they predict winning, or at least drawing, the match? Figure 4 graphically depicts shot taking efficiency (and profligacy). In it, the height of each stacked bar line displays the shots taken by paired competitors. The FIFA-designated home team always appears as a blue segment and its opponent in red. (Red bars appear only whenever the visiting team outshoots its opponent.) Two examples are provided. In one, the 8-shot blue home team, USA, defeats Ghana, which enjoyed an "excess shots" advantage of 13 attempts, while losing 2:1. Whether the focus is on winning or losing home teams, or on drawn matches, the shapes of these three outcomes clusters are rather similar.

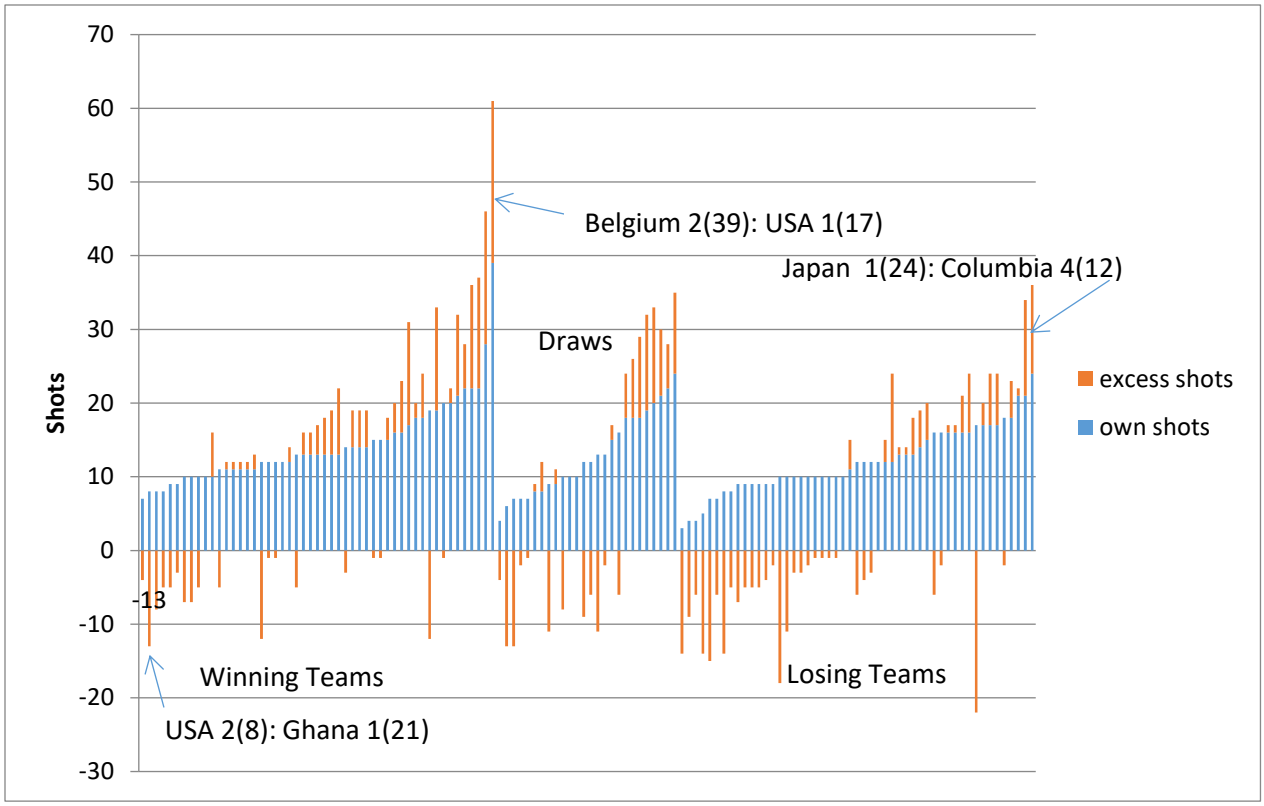

Figure 4. Shot Efficiency in Achieving Wins, Draws and Losses, by Match.

