Owen Wasilewski

Dr. John Clark

SPMG 4820 A

1 May 2024

**Performance variability across Professional Football (soccer)**

**Background and Hypothesis**

For the last ten or so years since I started watching soccer it feels like in the big European leagues, the same 2 or 3 teams are the only teams that have a chance to win the league every year.  Comparing this to MLS, we have seen 8 different champions in the last 10 years.  So, in this project, I want to test if MLS (Major League Soccer) was more random than the big 5 European leagues.

MLS has multiple reasons why it could be more random than the other leagues.  First, MLS has a salary cap, like most other American sports leagues.  This creates a level playing field for all 29 teams, so the same team doesn’t always win.  The big difference is you can sign 3 players who don’t count against the cap.  This allows teams to bring in players who would be unable to be signed in the system as it is.  So, players like David Beckham, Thierry Henry, and Leonel Messi can join the league and still get paid the money they deserve.  Next is the playoffs.  Unlike the European leagues where you need to finish first to win the league, in MLS you can technically be the 18th worst team and still win the MLS Cup.  This can change the way teams play early or later in the season based on if they play in the Concacaf Champions League in February and March or have already clinched a playoff spot in September and October.  Third is an unbalanced schedule.  In Europe, you play every team in your league home and away during the season.  In MLS you play every team in your conference, east or west, at least once home and away, and then you play the other conference usually 6 times.  This could mean you can play the worst 6 teams in the west and someone in your conference can play the 6 best teams in the west making an unbalanced schedule.  Finally, some simple statistics based on the teams on average in each league.  MLS has the second youngest average age, 25.7, out of the 6 leagues being tested.  Only being beaten out by Ligue 1 by .1 year.  The average transfer value per player is $1.54 million, which is the least and $6.5 million less than the next closest being Ligue 1 again.  I also wanted to look at team payrolls between the Premier League and MLS.  The top payroll in the Premier was $217 million and the least was $15 million last season.  But in MLS we see that the top was $40 million, and the least was $8 million.  That is a smaller gap between the two, showing a more competitive balance.

I will be testing 6 leagues, MLS, Premier League, La Liga, Bundesliga, Serie A, and Ligue 1, over the last 6 completed seasons, 2018-2023.  These 6 seasons were picked because the 2017-2018 season, completed in 2018, was the first season that expected goals, XG, was calculated.  I will be calculating 3 different point systems to see how accurate they are in each league.  Those 3 systems are Projected, Expected, and Poisson Points.  They will consider before a game, after a game, and after a season, respectively.  All 6 league will be averaged to 38 games, since some play 38, some play 34, and when we remove some teams, each metric will be comparable to each other.  The values I will get are the sum of the absolute difference of actual points and predicted points earned, divided by the number of teams in the league.

**Methodology**

The first predicted point system was Projected Points.  For this I used betting odds from BET365.  I turned betting odds into the probability of a home win, draw, and away win.  I then multiplied each outcome by the points that team would get for that outcome.  This being 3 for a win, 1 for a draw, and 0 for a loss.  I will use the first game of the 2018 MLS season as an example for each predictor, this being Toronto, home, vs Columbus, away.  The betting odds for a home win was 1.54, a draw 4.12, and an away win 5.85.  This than being turned into probabilities lead to 61.1% for a home win, 22.8% for a draw, and 16.1% for a home win.  This means that Toronto would have a projected points of 2.061 points and Columbus would have 0.711 projected points.  The drawback of this system was that the betting odds were collected the day before the game, so last minute changes to the expected starting lineup would not be accounted for.

The second predicted point system was Expected Points.  For this I used Expected Goals, XG, from Opta.  XG is a statistic that measures how good a shot was.  So, a penalty is .8 XG and a shot from far out is less than .01.  Using this is the best metric as of now to measure what team had more and better-quality shots.  To find the Expected Points, I put every team's XG into a Poisson distribution to get the probability that they score x number of goals in a game.  For each game, each final score was found by multiplying the probability of each team scoring that number of goals.  Using our example, Toronto had 1.4 XG and Columbus had 1.5.  Using those as the means in the Poisson distribution, it led to the probability of a Toronto win being 35.4%, a draw 24.7%, and a Columbus win 39.9%.  Then multiplying that by the points you would get for that result, Toronto had 1.309 Expected Points and Columbus had 1.443 Expected Points.  The drawback of this is that XG does not account for who takes the shot, so if the best player or the worst player is the one taking the shot, XG would be the same.

The third and final predicted point system was Poisson Points.  For this I used the goals scored and conceded by each team when they played at home and away during each season that was tested.  I put each game result into a Poisson regression and that gave me a formula to calculate another form of expected goals after the season is over.  The formula had an intercept and 3 betas.  First beta 1, was if you played at home, include the value given, or if you were away, just 0.  Beta 2 was your scoring output and Beta 3 was your opponent’s conceding variable.  These values were then treated like XG in the previous example and calculated probabilities once put into a Poisson distribution.  Using the same example again, Toronto had a value of 1.76 and Columbus was 1.19.  We see the effect of playing at home since Columbus was the slightly better team at the end of the season.  But this puts the probabilities back in favor of Toronto giving them a probability of winning at 50.6%, drawing at 23.5% and Columbus winning at 25.9%.  That gave us a Poisson Points value of 1.752 for Toronto and 1.013 for Columbus.  The biggest drawback of this was it was affected by blowouts.  If two teams of equal level in the standings played each other, and one of the teams won by a large amount it would throw off both teams where they should be, but that rarely happened.  Columbus ended up winning this game easily 2 to 0 so we see the predictors were off to a great start.

