Evaluating Parametric Methods for Modeling European Soccer Team Goals

Zhiwei Xiao, University of Michigan
Thea Sukianto, Boise State University

Carnegie Mellon Sports Analytics Conference 2020

Introduction of the Data

- European Pro Soccer Database for 2008-2016 seasons
- Source: Kaggle (a machine learning and data science community)
- Format: SQLite (accessible in R using dbConnect function)
  - Contains information of 11 leagues
  - More than 25000 matches / 1000 players
  - Tables: Country, **League**, **Match**, Player, player attributes, **Team**, team attributes
  - Our research focuses on the latest season 2015-2016

How can we best model soccer team ratings in European leagues?

Getting Started

- Start simple - two parametric models (GLMs)
- Ratings are based on goals
  - Offensive and defensive ratings
- **Response:** home, away team goals (counts)
- **Predictors:** home, away team names (categorical)
- Let’s see how the data is distributed...

Candidate Distribution #1: Poisson

- Suggested by existing work (e.g. Karlis and Ntzoufras 2003)
- Most interpretable
  - If **X** ~ Poisson(λ), then…
    - **X:** r.v., # of event occurrences in a certain interval **(goals per match)**
    - **λ**: Rate parameter **(mean and variance of the goals per match)**

Correlation?

Overdispersion?

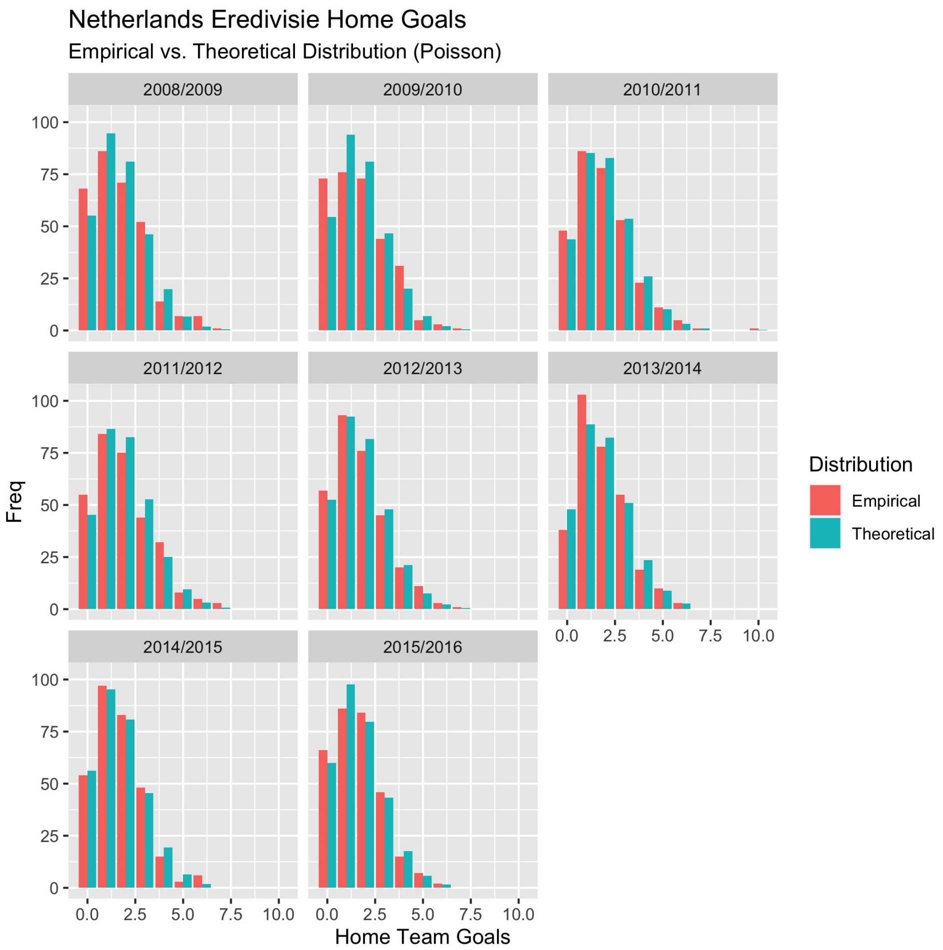
Zero-inflation?

Overdispersion?

- **Definition:**
    - presence of greater variability in a data set than would be expected based on a given statistical model
    - Often encountered when fitting simple parametric models like poisson distribution
        - Poisson has one free parameter and does not allow for the variance to be adjusted independently of the mean.
- **How to check: Computed variance to mean ratio for home/away team goals**
    - Done over different seasons and different leagues
    - Ratios were all very close to 1, which suggested we don’t need to worry about overdispersion.

Zero-inflation?

- **Zero-inflation:** Observed zero frequency > Expected
- Checking for zero-inflation
  - Plot observed vs. expected Poisson frequencies
  - Conduct a chi-square goodness-of-fit (GOF) significance test



Layout of Most Recent Work

Candidate Distributions

Model Evaluation

Poisson → Likelihood Estimation

Negative Binomial → Likelihood Estimation

Zero-inflated Poisson (ZIP) → Likelihood Estimation

Time-window Validation

Compare model coefficients

Compute Multiclass Brier Score

Plot Probability Calibration

Zero-inflated Negative Binomial (ZINB) → Likelihood Estimation

Implementation difficulties
(Attempted resampling, but MLE algorithm would not converge)

Log-Likelihood Comparison

Home Team Goals

Based on fitdistr()'s log-likelihood output:

Belgium: Poisson

England: Zero Inflated Poisson

France: Negative Binomial

Germany: Negative Binomial

Italy: Zero Inflated Poisson

Netherland: Zero Inflated Poisson

Poland: Negative Binomial

Portugal: Negative Binomial

Scotland: Negative Binomial

Spain: Negative Binomial

Switzerland: Zero Inflated Poisson

Away Team Goals

Similar log-likelihoods for each model (within +/- 1 unit)

