

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Goals in Soccer: Factors that Matter

Jan Vecer

CMStatistics 2016

December 10, 2016

Abstract

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

We present a state-of-the-art Bayesian statistical analysis for studying large soccer data that has recently become available. Our data was obtained from OPTA in the case of the English Premier League and from the official data providers of the German Bundesliga and the World Cup 2014. The aim is to identify factors that play a major role in the goal creation. The major factors fall in three broad categories, namely the speed, the precision, and the discipline.

Abstract

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Our approach is twofold. First, in the aggregate analysis section and the tracking section, we focus on the goal differential of the two teams, the difference between the home team and the away team scores. This approach is arguably more precise than focusing on probabilities of winning or losing the game as the corresponding models require non-linear transformation of data (in terms of logit or probit regressions). It also removes a high correlation of the tracking variables corresponding to the two teams. Second, we employ a Poisson regression when estimating the impact of precision (focusing on open crossing) on the scoring of an individual team. We study both the score differential and the absolute number of goals in the analysis of discipline and stoppages.

Abstract

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Our data analysis suggests that we are able to explain about 50% of the goal differential variability from the available game statistics. This gives a higher predictive power for the outcome of soccer games and identification of the factors that influence scoring. A detailed precision statistics explains about 30% of the variability, tracking statistics adds additional 17% and the discipline statistics the remaining 6%.

Models on Aggregate Data

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

In this section, we perform a simple linear regression analysis of the score differential

$$\text{Score Differential}_i = \text{Goals Home Team}_i - \text{Goals Away Team}_i$$

on factors that may explain it. This ignores variability of the individual teams that is analyzed in the subsequent sections. We have three sources of data that correspond to the English Premier League, the German Bundesliga and the World Cup 2014. As these are from different data providers, some of the variables are unique to the specific competition. Namely, OPTA data for the English Premier League give very detailed precision statistics that are missing in the German Bundesliga. In contrast, our German Bundesliga data have tracking statistics, which is missing in other competitions. The list of variables that is recorded in all competitions include the number of good passes, bad passes, corners, crosses, red cards, yellow cards and fouls.

English Premier League

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.42825	0.03771	11.35687	0.00000
ThroughDif	0.10167	0.00981	10.36646	0.00000
GPDif	0.00243	0.00028	8.53610	0.00000
BPDif	-0.01603	0.00326	-4.91397	0.00000
GCrossDif	0.02317	0.01117	2.07411	0.03820
BCrossDif	-0.05238	0.00524	-10.00500	0.00000
TacklesWDif	0.03723	0.00740	5.03031	0.00000
ClearancesDif	0.00983	0.00313	3.13495	0.00174
CornersDif	0.04965	0.01045	4.74944	0.00000
ThrowsDif	-0.01048	0.00491	-2.13549	0.03284
GDribblesDif	0.02406	0.00765	3.14497	0.00169
BDribblesDif	-0.01748	0.00829	-2.10891	0.03508
OffsideDif	0.05516	0.01356	4.06640	0.00005
RedDif	-0.72145	0.08849	-8.15247	0.00000
YDif	-0.08429	0.02182	-3.86332	0.00012

English Premier League

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Dependent variable: Dif

Residual standard error: 1.49083 on 1964 degrees of freedom

Multiple R^2	0.31093	Adjusted R^2	0.30602
----------------	---------	----------------	---------

$F(14, 1964)$	63.30262	p-value	0.00000
---------------	----------	---------	---------

Table : English Premier League regression. OPTA data, 2008–2014.

German Bundesliga

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.28337	0.04371	6.48336	0.00000
SprintKmDif	0.78151	0.08906	8.77557	0.00000
FastKmDif	-0.73116	0.06729	-10.86600	0.00000
DistanceDif	0.12233	0.01392	8.78724	0.00000
GPDif	0.00379	0.00026	14.59136	0.00000
BPDif	-0.03516	0.00346	-10.17049	0.00000
CrossesDif	-0.06537	0.00617	-10.59656	0.00000
CornersDif	0.02611	0.01179	2.21411	0.02697
RedDif	-0.37430	0.11299	-3.31282	0.00095
YDif	-0.10605	0.02617	-4.05275	0.00005

German Bundesliga

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Dependent variable: Dif

Residual standard error: 1.57334 on 1509 degrees of freedom

Multiple R^2	0.32543	Adjusted R^2	0.32141
----------------	---------	----------------	---------

$F(9, 1509)$	80.88799	p-value	0.00000
--------------	----------	---------	---------

Table : Bundesliga regression, matches from 2011–2016.