Logit analysis of these data confirms the lack of statistically significant relationships between quantity of shot taking, and degree of outshooting one's opponent, on winning (and drawing) a match. Thus may the USA defeat Ghana while taking only 38% of the match's shots, while Japan may shoot twice as often as Columbia but lose by a lopsided score. The logit analysis results appear in Table 3.

							95% CI for Exp(B)		
Dependent variable: match win									
		β	se	Wald	*df*	Sig.	Exp(B)	Lower	Upper
Shots		.075	.037	4.207	1	.040	1.078	1.003	1.158
	Constant	1.410	.522	7.285	1	.007	2.440		
Excess Shots*		.080	.038	3.451	1	.034	1.083	1.006	1.166
	Constant*	-.415	.274	2.297	1	.130	0.660		
Dependent variable: match win or draw									
Shots		.071	.038	3.451	1	.063	1.074	.996	1.158
	Constant	-.512	.521	.967	1	.325	0.599		
Excess Shots*		.059	.039	2.303	1	.129	1.061	0.983	1.145
	Constant*	.495	.268	3.406	1	.065	1.640		

* Because comparative (i.e., excess) shot figures and match goal differentials are based on values achieved by two teams, the table excludes all "away" team results as non-independent cases.

Table 3. Contribution of team's shot totals to match results.

In response to our first research question, we identify statistically significant associations with shot success (i.e., goal scoring) for both free kick set pieces and fastbreak scoring opportunities.

We conclude that none of the following variables exhibits a statistically significant relationship to shot success: proximity of nearest defender, shot productivity and shot efficiency.

To answer our second research question, *what soccer-related statistics are linked to match success*, we analyze the following variables.

Ball possession, Time of. Excluding thirteen drawn matches and games decided by penalty kick shootouts[1], winning teams enjoyed greater than 50% of time of possession in 38 of 51 contests. Figure 5 presents the goal differentials as predicted by excess possession in those 38 cases. We recoded 64 outcomes into win-draw and draw-loss variables and used the top halves of these pairs to ensure independence of cases. The resulting win-draw logit equation is 0.04 * minutes of possession minus 1.643, with a Nagelkerke R^2 value of 0.019. The corresponding findings for the draw-loss predictor were $\beta_1 = -0.0299$, $\beta_0 = 1.851$, and R^2 value of -0.147. We confirm Collett's finding that (in World Cup 2014) ball hegemony did not predict goal scoring differences (i.e., surplus goals).