Formula: Away goals <- home team + away team

1. **Poisson, Negative Binomial**

Belgium, Netherlands, Poland

2. **Poisson, Negative Binomial, Zero-Inflated Poisson**

England, France, Germany, Italy, Scotland, Spain, Switzerland

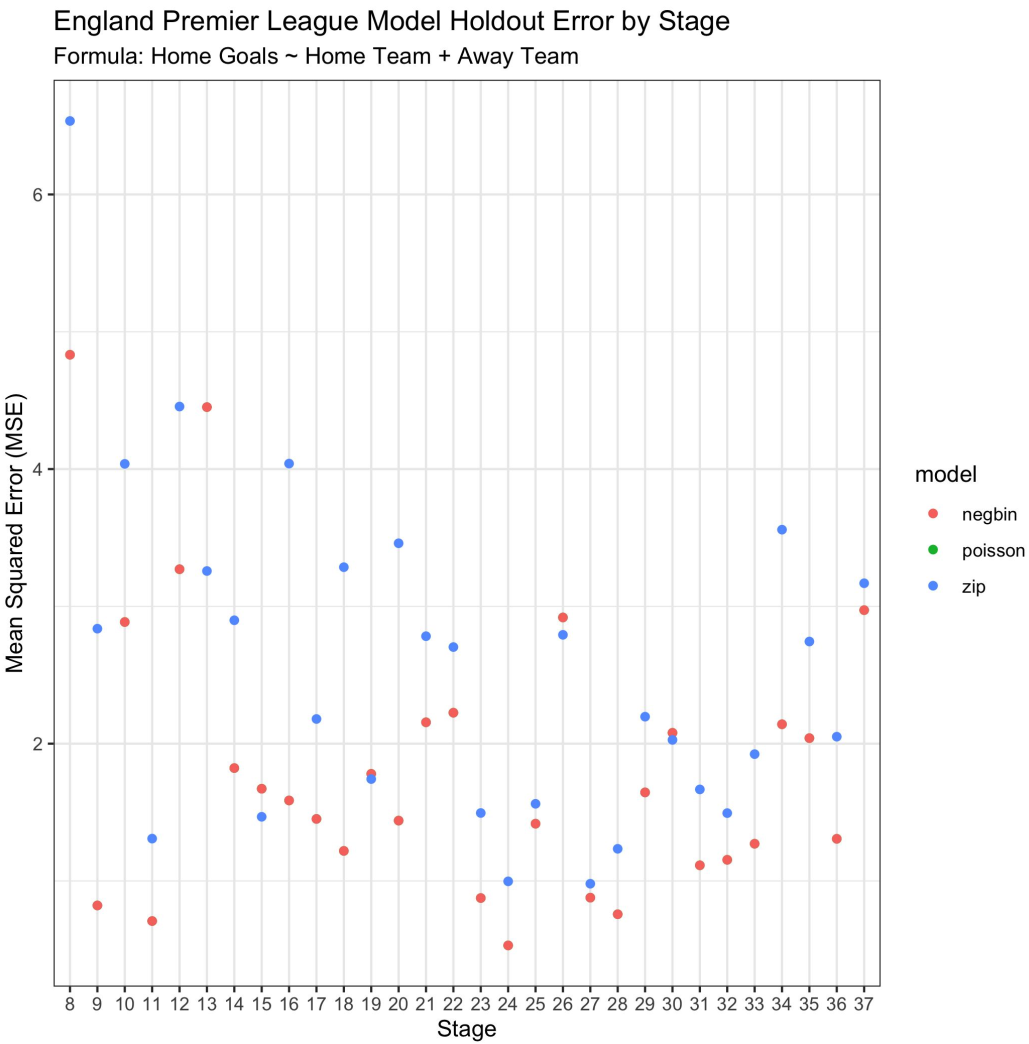
3. **Poisson, Negative Binomial, both zero-inflated alternatives**