World Cup 2016

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
InPossDif	0.00003	0.00002	1.35160	0.18175
RankDif	0.00167	0.00070	2.38405	0.02042
CrossDif	-0.05451	0.01915	-2.84575	0.00611
TacklesDif	0.13645	0.07277	1.87494	0.06584
RedDif	-1.98592	0.56116	-3.53898	0.00080
YDif	-0.11853	0.15901	-0.74538	0.45905

Dependent variable: Dif

Residual standard error: 1.57405 on 58 degrees of freedom

Multiple R^2	0.36132	Adjusted R^2	0.29525
$F(6, 58)$	5.46871	p-value	0.00000

Table : World Cup 2014.

Sprints

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

A sprint occurs if a player achieves a speed value of at least 4.0 m/s for more than two seconds (>50 frames) and if he achieves a value of at least 2 frames of speed greater or equal to 6.3 m/s within this period.

Additionally the time span between the first frame of speed greater or equal to 6.3 m/s and the last frame of speed greater or equal to 6.3 m/s within this period must last for minimum 25 frames. A new sprint can only be counted after the speed is less or equal to 4.0 m/s.

Fast Runs

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

A fast run occurs if a player achieves a speed value of at least 4.0 m/s for more than two seconds (>50 frames) and if he achieves a value of at least 2 frames of a speed greater or equal to 5 m/s within this period and if it does not qualify as a sprint (speeds of 6.3 m/s are not achieved in an interval lasting more than a second). Additionally the time span between the first frame of speed greater or equal to 5 m/s and the last frame speed greater or equal to 5 m/s within this period must last for minimum 25 frames. A new fast run can only be counted after the speed is less or equal to 4.0 m/s.

Tracking Data

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

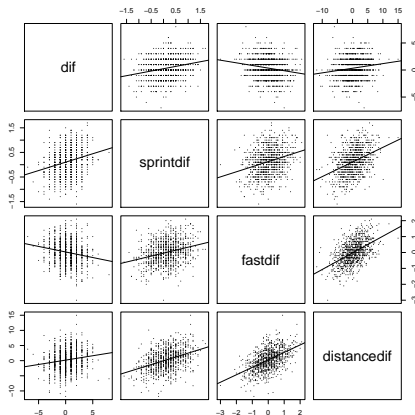


Figure : Visualization of score differential, sprint differential, fast run differential and distance differential with the corresponding regression lines.

Aggregate Model

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

$$\begin{aligned}\text{Score Differential}_i = & b_0 + b_1 \cdot \text{SprintKM Differential}_i \\ & + b_2 \cdot \text{FastKM Differential}_i \\ & + b_3 \cdot \text{Distance Differential}_i + \varepsilon_i.\end{aligned}\tag{1}$$

Aggregate Model

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.20688	0.05877	3.520	0.00045
SprintsKMDif	0.88727	0.11653	7.614	0.0000
FastKMDif	-1.01194	0.09289	-10.894	0.0000
DistanceDif	0.13858	0.01994	6.950	0.0000

Table : Impact of speed on the score differential.

Robust Regression

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value
(Intercept)	0.2132	0.0583	3.6557
SprintsKMDif	0.8027	0.1156	6.9420
FastKMDif	-1.0392	0.0922	-11.2750
DistanceDif	0.1535	0.0198	7.7569

Table : Impact of speed on the score differential, robust regression.

Bayesian Approach to Variability of the Teams

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

$$\text{Sprints Differential of a Team} \sim t_7(a, \sigma_u), \quad (2)$$

$$\text{Fast Run Differential of a Team} \sim t_7(b, \sigma_v), \quad (3)$$

$$\text{Distance Differential of a Team} \sim t_7(c, \sigma_w). \quad (4)$$

Bayesian Approach to Variability of the Teams

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

$$\begin{aligned}\text{Score Differential} \sim t_7(\mu &+ (a + u[\text{T1}[i]] - u[\text{T2}[i]]) * \text{Sprint D}[i] \\ &+ (b + v[\text{T1}[i]] - v[\text{T2}[i]]) * \text{Fast Run D}[i] \\ &+ (c + w[\text{T1}[i]] - w[\text{T2}[i]]) * \text{Distance D}[i], \sigma_y)\end{aligned}$$