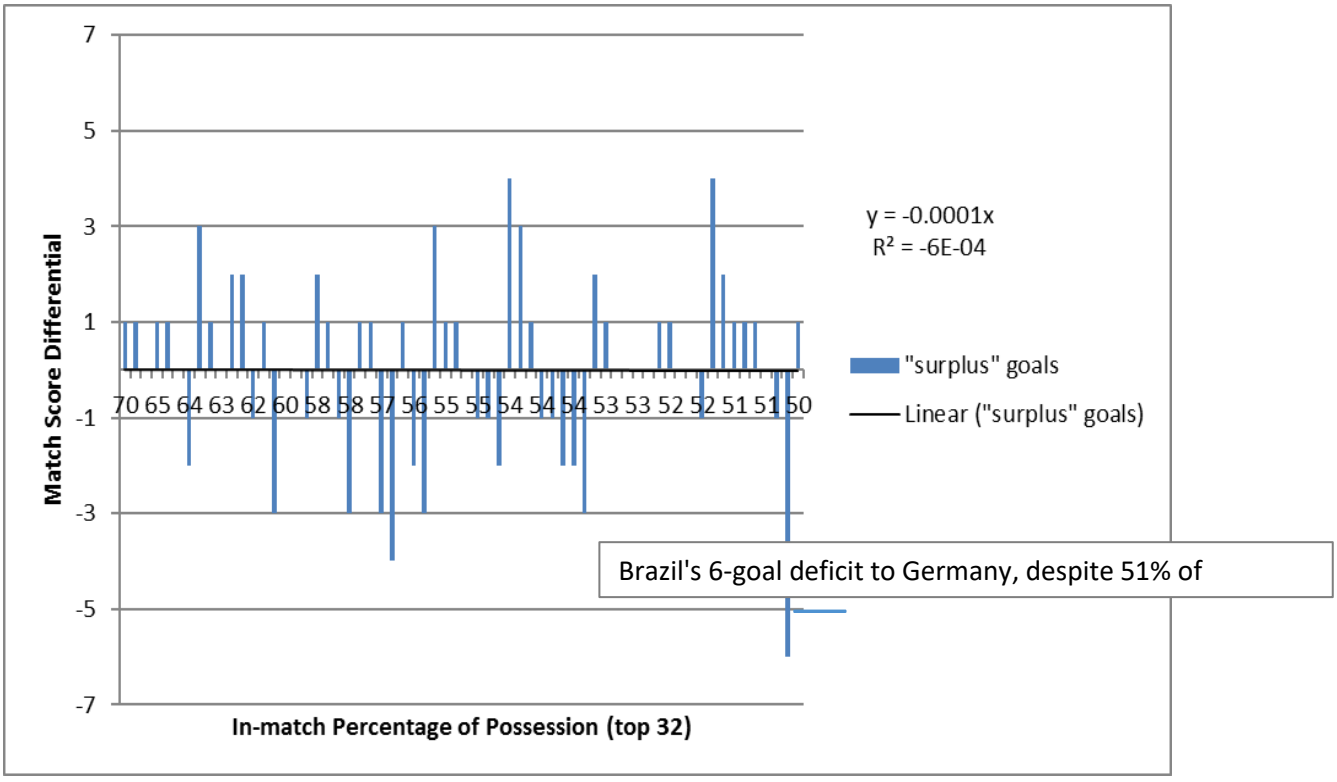

Figure 5. Effects of ball possession differentials on goal differentials.

[1] Excluded due to small sample size.

Fastbreak Scoring. We tested the relationship between minutes remaining in a match and the identity of the next team to score through a breakout goal that quickly overwhelms an opponent's defenses. Although the number of media coded breakout goals is only 11, confirming relationships do appear. Figure 6 indicates distribution of those goals. Nine fast break goals occurred in the last twenty minutes of matches, and only one trailing team (Portugal versus USA) scored any of them.

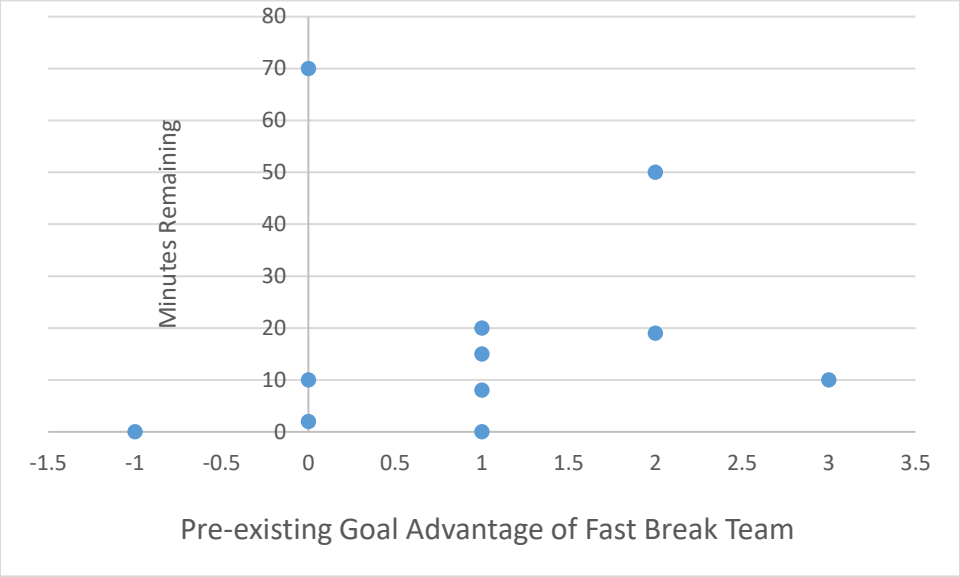

Figure 6. Fast Break goal scoring context.