Portugal

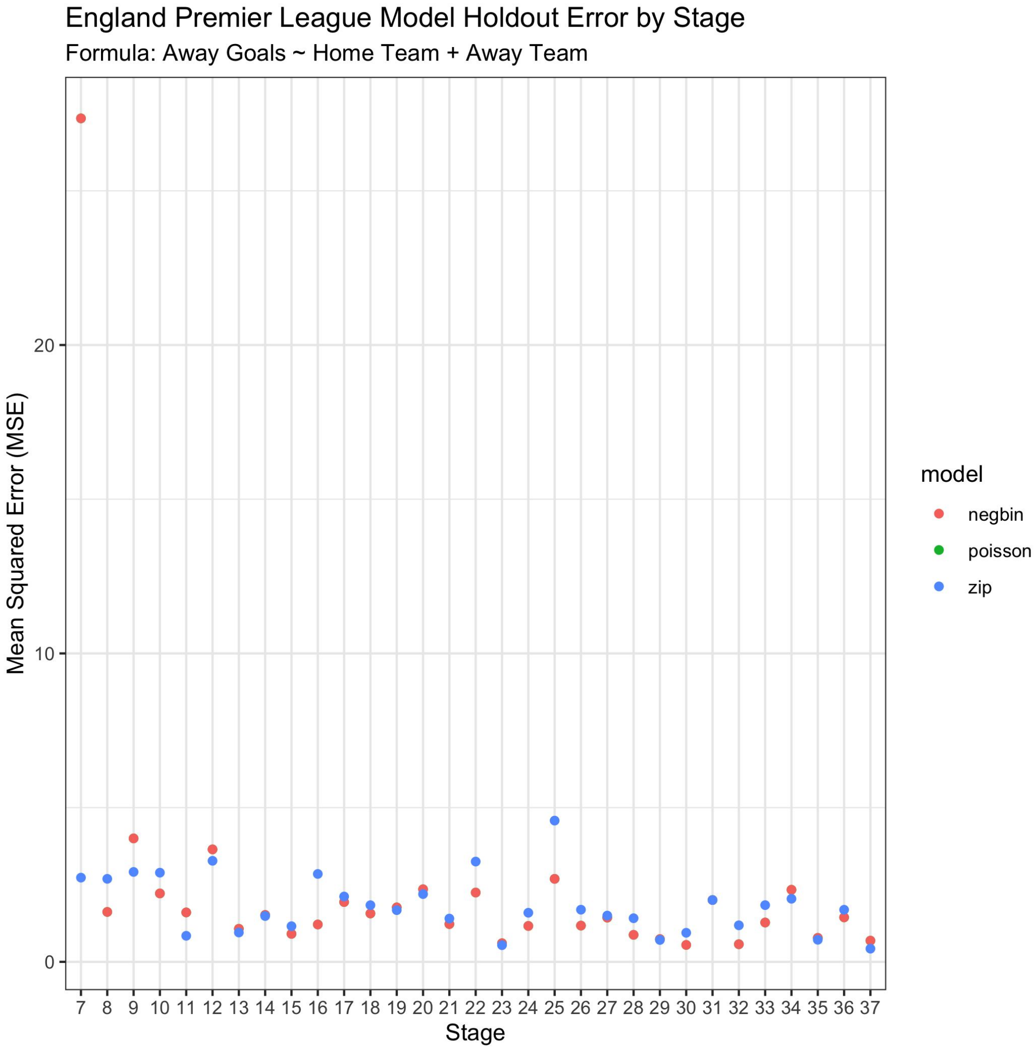
Model Validation

Model Validation Procedure

- Time window based train-test split
- Difficulty: finding the earliest group stage when all teams have played
  - Choose **week 6** for ease of implementation
- For group stage number **t,** train model on group stages **1, … , t** and test on group stage **t + 1**
- Calculate the mean squared error (MSE)



Not shown: Poisson/ZIP stage 7 test error = 8e22 (!)