Resulting Table

Teams	SprintDif	SD S.E.	FastDif	FD S.E.	DistDif	DD S.E.
Bundesliga	0.7854	0.1102	-0.9723	0.0978	0.1418	0.0127
1.FC Koln	0.8042	0.1635	-0.9843	0.1441	0.1479	0.0259
1.FC Nurnberg	0.7699	0.1774	-0.9482	0.1563	0.1490	0.0289
1.FSV Mainz 05	0.8208	0.1592	-1.0016	0.1358	0.1379	0.0238
1899 Hoffenheim	0.7729	0.1531	-0.9779	0.1423	0.1390	0.0238
Bayer 04 Leverkusen	0.7766	0.1511	-0.9626	0.1332	0.1432	0.0243
Borussia Dortmund	0.8279	0.1486	-1.0040	0.1352	0.1407	0.0239
Borussia Mgladbach	0.7583	0.1589	-1.0255	0.1358	0.1414	0.0234
Eintracht Braunschweig	0.7895	0.1772	-0.9837	0.1493	0.1438	0.0252
Eintracht Frankfurt	0.7245	0.1692	-0.9380	0.1414	0.1359	0.0247
FC Augsburg	0.7296	0.1660	-1.0098	0.1438	0.1422	0.0232
FC Bayern Munchen	0.9048	0.2197	-0.8671	0.1690	0.1452	0.0237
FC Ingolstadt 04	0.8022	0.1812	-0.9289	0.1511	0.1373	0.0254
FC Schalke 04	0.7672	0.1547	-1.0324	0.1509	0.1402	0.0234
Hamburger SV	0.8066	0.1555	-0.8623	0.1722	0.1419	0.0234
Hannover 96	0.7439	0.1604	-0.9805	0.1301	0.1384	0.0235
Hertha BSC	0.8006	0.1589	-0.9346	0.1401	0.1439	0.0239
SC Freiburg	0.7444	0.1866	-0.9803	0.1396	0.1395	0.0233
SC Paderborn 07	0.7400	0.1774	-0.9563	0.1463	0.1412	0.0250
SV Darmstadt 98	0.8064	0.1779	-0.9765	0.1492	0.1413	0.0252
SV Werder Bremen	0.8587	0.1773	-0.9860	0.1379	0.1435	0.0246
VfB Stuttgart	0.7635	0.1530	-1.0208	0.1485	0.1464	0.0239
VfL Wolfsburg	0.7654	0.1514	-0.9882	0.1324	0.1417	0.0232

Table : Impact of speed on the score differential, multilevel Bayesian model.

Bundesliga Sprint

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

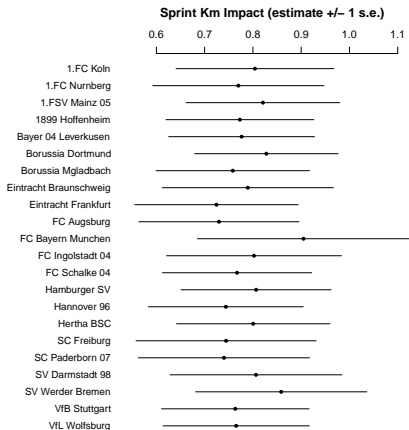


Figure : Estimated sprint coefficients for Bundesliga teams.

Bundesliga Fast Runs

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

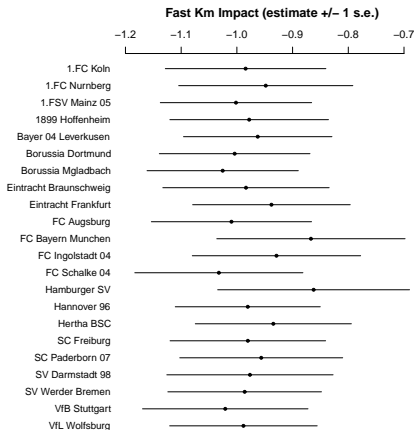


Figure : Estimated fast run coefficients for Bundesliga teams.

Bundesliga Distance

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

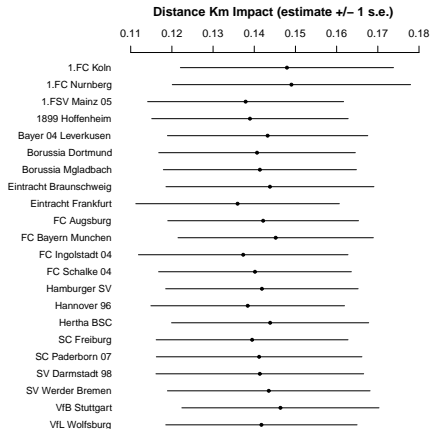


Figure : Estimated distance coefficients for Bundesliga teams.

