Calculating Team Ratings

The model contains two fixed parameters: an intercept term equal to the average number of goals per team-game and an adjustment for the game's location to account for home-field advantage. Additionally, it contains random effects terms for each of the two teams in the game.

Two separate models were trained per game, one from each team's "perspective," offensive and defensive effects for each team in the game are computed. For a match between Team A and Team B, the model for goals scored by Team A yields random effects terms for Team A's offensive ability and Team B's defensive ability; the model for goals scored by Team B yields random effects for Team B's offensive ability and Team A's defensive ability.

The dummy variable Location is used to encode information about the location of a match and treated as a numeric variable in the model: 1 for home matches, 0 for neutral-site matches, and -1 for away matches. This method of encoding the dummy variable is preferred to traditional one-hot encoding for three reasons: it results in a more practical interpretation of the associated beta value, it guarantees that neutral-site games will not have any home or away effect, and its flexibility allows for the potential encoding of "semi-home" or "semi-away" matches (e.g. a match between England and the United States played in Wales) with values between 0 and 1 in future models.

The two models were trained in R using the lme4 package using all international matches between January 1, 1993 and the beginning of the 2022 Qatar World Cup. The start date of 1993 was chosen because it marks the most recent time FIFA updated its Rules of the Game; the rule changes caused a significant uptick in scoring in international matches, so including pre-1993 data would require undesirable additional transformations of the response variable to control for its non-stationary mean over time.

Exploratory data analysis of goals scored and allowed by teams at the game level suggested that the data broadly follow a Poisson distribution, a conclusion supported by recent soccer match prediction literature. Since the Poisson-distributed count data are clearly not suitable for traditional linear regression due to heteroscedastic errors and the fact that the response can only take integer values, a Poisson regression model was fit to the data with a log link function used to ensure that the transformed response took only positive values. The Poisson model allowed variance to increase along with the mean since for Poisson-distributed random variables the two values are equal. It also ensured score predictions greater than zero for all games, a necessary property not satisfied by other hypothesized generalized linear models such as gamma regression.

In order to account for the time-dependent nature of the observations, an exponential decay filter was implemented to increase the importance of more recent games to the model's estimates. This filter weighted games by a factor of one-half to the power of d/H where d is the number of days elapsed between the current date and the match date and H is the half-life for each game's importance. Backtesting and evaluating model performance suggested the best value for H to be 2,190 days or six years. Practically, this means matches played six years ago from the current date hold 50 percent as much weight in the model as matches played on the current date.

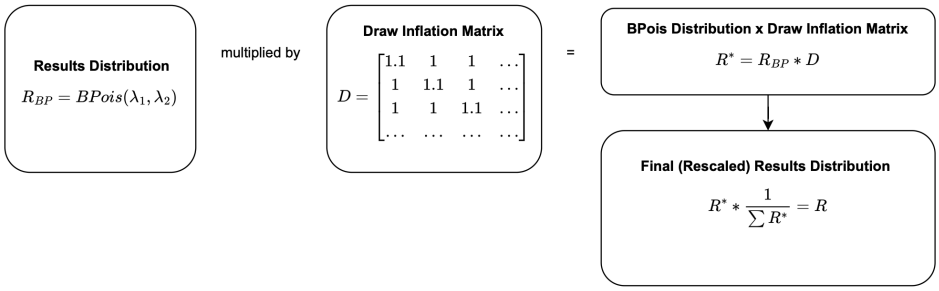
Finally, fitting the mixed-effects Poisson models in R yielded offensive and defensive coefficients for each team which could be used to rank teams using net effects as well as generate individual match predictions.

Calculating Projections

Previous research into soccer match prediction suggests that final scores typically follow a bivariate Poisson distribution inflated along the diagonal by a draw-inflation matrix. This diagonal inflation, which is estimated to produce an increase of approximately 10% in the number of expected draws compared to a typical bivariate Poisson distribution, is hypothesized to occur because of the observed conservative and defense-oriented strategies of teams in close games late — a phenomenon linked to loss aversion. Without this adjustment, the match-prediction model consistently overestimates the number of wins and losses in relation to draws.

Using the predictions for team and opponent score using the team-strength model outlined in the previous section yields two lambda parameters representing the expected number of goals scored by each team in the match. Multiplying the bivariate Poisson distribution produced by those two parameters by the draw-inflation matrix, equal to 1 in all locations not along the diagonal and 1.1 along the diagonal, yields a results matrix with the probabilities of each potential final score. This results matrix is rescaled by a factor of 1 divided by the sum of the results matrix to ensure that it represents a valid probability mass function which sums to 1.

Generating the Final Results Distribution from Predicted Scores



The results matrix R can be translated into soccer terms quite easily. Assuming the x-dimension represents goals scored by Team A and the y-dimension represents goals scored by Team B:

- Team A's win probability equals the sum of all values below the diagonal.
- Team B's win probability equals the sum of all values above the diagonal.
- The probability of a draw equals the trace of the matrix, or the sum along the diagonal.