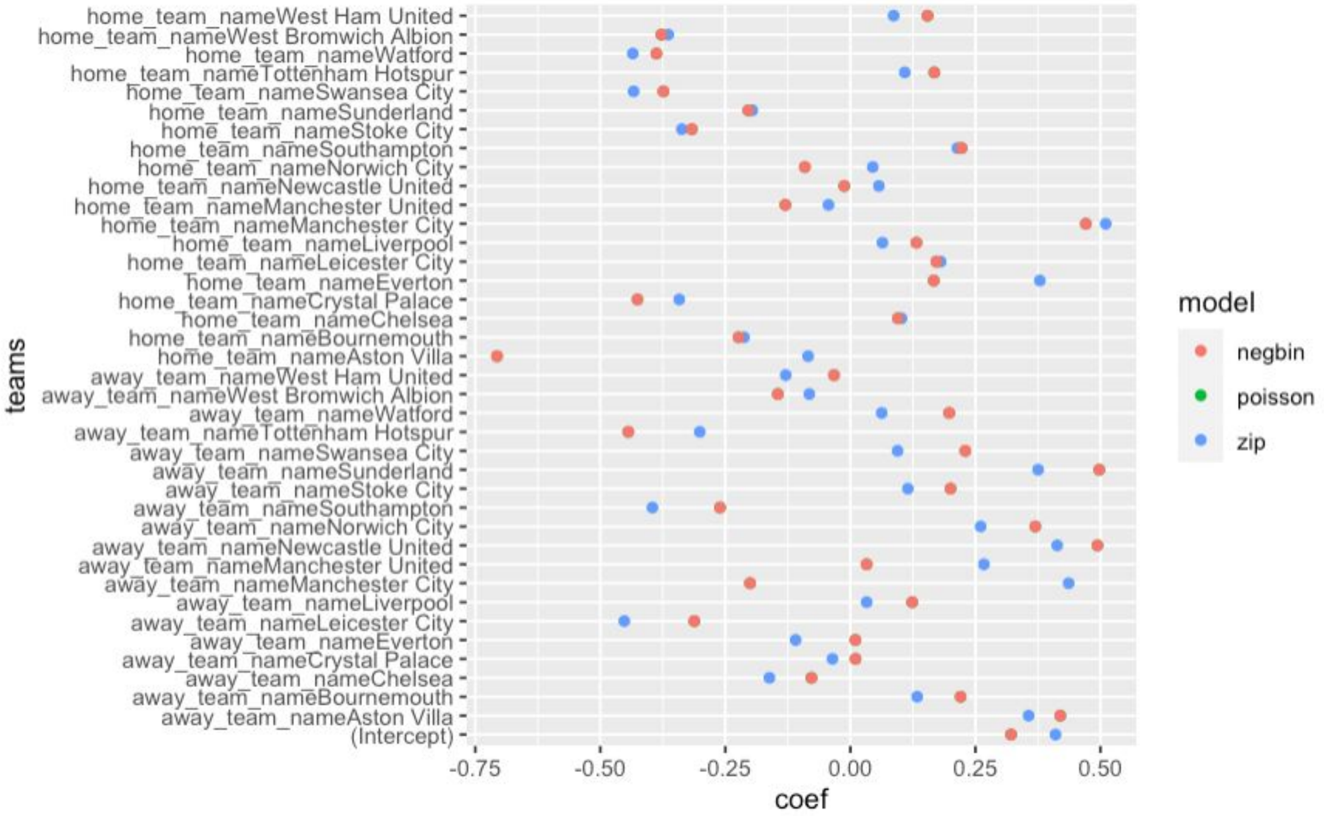


Stage 7 negative binomial test error is higher

Coefficients Comparison

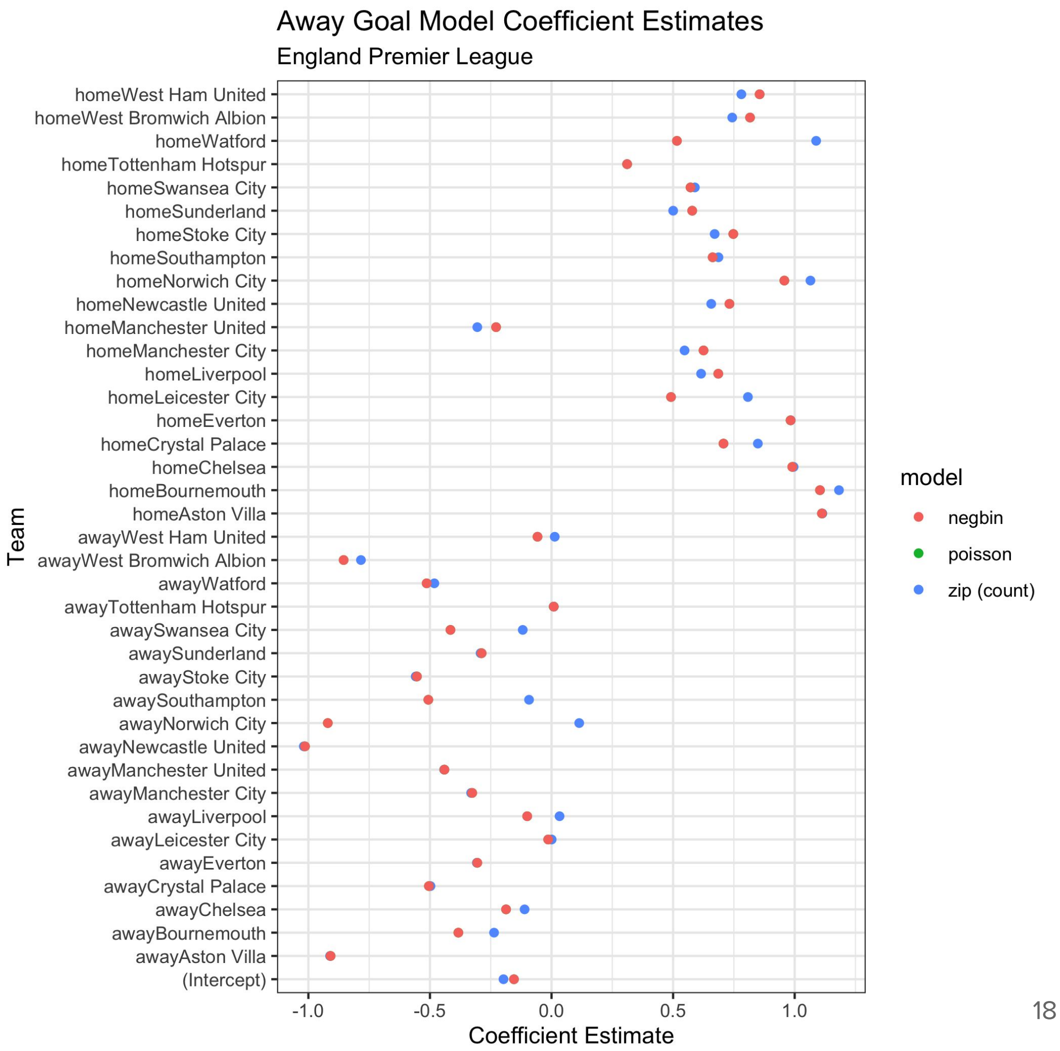
Coefficient Comparison for England Premier League's Home Team Goal

Formula: home_team_goal ~ away_team_name + home_team_name



Coefficient Comparison:

England Premier League away goals



Brier Score

Introduction to Brier Score

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2$$

- N is the overall number of instances of all classes. R is the number of possible classes in which the event can fall, so in our research R = 3, because we have three possible classes: home team wins, away team wins, a draw
- f_ti here is the probability that’s forecast by our models.
- O_ti here is the actual outcome of the event at instance t (0 if it doesn’t happen, 1 if it does happen)
- The lower the Brier score is for a set of predictions, the better the predictions are calibrated.