Negative Impact of Open Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

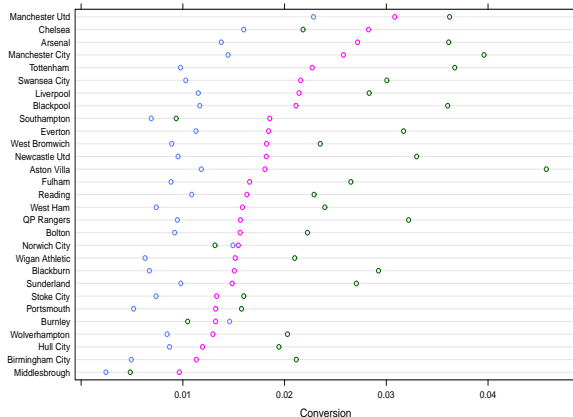


Figure : The fraction of open crosses (blue), final third entries (red) and outside the box shots (green) that results in a goal for individual attacking teams, EPL games in seasons 2008 – 2013.

Goal Model

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Goal model:

$$\text{Goals}_i \sim \text{Poisson}(\exp((\beta^I + u_{j[i]}^I) + (\beta^C + u_{j[i]}^C) \cdot \text{Cross}_i + \beta^H \cdot \text{Home}_i)), \quad (5)$$

where

$$u_j \sim N(0, \Sigma_u) \quad (6)$$

represents random effects.

English Premier League Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.2631	0.0628	4.1901	0.0000
OpenCross	-0.0170	0.0022	-7.7247	0.0000
Home	0.3841	0.0210	18.3115	0.0000

Table : Regression estimates for the English Premier League (10 seasons 2006–2016, 3800 games).

EPL Random Effects

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Groups	Name	Std.Dev.	Corr
Team	(Intercept)	0.31632	
	OpenCross	0.00721	-0.75

Table : Correlation matrix for EPL random effects.

EPL Team Crossing

Goals in Soccer: Factors that Matter

Precision:
Major
Negative
Impact of
Open Crossing

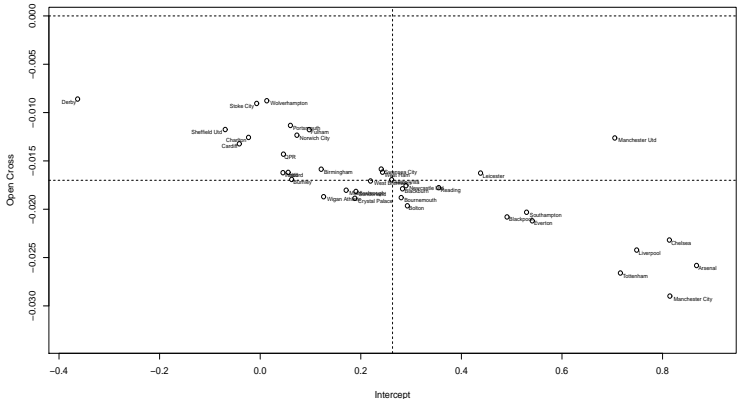


Figure : EPL teams, intercept and cross slope. The points follow normal distribution, having a negative correlation of -0.75.

EPL Team Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

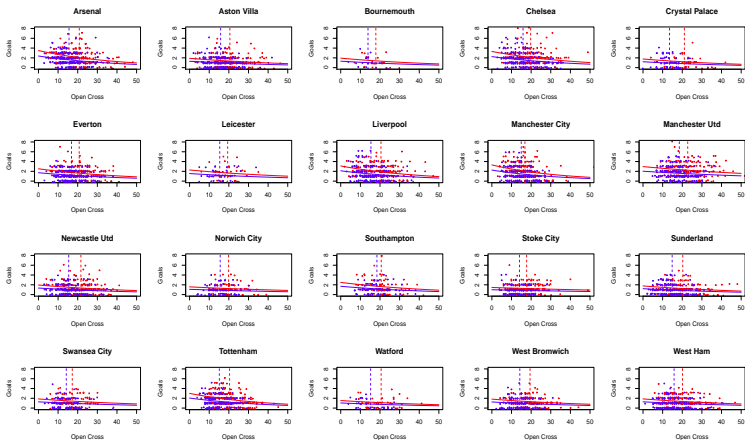


Figure : Goals versus open cross for individual EPL teams that played in the 2015/16 season.

Bundesliga Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.2909	0.0657	4.4399	0.0000
OpenCross	-0.0180	0.0040	-4.4687	0.0000
Home	0.2704	0.0262	10.3239	0.0000

Table : Regression estimates for the German Bundesliga (7 seasons 2009–2016, 2142 games).

Bundesliga Random Effects

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

Groups	Name	Std.Dev.	Corr
Team	(Intercept)	0.2688	
	OpenCross	0.0103	-0.44

Table : Correlation matrix for Bundesliga random effects.

