A brief history of statistics in soccer: Why actual goals remain king in predicting who will win

December 30 2022, by Laurence Shaw

Credit: Csongor Kemény from Pexels

In 2017, BBC's Match of the Day introduced a new statistic in their post-match summaries of Premier League matches. Expected goals, or xG, is designed to tell us how many goals a team should have scored based on the quality of the chances they created in a game. It is loved by amateur and professional statisticians alike who want to use data to analyze performance.

The BBC regularly uses xG in its Premier League coverage, but this metric was absent from both BBC and ITV coverage at the recent men's World Cup. A brief look into what xG is and the history of using data to predict soccer matches may give us some insight into why they decided not to use it.

The concept of expected goals originally came from ice hockey but is easily appliable to soccer. xG is calculated by looking at every shot that a team took in a match and assigning it a probability of being scored.

This probability is calculated by looking at shots from similar situations in historical matches and calculating what percentage of them resulted in a goal. By adding the probabilities together for all shots that a team takes, we get their expected goals for the entire game.

Consider the Premier League match between Tottenham and Liverpool in November 2022, which Liverpool won 2-1. Liverpool only achieved an xG of 1.18 from 13 shots in the match, while Tottenham managed an xG of 1.21 from their 14 shots.

In the post-match interviews, Tottenham manager Antonio Conte claimed that Tottenham were unlucky to lose given their performance. An xG score line of 1.21 vs. 1.18 suggests a very even game and would seem to back up Conte's point.

However, Liverpool manager Jürgen Klopp suggested that the quality of Mohamed Salah, who scored two goals from three shots with a combined xG of 0.67, was the difference in this match. This exposes one of the major weaknesses of xG. It takes no account of who the striker or goalkeeper is. But is this weakness enough to make xG unreliable as a resource for predicting future games?

Soccer prediction before xG

The obvious piece of data to use when analyzing soccer is goals. Indeed, this was the only information used in the 1997 model of Mark Dixon and Stuart Coles, which predicts future soccer matches by assigning each team attacking and defensive rating.

The Dixon-Coles ratings are calculated using the number of goals scored and conceded in previous matches, taking account of the quality of the opposition. The ratings of two different teams, along with a home advantage boost, can them be combined to predict the score of an upcoming match between them.

Given the number of statistics available in soccer, a model that only uses goals to predict future games may seem remarkably simple, but its effectiveness lies in understanding what makes for good statistical analysis: high quality data, and lots of it.

Goals are the highest quality data available in soccer prediction, since they are the only thing that actually affects results. This explains why other traditional metrics such as number of shots or possession percentage are not used in the Dixon-Coles model.

A shot could be a penalty, which players expect to score, or a speculative effort from distance—yet both count equally as shots on goal. Similarly, a team could have lots of possession but not in an area of the pitch that gives them chances to score goals.

As far back as 1968, a statistical study was unable to find any link between shots, possession or passing moves and the outcomes of soccer matches. This supports the idea that goals are the only factor worth considering.

Why might xG be useful?

The weakness of Dixon-Coles comes in the quantity of data. There were 1,071 goals scored in the 2021/22 Premier League season, which may seem like a lot. However, this is only 2.82 goals per game. To counteract this lack of information per game, Dixon and Coles used three years' worth of data to make their predictions, despite most teams going through wholesale changes in playing and management staff over this period.

Increasing the quantity of data over a shorter timescale is where xG data has an advantage over goals alone. Essentially, it is an attempt to find balance between the quality of goal data and the quantity of shot-based data. This is a classic conundrum in statistics known as the bias-variance trade-off.

Take the Liverpool vs. Tottenham game mentioned earlier. The three goals scored are the only pieces of information that the Dixon-Coles model can extract from this match, whereas an xG-based model would get information from all 27 shots taken—with the added quality of having some indication of how likely those shots were to result in a goal. However, not taking account of who is involved in a shot does place a limit on the quality of this xG data.

Despite being 25 years old, the Dixon-Coles model is still the gold standard of soccer prediction, as found in this 2022 study. While xG provides good information about the balance of play in a single match, no xG model has been shown to be superior to Dixon-Coles in predicting the future.

Until that happens, doubts about its weaknesses will remain and actual goals must retain their place as the only truly reliable indicator of how good a team is.

Provided by The Conversation

Citation: A brief history of statistics in soccer: Why actual goals remain king in predicting who will win (2022, December 30) retrieved 23 June 2024 from https://phys.org/news/2022-12-history-statistics-soccer-actual-goals.html