England Premier League as an Example

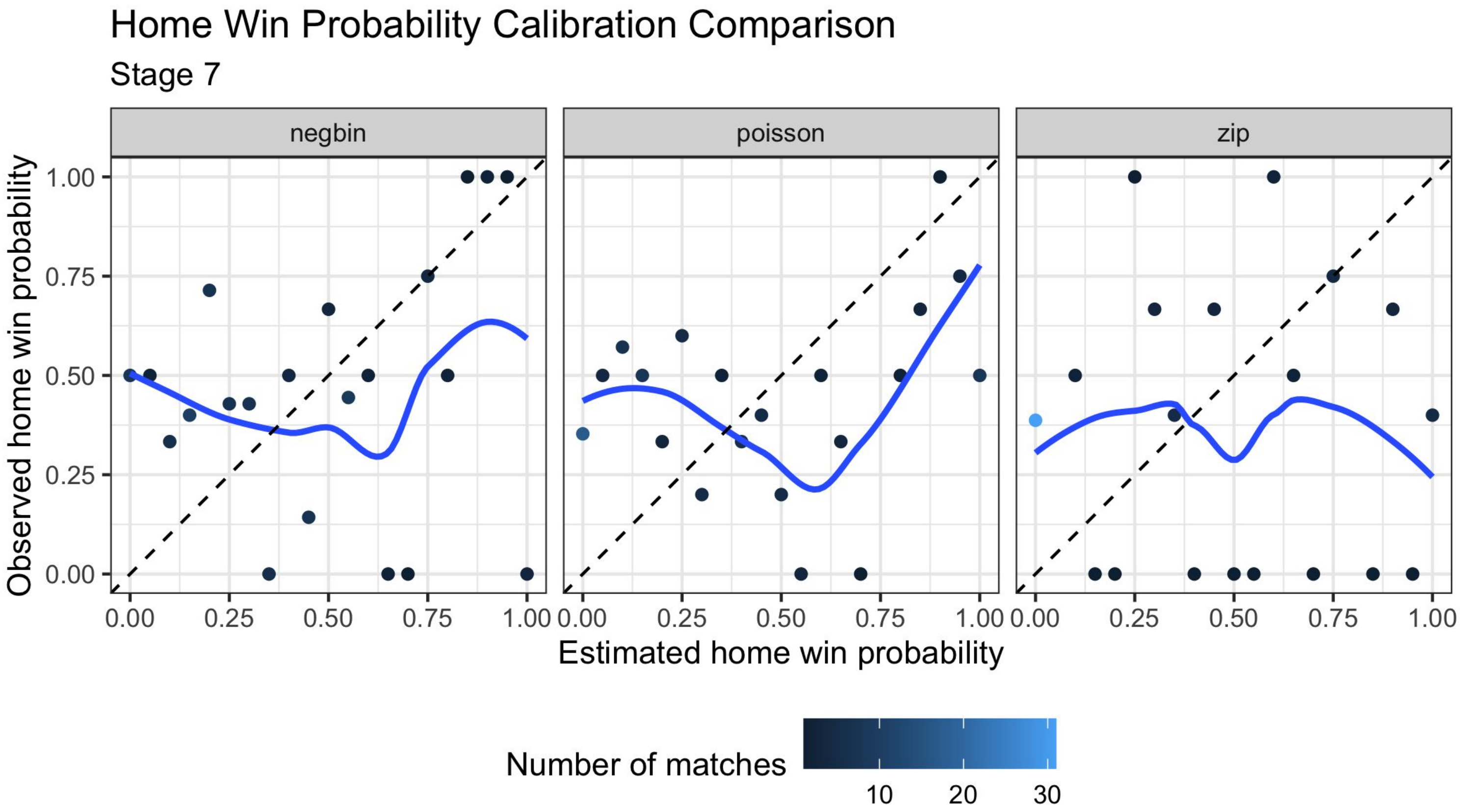
League	Distribution	In-Sample Multiclass Brier Score
England Premier	poisson	0.1936225
	Negative Binomial	0.1938776
	Zero-Inflated Poisson	0.1814058

We found that zero-inflated Poisson always yielded the lowest Brier Score.

