# An In-Depth Look at NCAA Pac-12 Women's Soccer

By Joey Maurer • 14 Jan 2018 • 8 min read

Courtesy UCLA

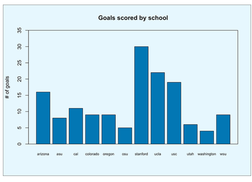
Women's collegiate soccer is thriving on the west coast. Less than two months ago, the championship game of the College Cup saw two Pac-12 powerhouses go head-to-head in Orlando, Florida. Stanford emerged with a 3-2 victory over UCLA to secure their second title since 2011 and the conference's fourth in that seven-year span. Both teams are set up extremely well heading into next season, and the road to a national championship appears to run through California.

The following is a comprehensive introduction to the statistics of NCAA women's soccer, focusing specifically on Pac-12 teams during the 2017 season. I scraped and organized the data using Python and did the analysis/visuals in R. A couple issues in the data: Stanford's website does not list minutes played, so all analysis using the minutes variables excludes Stanford players. Also, Oregon State changed the format of their website before the last conference game, so I was unable to get data from their 1-0 win over Oregon. All warnings aside, let's jump into it.

Soccer is a particularly hard sport to quantify; there are very few goals per game and very few games in a season. Without passing data or advanced player tracking technology, we do not have much to work with aside from basic counting statistics. Games played, games started, and minutes played can tell us about the usage of a player, while goals, assists, points (goals*2 + assists), shots, and shots on goal represent offensive production. But there are no specific measurements for defense. Using minutes played per game and percentage of games started (GS/GP) as a proxy can tell us how much a coach trusts a player, but it does not directly measure performance.

In short, counting stats cannot be solely relied upon to judge a player's value. The eye test is more applicable in soccer compared to most other major sports. However, numbers can still be applied in many situations, a few outlined below. Let's overview the dynamics of the Pac-12 conference teams during the 2017 season.

Parity in college soccer is relatively low. Stanford was the top team in the country and on a tier of its own in the Pac-12, going a perfect 11-0-0. UCLA and USC followed closely behind, and were legitimate contenders for a national championship. The middle ground of the conference made up schools that at least made the tournament: Arizona, Cal, Colorado, and Washington State. The rest fill out the bottom. A large disparity exists in the talent level between the best and the worst, resulting in few upsets and relatively predictable games. One must be very careful when making comparisons across teams. Here is why:



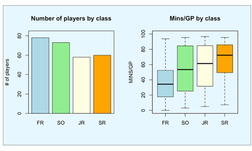

Notice the large discrepancies. If you're on a good team, you're going to have the ball a lot. Teams like Stanford and UCLA dominate possession and get a lot of high quality opportunities (that lead to goals) while limiting their opponent's. Good players on bad teams simply do not have the quality of teammates around them to facilitate the same volume of scoring chances.

Additionally, positions must be taken into account when comparing players. Forwards and midfielders, for example, often have different roles. Jessie Fleming, a member of Canada's national team and 2016 Olympian, is often described as the "straw that stirs the drink" for UCLA. She finished fourth on the team with 5 points in 10 conference games (for context, UCLA's top scorer was Hailie Mace with 15 points in 11 games). Point is that Fleming was not relied upon for offense. She played as a holding midfielder who would control possession and set up the attack from deep. Stanford's Andi Sullivan holds a similar pedigree as one of the best players in the country, and she did not score at all in conference play. This is why we will not focus too much on offensive numbers, but rather explore other variables that can give more insight.

Enter the usage statistics mentioned above; games played, games started, and minutes played. Generally, good players will start most or all of their games played, and will play a lot of minutes. Of course, there will still be variability between teams and some schools may use players in situations where they otherwise wouldn't if they had a better option. Still, usage statistics are less volatile and more encompassing than pure offensive numbers.

# Usage by Class

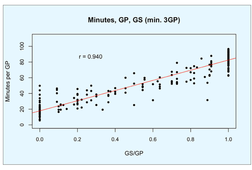
How does the distribution among freshmen, sophomores, juniors, and seniors look? Pretty much how you would expect. The more experienced seniors play the most minutes on average. There are some superstar freshmen and some substitute-level seniors, which is why they all have similar ranges. But the overall trend is the further along in college a player is, the more minutes they are likely to play.