The Wilcoxin-Mann/Whitney test reports no statistically significant relationship between the team trailing in a match and the team that scores the next goal. The chi-square analysis with one degree of freedom = 0.061, p = 0.806 fails to support the comeback theory. Using the independent samples Mann-Whitney U method, we next tested the null hypothesis that the distribution of match minute is the same across the distinct categories of fast-break and non-fast-break goals. The test reports an asymptotic significance (2-sided) of p=0.015 with z-score of 2.431, permitting rejection of H_0. The mean rank of fast-break goals is 121.09, while of other goals it is 83.59. Fast break goals do occur significantly later in matches.

Successful pass percentage. We turn to the association between teams' percentages of successful passes and their match results. Summary statistics converge (variance < 0.006) for each outcome near the overall population means (0.7507), indicating that the range of 128 passing success percentages does not improve ability to predict match outcomes. Ordinal regression (Chi-square = 0.861, Sig. = 0.353) confirms the absence of model fit.

Figure 7 presents the distribution o. At the match-by-match level of analysis, do increasing percentages of completed passes predict winning, or at least drawing, the match? In it, the height of each bar line displays the completion percentages, grouped by match outcomes. Three paired examples are provided. Teams completed as many as 89% of their passes in matches they won or

as few as 58%. Or, as with England, a nation could complete as many as 86% of its passes and still suffer defeat. Whether the focus is on winning or losing home teams, or on drawn matches, the shapes of these three completion ranges are rather similar.