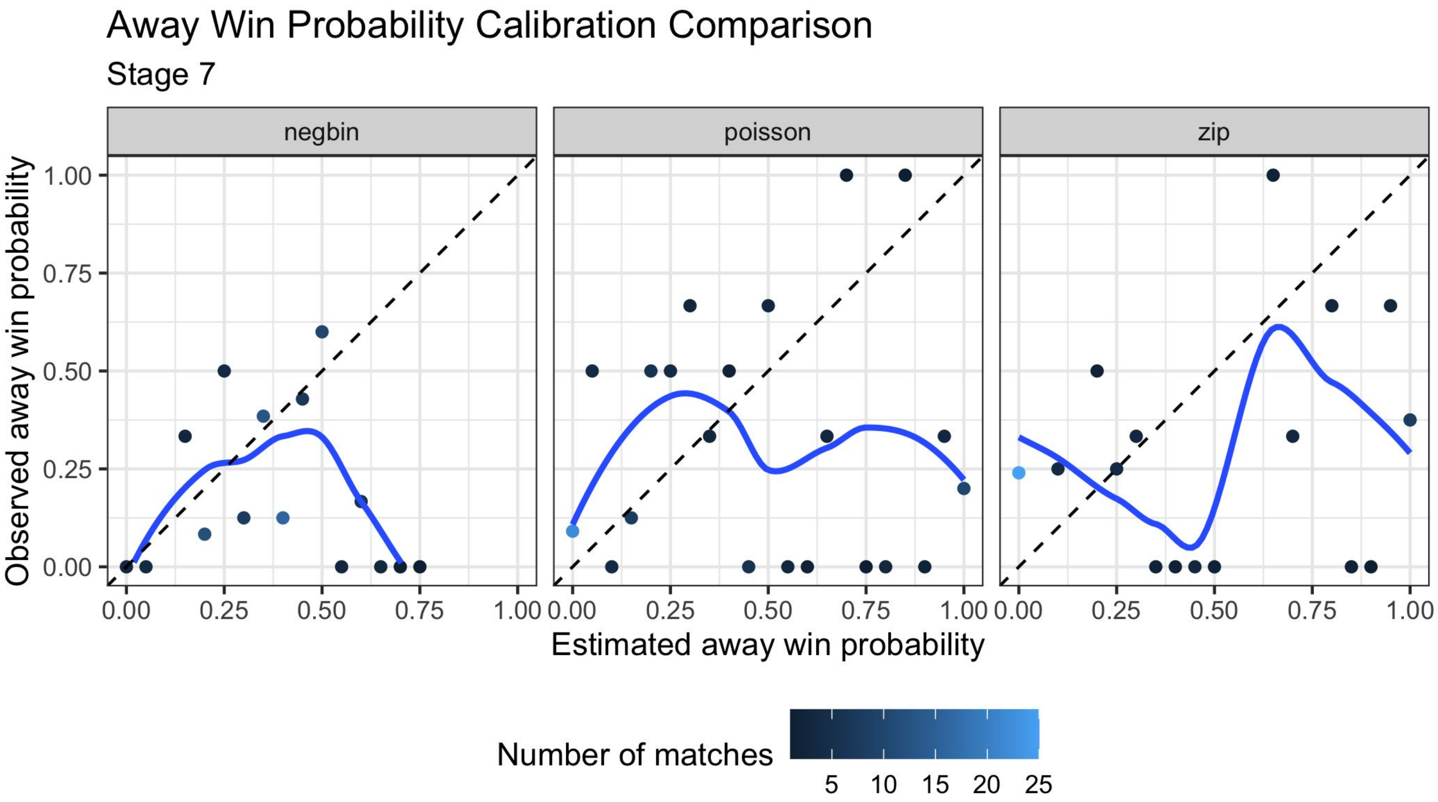
Calibration Curve

Probability Calibration

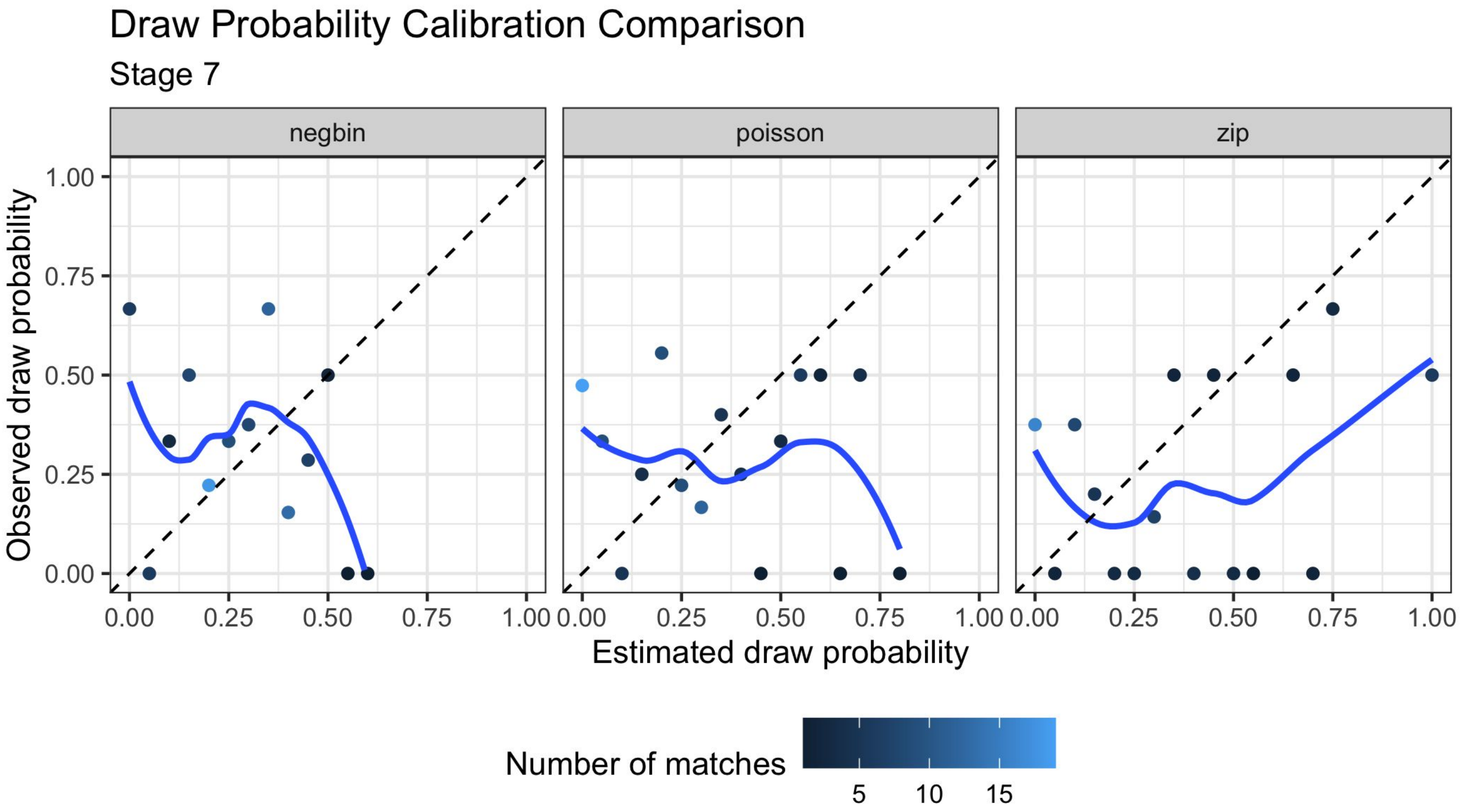
- Time window train-test split over group stage weeks 6-29
- For each league, fit GLMs and store parameters
- Use parameters to simulate test set matches (n = 500)
- For each stage across leagues, compare observed vs. expected win probabilities

