Major League Soccer Defensive Quality

In recent times the Major League Soccer has taken tremendous strides to improve the game. There has been a strong drive to bring in ageing stars such as David Beckham (31 years old), David Villa (32 years old), Kaka (32 years old). The premise of this strategy is that these stars, who are slightly past their prime, will be able to improve the fan base and the overall appeal of Major League Soccer. The majority of these players are attacking superstars, mainly because for fans, the most exciting part of the game always occurs in the scoring of goals. However games are not won by scoring a lot of goals but rather by scoring more goals than your opponent.

An often-neglected area of the game is improving the defensive quality of a team. Below is a table that shows the number of players per position and their salary information by position.

totals	number	money	avg	median
Defenders	160	$23,395,767.12	$146,223.54	$99,500.00
Midfielders	188	$68,831,483.00	$366,124.91	$108,450.00
Forwards	126	$44,172,764.12	$350,577.49	$125,000.00

As you can see defenders pale in comparison to their attacking teammates in regards to salary. As a result, I decided to dive into the defensive statistics of the top MLS teams (MLS teams in the Conference Semifinals). Not-surprisingly all of the western conference and all but one of the eastern semifinalist were the top defensive teams in the league:

West

Seattle Sounders - 36 goals conceded
Vancouver Whitecaps – 36 goals conceded
FC Dallas – 39 goals conceded
Portland Timbers – 39 goals conceded

East

Montreal Impact – 44 goals conceded
DC United – 45 goals conceded
New England Revolution – 47 goals conceded
Columbus Crew – 53 goals conceded (had the second highest goals scored with 58)

Objective of this project/exploratory analysis is to try and identify the key defensive factors that exist amongst the best defensive teams that are correlated to goals scored. Specifically, I will be looking for how opponent's goals scored are related to several dependent variables or defensive factors so that I can look at how well these defensive factors can predict the conceding of goals by a team.

METHODS

Study Sample / Data Screening

For my analysis, the data for this project is provided by OptaSports, a data-warehousing company that specializes in collecting professional soccer statistics. The original data consisted of all 20 teams regular season games (304 observations). As mentioned above, I studied the 8 teams that have reached the semi-finals with a total of 108 observations.

Variables:

Homegoals: These are goals scored by the team analyzed. As mentioned above it is important because in order to win a game, a team must outscore the opponent.
AwayGoals: These are goals scored by the opposition. These are extremely important as the less goals a team concedes the easier it is to recognize a strong defensive team.
AwayAttempts: These are the amount of shots an opposition team attempts. These are important because an attempt can lead to a goal scored.
AwaySOG. These are the amount of shots on goal/target. A shot on goal can result in a goal scored whereas a shot off target has a 0% chance of scoring.
HomeBlocks: These are the amount of blocked shot attempts by the analyzed team.
Away Corners: These are the amount of corners which can lead to goals by the opponents.
Away Cross: These are the amount of crosses by the opponents which can lead to goals.
Away Offsides : The amount of times the defensive unit of the analyzed team plays the opponents offside (eliminates an opportunity to score a goal).
Home Fouls: The amount of fouls performed by the analyzed team. This is important as a foul can lead to a free kick/opportunity to score a goal. So generally a team will want to minimize fouls.
Home Yellow: the amount of yellow cards awarded to the analyzed teams.
Home Red: The amount of red cards awarded to the analyzed team. With a red card there are 1 less (per red card) players on the field, so it becomes harder to defend and attack.
Home duels/away duels: This is a coefficient that compares the amount of duels won by the analyzed team vs opposition.
Home Tackles: The amount of tackles won by the analyzed team.
Home Clearances: The amount of times the analyzed team successfully clears the ball out of their danger zone (18 yard box)
Away PassPct: The percentage of passes the opponent makes with the passes. A lower number is desirable as that means the analyzed team is intercepting more of the opponents passes.