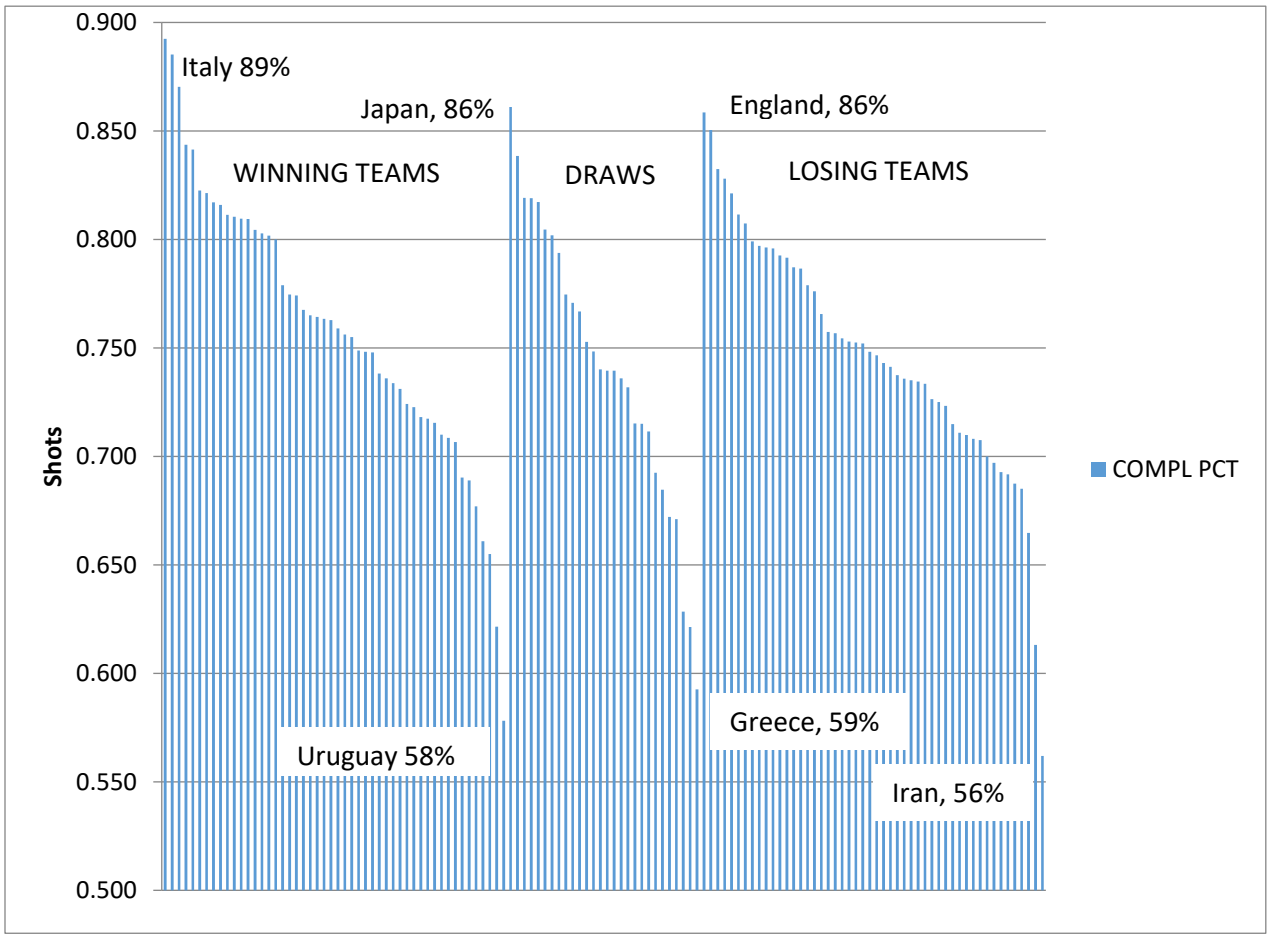

Figure 7. Effect of teams' passing success on match-by-match results.

In response to our second research question, we conclude that neither time of ball possession (a.k.a. ball hegemony) nor percentage of successful passes exhibits a statistically significant relationship to match success. We build u pon our previous finding relating to fastbreak scoring opportunities. We find that fastbreak goals are scored disproportionately late in matches and disproportionately by teams that are tied or leading the match when the fastbreak goal is scored.

To answer our third research question of *what role does player position variance have on assists or scoring*, we analyzed player positioning.

Like Bialkowski and her colleagues (2014), we challenged the attention paid to players' mean positioning. We examine the goodness of fit between the mean positions provided in FIFA's 15-minute actual formation summary charts and the instantaneous positions from which goal scorers ("S") scored and their helping teammates ("H") passed to them.

On Figure 8 for example, we superimpose the actual positions of scorers (marked by "x") or helpers (by dots) for Germany's first, fourth and fifth goals in its victory over Brazil. We used a digital ruler, field markings and a protractor to estimate distances. Thus, Toni Kroos has moved (from his average position in the first 15 minutes of match 61) more than 40 yards, both towards Brazil's goal and towards his opposite sideline, in order to strike the corner kick that results in Thomas Müller's 11th-minute goal.