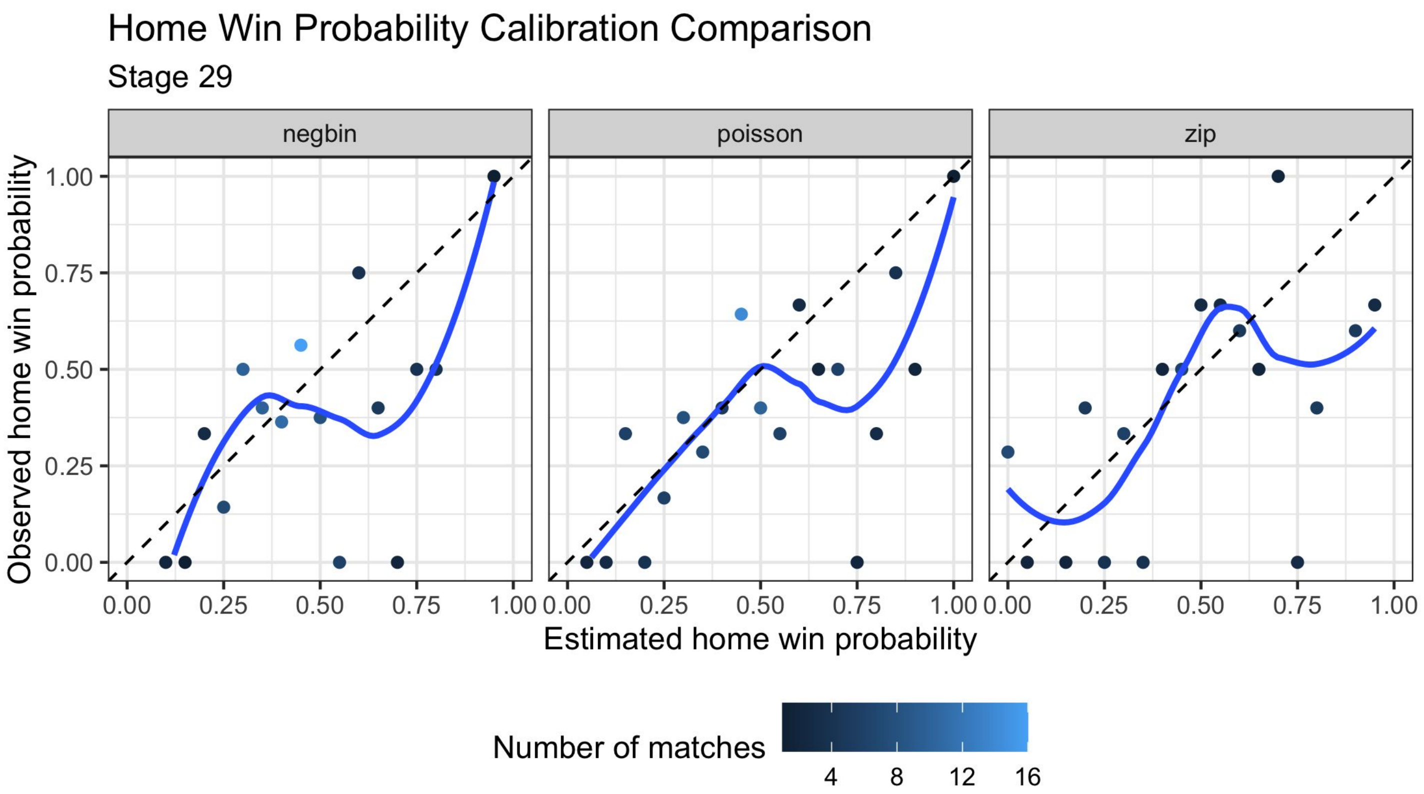


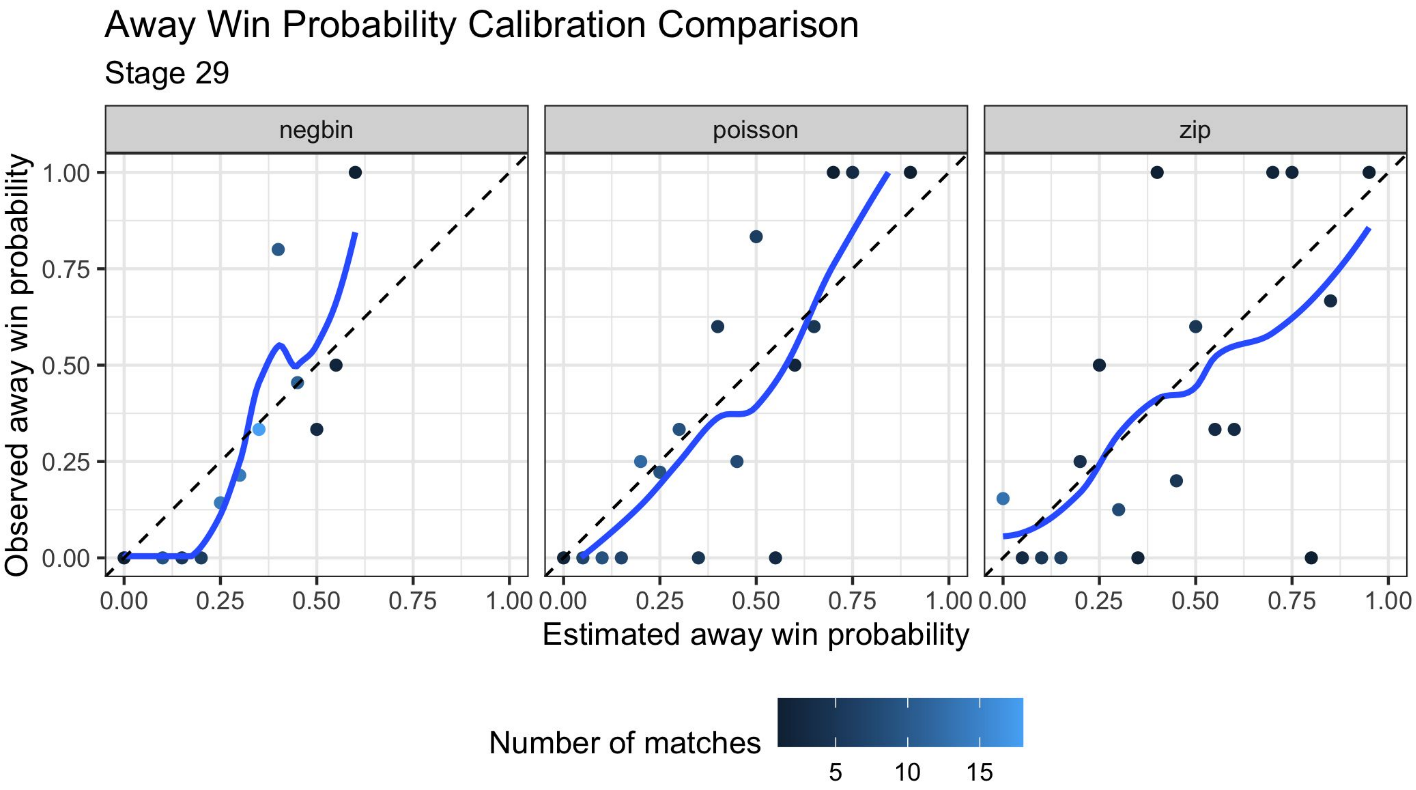
Home Win Probability Calibration Comparison
Stage 7
negbin
poisson
zip
Observed home win probability
0.00
0.25
0.50
0.75
1.00
0.00 0.25 0.50 0.75 1.00
Estimated home win probability
Number of matches
10 20 30