Exploratory Analysis

With 14 variables and 1 response variable this dataset is a good fit for a principle component analysis regression model.

I started first by performing a scatter plot and a correlation procedure to look for correlations within the variables. Surprisingly, there weren't many strong correlations amongst the variables. For each variable there were only about 2 to 3 strong correlations with another variable. Such as away attempts + away shots (0.63) and away cross + away corners (0.6512) .

Correlation Matrix										
	hgoal	aattempts	asog	hblock	acorner	across	aoff	hfoul	hyel	hred
hgoal	1.0000	0.0095	0.1083	-.1208	0.0721	0.2135	-.0663	0.0327	-.0252	-.0556
aattempts	0.0095	1.0000	0.6300	-.0637	0.4409	0.3472	0.1062	0.0038	0.0379	0.0815
asog	0.1083	0.6300	1.0000	0.0412	0.2992	0.1517	0.1845	0.1087	0.1588	-.0540
hblock	-.1208	-.0637	0.0412	1.0000	0.0150	-.0802	0.1025	-.0611	-.0980	-.1042
acorner	0.0721	0.4409	0.2992	0.0150	1.0000	0.6512	0.1104	-.1202	0.0746	0.0273
across	0.2135	0.3472	0.1517	-.0802	0.6512	1.0000	0.1374	-.1417	0.0935	0.0157
aoff	-.0663	0.1062	0.1845	0.1025	0.1104	0.1374	1.0000	-.0042	-.0973	0.1587
hfoul	0.0327	0.0038	0.1087	-.0611	-.1202	-.1417	-.0042	1.0000	0.3091	0.0185
hyel	-.0252	0.0379	0.1588	-.0980	0.0746	0.0935	-.0973	0.3091	1.0000	-.0035
hred	-.0556	0.0815	-.0540	-.1042	0.0273	0.0157	0.1587	0.0185	-.0035	1.0000
htackle	-.0145	0.0015	-.1110	0.0494	-.0104	-.0493	-.0339	0.0432	-.1153	-.0442
hclear	0.0492	0.2128	-.0506	-.0521	0.4640	0.6316	0.0116	-.0110	-.0192	-.0415
apasspct	0.2603	0.1571	0.1980	-.1388	0.0911	0.2783	0.0942	0.0118	0.0282	0.1631
Duels	-.0241	-.1172	-.2426	0.0889	-.0029	0.0321	-.0352	-.2007	-.2297	-.0248

Principal Component Analysis

After running my first principal component analysis, I got the below statistics and eigenvalues:

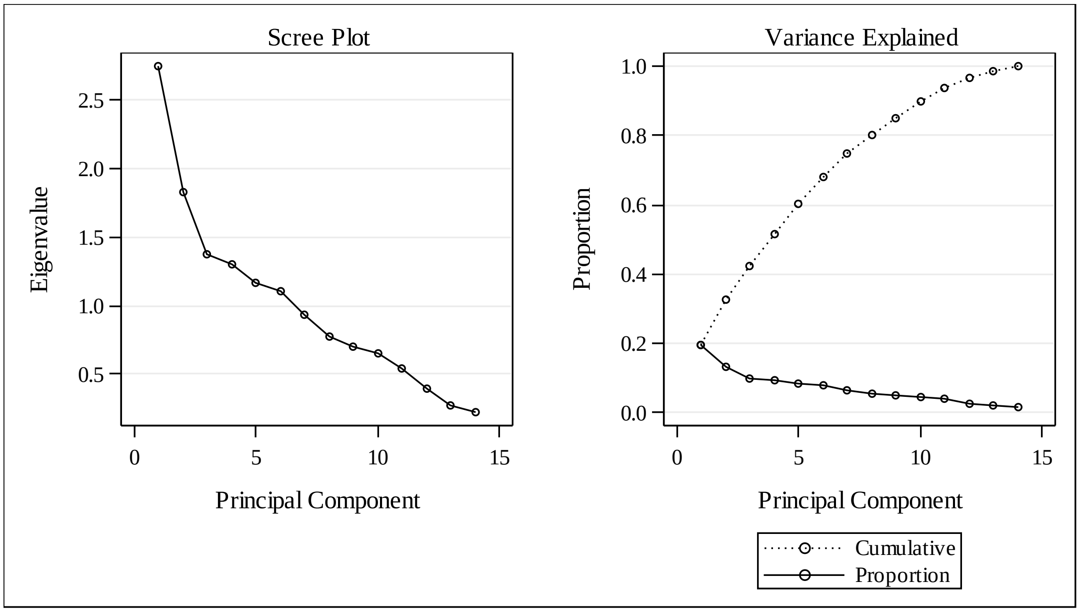
Simple Statistics								
	hgoal	aattempts	asog	hblock	acorner	across	aoff	hfoul
Mean	1.610169492	10.94067797	3.737288136	3.415254237	4.220338983	16.49152542	1.567796610	12.47457627
StD	1.176943007	4.04524068	2.006093138	2.157579647	2.432457142	7.36676037	1.630263015	3.51726356