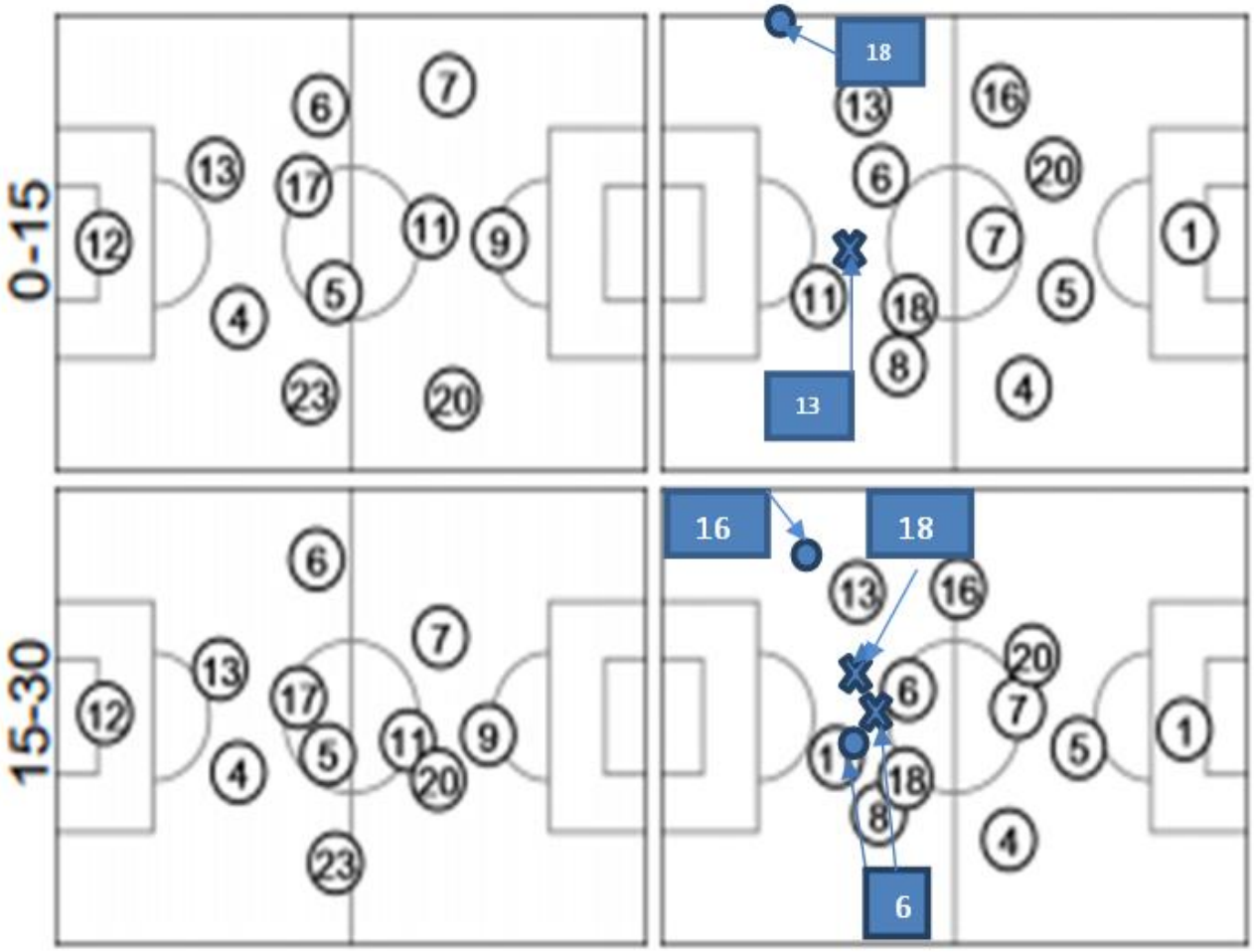

Figure 8. Average versus actual positions of three German players on four goals against Brazil.

Table 4 presents the variance in yards between the 15-minute average positions and the actual positions of German scorers as they shot and of helpers as they made the final pre-goal passes (assists) in match 61. Δx indicates those positional variances in yards, along the axis that connects the goals. Δy indicates positional variances along the axis that connects the sidelines.

Goal Scorer (Helper)	Scorer Δx	Scorer Δy	Helper Δx	Helper Δy
1: 13 Müller (18 Kroos)	-27	-24	-42	45
4: 18 Kroos (6 Khedira)	-27	12	-27	-9
5: 6 Khedira (8 Őzil)	-30	-3	-30	15

Table 4. Differences in yards between players' actual and average positions (Germany sample).

Table 5 compiles these statistics of variance from average positions across all matches of the tournament. We converted to absolute values to preclude cancellation of values as opponents attack in opposite directions. Thus these summarize direction-neutral vectors connecting six 15-minute average positions to actual positions in all 154 non-penalty and non-own goals.