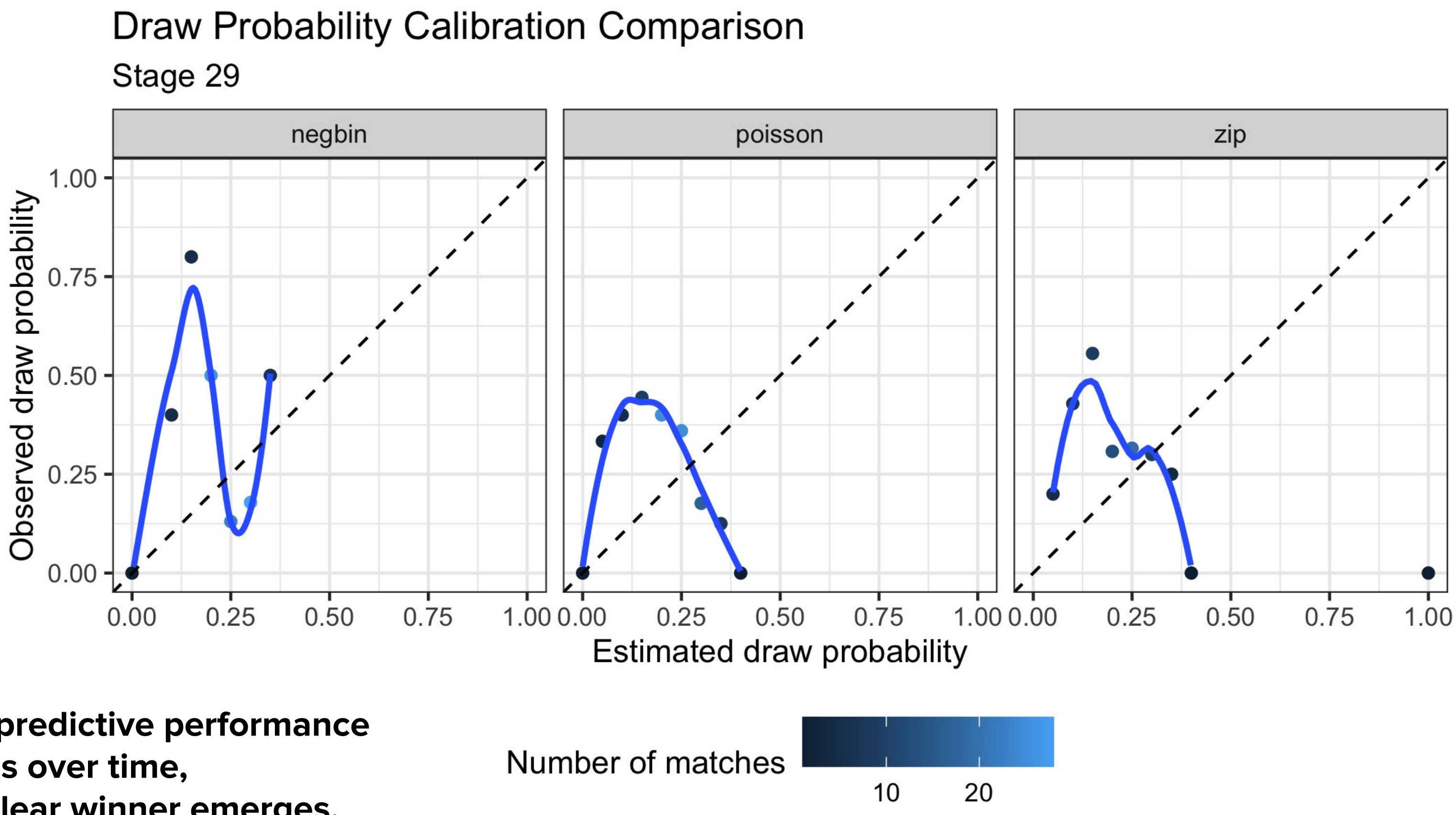
Away Win Probability Calibration Comparison
Stage 7
negbin
poisson
zip
Observed away win probability
Estimated away win probability
0.00
0.25
0.50
0.75
1.00
Number of matches
5
10
15
20
25


Draw Probability Calibration Comparison
Stage 7
negbin
poisson
zip
Observed draw probability
Estimated draw probability
0.00
0.25
0.50
0.75
1.00
Number of matches
5
10
15


Home Win Probability Calibration Comparison
Stage 29
negbin
poisson
zip
Observed home win probability
Estimated home win probability
0.00
0.25
0.50
0.75
1.00
Number of matches
4
8
12
16



Away Win Probability Calibration Comparison
Stage 29
negbin
poisson
zip
Observed away win probability
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
Estimated away win probability
Number of matches
5 10 15



Overall predictive performance improves over time, but no clear winner emerges.

Future Work & Promising Directions

- Figure out whether these results are generalizable to other European leagues.
- Calculate out of sample Multi-Class Brier Score.
- Study the Time effect on rating
- Consider other factors than goal count -> Use Bayesian prior
    - FIFA rating
    - Player injuries
    - Goal type
    - Expected Goals (xG) by FiveThirtyEight
- Modeling goal difference instead of just home/away team goals