Simple Statistics						
	hyel	hred	htackle	hclear	apasspct	Duels
Mean	1.677966102	0.1016949153	15.05084746	20.45762712	0.7486440678	1.083662495
StD	1.100748432	0.3035355907	4.85137650	9.15548882	0.0546146603	0.232605443

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.74609226	0.91916444	0.1961	0.1961
2	1.82692782	0.45527525	0.1305	0.3266
3	1.37165256	0.07425973	0.0980	0.4246
4	1.29739283	0.12839127	0.0927	0.5173
5	1.16900156	0.06677103	0.0835	0.6008
6	1.10223053	0.16563618	0.0787	0.6795

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
7	0.93659434	0.16704503	0.0669	0.7464
8	0.76954931	0.06503634	0.0550	0.8014
9	0.70451298	0.05522348	0.0503	0.8517
10	0.64928949	0.10386592	0.0464	0.8981
11	0.54542358	0.15095828	0.0390	0.9370
12	0.39446530	0.12700575	0.0282	0.9652
13	0.26745955	0.04805165	0.0191	0.9843
14	0.21940790		0.0157	1.0000

Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
hgoal	0.144945	-.008649	-.413709	0.255549	0.144514	-.495005	0.331157	-.025768	0.340545
aattempts	0.422950	-.144601	0.260469	0.066143	0.199147	0.005536	-.412734	-.075105	0.178046
asog	0.333447	-.362312	0.308207	0.109186	0.318175	-.174183	-.135741	-.009244	-.007279
hblock	-.069376	0.115720	0.489293	-.151010	0.229915	-.032654	0.521414	0.515710	0.286011
acorner	0.474400	0.146639	0.080439	-.195126	-.025524	0.055226	-.036047	0.111288	0.051089
across	0.493165	0.218834	-.153780	-.115660	-.135848	0.013585	0.156550	0.061159	-.103786
aoff	0.136899	-.028920	0.419986	0.299686	-.218190	0.189105	0.485347	-.422307	-.350195
hfoul	-.022663	-.376632	-.225881	-.053023	0.293734	0.456861	0.295454	-.317224	0.365846
hyel	0.089443	-.385561	-.261038	-.280610	0.094629	0.322154	0.083471	0.444016	-.409713
hred	0.048191	-.064573	0.024811	0.435287	-.481203	0.422985	-.144147	0.324326	0.451477
htackle	-.046326	0.309330	-.055546	0.262283	0.569223	0.358680	-.096566	-.055398	-.062323
hclear	0.353613	0.316640	-.187878	-.298518	-.100472	0.194376	0.122316	-.203462	0.162092
apasspct	0.230067	-.043506	-.248687	0.554915	0.065439	-.054105	0.140869	0.276518	-.305535
Duels	-.078366	0.521162	0.009342	0.132163	0.231224	0.154132	-.095653	0.096372	-.073522



The first 5 components explain over 60% of the variance. Additionally the components have correlations that resemble the original correlation analysis.