Descriptives	Scorer Δx	Scorer Δy	Helper Δx	Helper Δy
Minimum	3	0	0	0
Maximum	71	27	68	53
Mean	**27**	**10**	**25**	**10**
s.d.	15	6	20	13
Mode	25	9	0	0

Table 5. Differences in yards between players' actual and average positions (entire tourney).

Several of the Δx statistics might at first surprise. On average, scorers shoot for goal (and score) about 27 yards closer to their opponent's goal than their circled average positions. On average, scorers drift 10 yards across the field to attain scoring position. Helpers exhibit similar Δx and Δy repositioning from average positions, and the standard deviations of their wanderings along both axes are even greater than scorers'. However, closer analysis of scorers' (x,y) shot-taking coordinates finds them in 54 cases at the end of free-kick, corner kick or fast break goals, events which may quickly (but only briefly) take them out of their average positions. This explanation is consistent with the breakout observations of Lucey et al. (2015) and Jarmoszko and colleagues (2016). Similarly, individual analysis of helpers' variances finds 19 cases in which they executed corner kick assists, briefly stationed far from their average field positions. In sum, such average positions data contribute little to understanding the geography of actual goal scoring.

In response to our third research question, we sampled maps indicating average actual positions during a match (divided into six fifteen-minute intervals) and compared these to coordinates of players' actual positions as goals were created. We conclude that average maps often indicate player positioning that is far removed from actual positions as goal scorers shoot and helper teammates collect assists on their own goals.

CONCLUSIONS

We built first on Jarmoszko's analysis of FIFA World Cup 2014, but reexamined five characteristics of play not for their relationships to match outcomes but more precisely to goal

outcomes. These consist of pitch location of shooter, pitch location of nearest defender, nature of set piece, footedness of shooter and shot, shot productivity and shot efficiency. We found that, as to both free kick set pieces and fastbreak scoring opportunities, causal relationships to goal scoring were evident.

Perhaps we should not be surprised that time of possession does not help to explain match outcomes. After all, some club sides and national teams have enjoyed tremendous successes by NOT possessing the ball very long. For example, Swiss and Italian systems (*verrou* and *catenaccio*, respectively) and the Inter Milan club sides featured resolute defending and then fast counterattacking surges once they recovered ball possession. Similarly, numerical advantages in shots at or on goal often do not favor winning sides, and low scoring wins by teams that manage to make one or a few shots count occur regularly.

Jarmoszko et al. focused on the association of fast breaks and match outcomes, and Lucey et al. highlight them too. Yet these sorties by definition count for very little in time of possession. What's more, they may actually *reflect, rather than promote*, already-favorable match outcomes. As trailing teams press to score an equalizing goal, they may leave themselves exposed to counter attack and to an increased score deficit.

We found that robust passing success rates need not lead to match success. The brevity (mean = 7.5 touches in 9 seconds) of scoring possessions reminds that brief outbursts can (and regularly do) overcome the prolonged possession passing routines of opponents that favor such buildups. It would be interesting to contrast all failed buildups with these statistics on successful ones, but we lack those data.

We studied for explanatory value the mapping of players' actual and average positions on the pitch. Cloudlike heat maps, and the average position charts derived from them, indicate where – but emphatically not WHY – a player usually operates on the field. In terms of explaining goal scoring and match outcomes, these diagrams underemphasize the exceptional moments and spaces in which goals often develop – far from individuals' 15-minute clusters and averages.

In a number of cases, however, descriptive statistics suggest promising relationships worthy of further research. We found evidence of a pattern in which already-leading teams added a goal to their leads during closing quarters of matches. We found a strong correlation between touches made and seconds elapsed in the buildup to non-penalty goals. We believe that the flexibility of "two footed" shooters contributes to their disproportionate goal scoring success. Data constraints have made variance testing unavailable in some cases. We would like to test the corresponding relationships between length of buildup (whether it is measured in touches or in seconds) and the distribution of *non-scoring* possessions.