Three quarters of freshmen come in as substitutes or only play about half of the game. Also note that freshmen average the least GP per player out of the four classes, meaning they are most likely to serve as 'healthy scratches' and not play at all. There is typically an adjustment period between club soccer and college, and it's hard to displace upperclassmen that know the system and have the experience.

However, top recruits may be expected to step into a starting role immediately. 10 freshmen started at least 10 games last season. 6 of them were from Stanford, UCLA, or USC. It's no secret that the top schools usually get the top incoming players. Recruiting is one of the major responsibilities of a coach and is crucial for sustained success. Stanford is a prime example of this. Freshman Catarina Macario was named the ESPNW Player of the Year and took home a myriad of other honors and awards after a historic first season. By sophomore year, many of the top talents have established themselves as starters and the interquartile range condenses leading up to seniors.

Games started divided by games played as another useful usage tool as it correlates extremely well with minutes per game played:




Perhaps a model can be made in the future to 'fill in' the missing values of Stanford players' minutes. Just by knowing games played and games started, you can make a pretty accurate guess of minutes per game. Taking into account position and year can paint a clear picture about the usage of a player. An example:

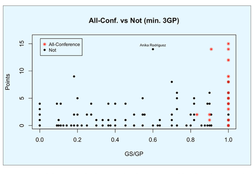
Player X (UCLA): 10/11 GS/GP, senior, defense

Her GS/GP ratio is ~90%. Looking at the graph, we can guess she probably plays between 60-80 minutes per game. This implies that she is a regular starter and as she plays defense for UCLA, she occupies a pretty big role on a top team in the conference.

Player X is Mackenzie Cerda. Sure enough, she plays 71 minutes per game and is an experienced starter on the back line for UCLA.

# All-Conference Teams

A useful set of information is released at the end of conference play, the All-Pac-12 Conference Teams. Officials for the conference who watch a lot of soccer come together to compile a first team, second team, and third team consisting of 36 players as well as an All-Freshman team of 12. First years are eligible for the All-Conference teams. While this is a subjective ranking, it provides a solid baseline for cohort analysis. Here is a scatter plot distinguishing All-Conference players based on usage and production.




It's not the prettiest graph because I had to use the discrete variable of GS/GP to include Stanford, but the takeaway is clear. A crucial requisite for being included in one of the All-Conference teams is that a player must be a regular starter for their school. All of the red stars are to the right of that 80% GS/GP threshold. Being a starter appears to take precedence over points scored, which brings us to the case of Anika Rodriguez.

Rodriguez is a forward for UCLA who was tied for third in the Pac-12 with 14 points during conference play. Surely this warrants an inclusion into one of those 36 spots? Apparently not. Rodriguez started 6 of 10 conference games and played less than 50 minutes five times in that span. Perhaps the committee interpreted this as insufficient importance to her team. Rodriguez was an absolute force during the NCAA tournament and an integral part of UCLA's run to the championship game, but her inconsistent usage during conference play was probably why she did not make any of the teams.

Which begs the question, had she played for a lesser team and still put up good numbers, would she have been included in one of the three teams? I'd lean towards 'yes'. She should clearly be the main offensive threat on a team like Oregon State or Washington and her minutes and GS/GP would shoot up. That is the issue with using a subjective ranking system like this. If you were going by simply 'who are the best 36 soccer players in the conference', the majority would be from Stanford, UCLA, and USC. But usage and playing opportunity clearly hold weight in these All-Conference teams, which can hinder the perceived value of some of the more talented players buried behind superstars on top schools.

There is a lot more to explore that I will perhaps look at in future projects, such as a deeper analysis into the makeup of teams and how that translates to success over the course of a season. Or the turnover rate (quality of seniors lost/freshmen gained/transfers) and how that relates to placement the following year. And of course, there is always more data to be acquired. And perhaps scraping box scores to examine events in individual games on a per-minute basis, or expanding to other major conferences to get a sense of college soccer as a whole. Suffice to say, this was just scratching the surface of data analysis that can be applied to NCAA women's soccer.