Regression Analysis of the Principal Components

I performed the regression analysis with the response Away Goals and given principal component attributes as explanatory variables:
Y = y0 + y1W1 + y2W2 + y3W3 + y4W4 + y5W5

Below is the SAS Output for the procedure:

Number of Observations Read	118
Number of Observations Used	118

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	13	37.23667	2.86436	3.26	0.0004
Error	104	91.34808	0.87835		
Corrected Total	117	128.58475			

Root MSE	0.93720	**R-Square**	0.2896
Dependent Mean	1.05932	**Adj R-Sq**	0.2008
Coeff Var	88.47183		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	1.05932	0.08628	12.28	<.0001
Prin1	1	0.05822	0.05229	1.11	0.2681
Prin3	1	0.32839	0.07398	4.44	<.0001
Prin4	1	0.19451	0.07607	2.56	0.0120
Prin5	1	0.10545	0.08014	1.32	0.1911

Due to the large p-values on Principal Components 1 4 and 5 we will reject those for the regression equation and below is the final model:
Y = 1.05932 + .32839(W3) + .19451(W4)

Conclusion

Interpretation of PCR Coefficients
Below is a table that explains the regression equation with the principal component variables.

Away Goals = 1.05932

+ .32839(-.413709*hgoal + .260469*aattempts + .308207*asog + .489293*hblock + .080439*acorner - .153780*across + .419986*aoff - .225881*hfoul - .261038*hyel + .024811*hred - .055546*htackle - .187878*hclear - .248687*apasspct + .009342*Duels)

+ .19451(.255549*hgoal + .066143*aattempts + .109186*hblock - .151010*acorner -.115660*across + .299686*aoff - .053023*hfoul - .280610*hyel + .435287(hred) + .262283*htackle - .298518*hclear + .554915apasspct + .132163*Duels)

Statistical Conclusion

Based on the Eigenvalues matrix, one can see that the below listed variables mostly contribute based on the direction of their maximum variance to the principal component 3 and 4.

"Home goals", "Away Shots on Goal", "Home blocks", and "Away Offsides" (Prin3)

"Home red cards", and "Away pass percentage" (Prin4)

Given the parameter estimates for the explanatory principal components and their p-values, we have a statistically significant correlation with the prin3 and explains about 29% of the awaygoals. (Rsquare = .2896). Other principal components "prin1, prin2, prin4, and prin5 are estimated to not be statistically significant with a p-value > 0. As we know, the statistical association from these observational data cannot be used to establish a causal interpretation. However, based on the parameter estimates, we can see that there is a weak correlation between the given attributes.

Appendix

SAS Code

```
/*import wizard with original dataset*/

proc print data = stat;run;

ods rtf;
proc sgscatter data = stat;
matrix agoal asog hblock acorner across apasspct duels ;
run;

proc corr data = stat nosimple;
var hgoal agoal aattempts asog hblock acorner across aoff hfoul hyel
hred htackle hclear apasspct duels;
run;

proc univariate data = stat;
var agoal;
histogram agoal;
qqplot agoal;
run;
ods rtf;

title 'Principle Component Analysis for STATS';
proc princomp data = stat out = statpc;
var hgoal aattempts asog hblock acorner across aoff hfoul hyel hred
htackle hclear apasspct duels; run;

/*proc print data = statpc; run;*/

title 'Regression with Principle Components';
proc reg data = statpc;
model agoal = prin1 prin3 prin4 prin5; run;

ods rtf close;
/*exploratory with log*/
data logstat;
set stat;
logagoal = log(agoal+1);
run;
proc print data=logstat; run;

proc univariate data = logstat;
var logagoal;
histogram logagoal;
qqplot logagoal;
run;
```