Bundesliga Team Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

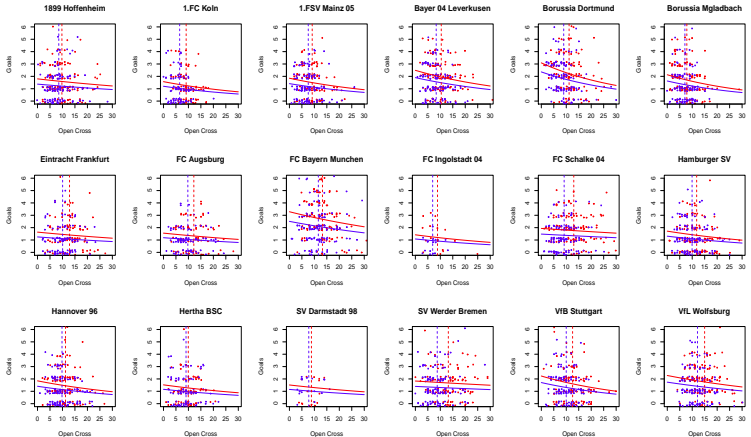


Figure : Goals versus open cross for individual Bundesliga teams that played in the 2015/16 season.

World Cup 2014 Crossing

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.643187	0.180940	3.554702	0.000378
Cross	-0.018784	0.009054	-2.074544	0.038029

Table : Regression estimates for World Cup 2014 (64 games).

EPL Model for Score Differential

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.6638	0.1260	5.2701	0.0000
Red1	-1.0824	0.1170	-9.2551	0.0000
Red2	0.7487	0.0957	7.8232	0.0000
Yellow1	-0.2430	0.0275	-8.8293	0.0000
Yellow2	0.0569	0.0252	2.2563	0.0241
Fouls1	-0.0015	0.0086	-0.1771	0.8595
Fouls2	0.0003	0.0081	0.0375	0.9701

Table : Impact of the discipline on the score differentials in the English Premier League.

EPL Home Team Scoring

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.6362	0.1814	3.5070	0.0005
Red1	-0.4546	0.0885	-5.1345	0.0000
Red2	0.2189	0.0529	4.1412	0.0000
Yellow1	-0.0689	0.0181	-3.8019	0.0001
Yellow2	0.0072	0.0157	0.4603	0.6453
Fouls1	-0.0016	0.0053	-0.2985	0.7653
Fouls2	0.0002	0.0058	0.0410	0.9673
Throws1	-0.0249	0.0030	-8.2138	0.0000
Throws2	-0.0030	0.0030	-0.9909	0.3217
SubIn1	0.1008	0.0302	3.3425	0.0008
SubIn2	0.1488	0.0312	4.7640	0.0000
Corners1	0.0014	0.0068	0.2054	0.8372
Corners2	-0.0235	0.0076	-3.0709	0.0021

Table : Impact of various stoppages on the scoring of the home team, English Premier League 2008-2013.

EPL Away Team Scoring

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.0310	0.2088	0.1486	0.8819
Red1	0.3993	0.0684	5.8358	0.0000
Red2	-0.2536	0.0768	-3.3021	0.0010
Yellow1	0.1325	0.0198	6.6824	0.0000
Yellow2	-0.0471	0.0188	-2.5009	0.0124
Fouls1	-0.0085	0.0062	-1.3551	0.1754
Fouls2	0.0017	0.0067	0.2575	0.7968
Throws1	0.0086	0.0034	2.5299	0.0114
Throws2	-0.0162	0.0036	-4.5001	0.0000
SubIn1	0.0731	0.0343	2.1350	0.0328
SubIn2	0.0163	0.0331	0.4920	0.6227
Corners1	-0.0090	0.0080	-1.1173	0.2638
Corners2	0.0099	0.0086	1.1424	0.2533

Table : Impact of various stoppages on the scoring of the away team, English Premier League 2008-2013.

Bundesliga Model for Score Differential

Goals in
Soccer:
Factors that
Matter

Jan Vecer

Models on the
Aggregate
Data

Tracking
Statistics

Precision:
Major
Negative
Impact of
Open Crossing

Discipline and
Stoppages

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.8342	0.1951	4.2756	0.0000
Red1	-1.1290	0.1502	-7.5149	0.0000
Red2	0.7228	0.1393	5.1893	0.0000
Yellow1	-0.2024	0.0394	-5.1389	0.0000
Yellow2	0.2043	0.0370	5.5259	0.0000
Fouls1	-0.0243	0.0108	-2.2499	0.0246
Fouls2	-0.0118	0.0101	-1.1650	0.2442

Table : Impact of the discipline on the score differentials in the German Bundesliga.