Overall, 2.67 goals were scored per match in 2014 FIFA WC Brazil, excluding penalty kicks made after 120 minutes of play. That equates to one goal each 35 minutes. In soccer, goals truly are extraordinary events! Exhilarating goal scoring moments consummate attacks that consume most commonly (and on average) *six* seconds! In soccer, most goals truly are meteoric events!

Should the Beautiful Game expect better answers? Perhaps Big Data will identify more-hidden relationships.

LIMITATIONS, AND OPPORTUNITIES FOR FUTURE RESEARCH

FIFA's official documents present aggregated data, not raw second-by-second event data. Similarly, Opta Inc. also limited granular player position data to specific intervals. For the sake of research into competitive strategies and tactics, it would have been helpful to have access to such data after the tournament. As such, player positioning was manually measured and input from the positional interval maps. This time-consuming and manual measurement process may have introduced transcription or slight measurement errors into our models. When possible we chose to conservatively round toward values that minimized variance.

Our data, excluding video and purchased elements, are available for use by other researchers. We encourage others to similarly share their data to better understand, predict, play and manage the Beautiful Game.

REFERENCES

Anderson, C. & Sally, D. (2013) The Numbers Game: Why Everything You Know About Soccer is Wrong. Penguin Books: New York.

Bialkowski A, Lucey P, Carr P, Yue Y, Sridharan S, & Matthews I. (2014) . Large-scale analysis of soccer matches using spatiotemporal tracking data. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pp. 725-730. Shenzen.

Collett, C. (2013) "The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010." *Journal of sports sciences* 31.2, 123-136.

FIFA (2014). http://www.fifa.com/worldcup/archive/brazil2014/statistics/index.html (accessed August 3, 2017).

Gardner, P. (1996) The Simplest Game. The Intelligent Fan's Guide to the World of Soccer. Macmillan: New York.

Groll, A., Schauberger, G. & Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2), pp. 97-115.Hoff, J. & Haaland, E. (2002), Bilateral motor performance effects from training the non-dominant foot in competitive soccer players. In *Science and Football IV*, W. Spinks, T. Reilly and A. Murphy (Eds.), 288-293 (Routledge: London and New York).

Horton, M, Gudmundsson, J., Chawla, S. & Estephan, J. (2014), Classification of Passes in Football Matches using Spatiotemporal Data..arXiv:1407.5093 [cs.LG]

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of sports sciences*, 23(5), 509-514.

Jarmoszko, A.T., Labedz, C. & Schumaker, R. (2016) Toward a model of collective intelligence: Examining data from the 2014 soccer World Cup. *Collective Intelligence 2016 conference*, New York University.

Koopman, S. J. & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *J. R. Stat. Soc. A*, 178: 167-186. doi:10.1111/rssa.12042.
Lucey, P., Bialkowski, A., Monfort, M., Carr, P. & Matthews, I. (2015) "Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data." In *Proc. 8th annual MIT Sloan sports analytics conference*, pp. 1-9. 2014.
Rathke, A. (2016). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*. 12(2proc), S514-S529. Doi:https://doi.org/10.14198/jhse.2017.12.Proc2.05.
Reep, C. & Benjamin, B. (1968). Skill and chance in association football. Journal of the Royal Statistical Society, A, 131: 581–585.)
Rue, H. & Salvesen, O. (2000), Prediction and Retrospective Analysis of Soccer Matches in a League. Journal of the Royal Statistical Society: Series D (The Statistician), 49: 399-418. doi:10.1111/1467-9884.00243.
Sumpter, D. (2016) Soccermatics: Mathematical Adventures in the Beautiful Game. Bloomsbury Sigma: London.
Xaviour, G. & Anjali, O. (2017). "Comparison of preferred foot and non-preferred foot soccer technique of junior players." *International Journal of Physical Education, Sports and Health*. 4(4), 234-238.
Zambom-Ferraresi, F., Rios, V., & Lera-López, F. (October 2018) "Determinants of sport performance in European football: What can we learn from the data?" *Decision Support Systems*. 114 (October 2018), 18-28.