**Results**

Below are the results in a table based on each Point system.  The columns are each league tested and the rows are different ways I tested each.  The first row is the full schedule played normally.  The second and third rows removed the best 4 teams, teams that would qualify for the champions league in Europe, and the bottom 3, teams that would be relegated in each European league.  This was used to remove some of the outliers that either beat everyone or lost to everyone and see how the league turned out without them.  The final row was used to remove the covid seasons, 2020 and 2021.  I wanted to see how big of an impact covid had on each of the leagues, especially MLS.  First is Projected Points, second is Expected Points, and third is Poisson Points.  Blue highlights signify the most random in that row and red signify least random.

A screenshot of a table

Description automatically generated

When we look at the results for Projected Points, we see that MLS is by far worse than the other leagues.  During the full schedule and minus relegated teams, it is in a league of its own.  But when we get rid of the top 4 teams and covid seasons, we see while MLS is still most random, it is a lot closer.  We also Serie A and La Liga as the least random variables in this category.  While the Premier League is all over the place.

A screenshot of a table

Description automatically generated

Looking at the results for Expected Points we see that MLS is in the middle for full schedule and minus relegated teams, and it is close to being the least random in minus top 4, while they are the least random without the covid years.  Overall, the Bundesliga seems to be the least random in this predictor, while Serie A is the most random right after being the least random in the betting odds.  The Premier League continues to be all over the place, even having a most and least random spot in this predictor.  Finally, we see that Ligue 1 is the most random without covid season by 0.0000012 over Serie A, so hopefully this was a rounding error.

A screenshot of a table

Description automatically generated

Finally, looking at Poisson Points, we see that every number is by far lower than in the previous two making this the most accurate at predicting points.  This makes sense since it considers the whole season at once and not game by game.  MLS doesn’t have the most or least random in any of the rows for this predictor but besides bottom 3 it is the second most random in the other 3. We see Ligue 1 be the least random and La Liga be the most random.  The Premier League seems to balance out here and does not seem to be all over the place in its randomness.

Overall, taking all 3 predictors into consideration, MLS is the most random league out of the 6.  When we had the most random season out of the 36 tested, we saw that the 2020 MLS season was the most random, when we averaged the 3 full schedule predictors together.  This would make sense the season was much shorter, making the need to balance it 38 games give a bigger penalty.  Also, an unbalanced schedule, to make sure teams didn’t leave their pockets of the country.  This would leave me to put an asterisk next to MLS being the most random, as when we get rid of the covid seasons, it is a real toss up to truly what the most random league is.  Out of the 5 European leagues I think it’s clear that Bundesliga is the least random league.  This would make sense to me because before the 2024 season, not included, the league has been won by the same team, Bayern Munich, since the 2012-2013 season.  This year Bayer 04 Leverkusen has had to not lose a single game in any competition, at the time of writing, this season to win over Bayern.

**Future Considerations and Improvements**

Considering the result I got, I think this would be a great project to revisit in 4-6 years for multiple reasons.  First, Serie A has finally been won by different times for the last couple seasons, leading to more random results and leagues like the Bundesliga and Premier League might be heading that way with some retirements and complacency.  Second, once we get more seasons with XG as a stat, I can get 10-12 seasons included and maybe completely remove the covid seasons.  Last, we might see a new widely spread way to rate performances for a whole team per game, that could provide a better predictor than XG. Another thing I could do with this project is add other smaller leagues closer in level to MLS.  These could include the Mexican, Dutch, Portuguese, Turkish, and the 2nd tier in England.  I mention the second tier as that is usually the league that players who played in both the English Championship and MLS say feel the most similar competition wise.

Finally, something that could maybe improve the model would try to put XG into a Poisson Regression model like I did for normal goals in the Poisson points.  This allows the XG to be spread throughout the season instead of just one game at a time.  At worst it might lower the difference very slightly but still create a better overall predictor.

**Work Cited**

# Illien, Mael. “Soccer Analysis and Modeling.” *Rpubs,* December 9, 2020

# <https://rpubs.com/maelillien/poisson-soccer>

# Lee, Andy. “Did Liverpool Deserve to Win the Premier League?” *Toward Data Science*, July 8,

# 2020, <https://towardsdatascience.com/did-liverpool-deserve-to-win-the-premier-league-eda7a4e1b9ca#h1>

N.A., “Historical Football Results and Betting Odds Data.” *football-data,* May 1, 2024

<https://www.football-data.co.uk/data.php>

N.A. “Major League Soccer’. *Transfer Market,* May 1, 2024,

<https://www.transfermarkt.us/major-league-soccer/>

Sports Reference LLC. *FBRef.com - Football Statistics and History,* May 1, 2024,

<https://fbref.com/en/>