Estimating the Change in Soccer's Home Advantage During the COVID-19 Pandemic using Bivariate Poisson Regression

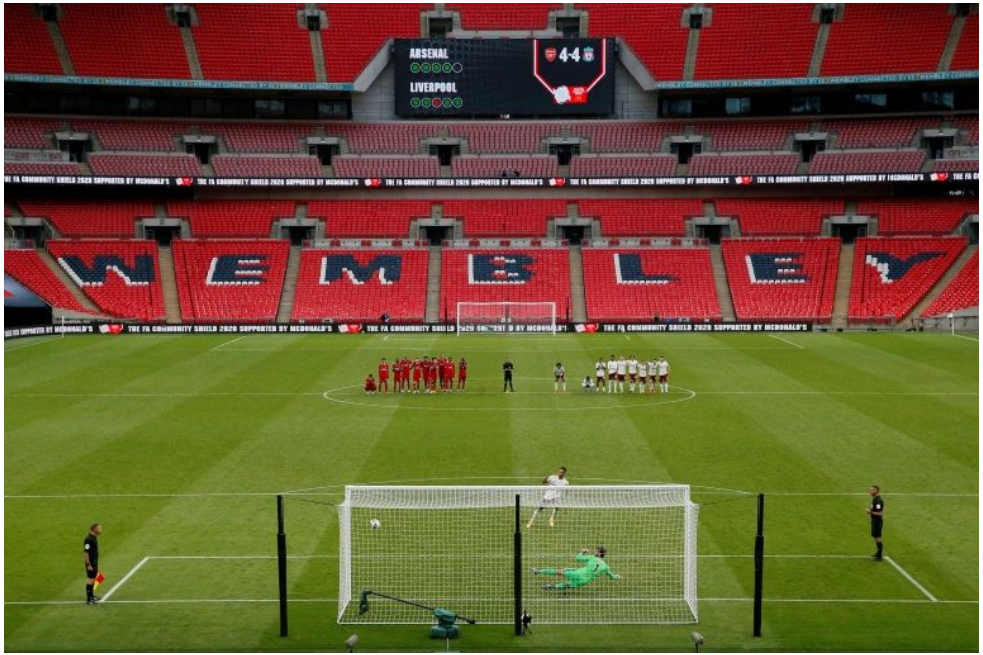
Luke Benz
Medidata Solutions

Michael Lopez
National Football League
Skidmore College

Harvard Sports Analytics Lab Seminar
April 30, 2021

Overview

- COVID-19 hit ~ ⅔ of the way through European soccer season.
- Soccer = first major sport to return to play (returned in May/June 2020 without fans)
- What happened to home advantage (HA) in games without fans?
- How did HA change w/ respect to **goals** and **yellow cards**.



Existing Approaches

- Many of the first papers on the topic make two assumptions:
  1. Home advantage is the same in all leagues and any effect of playing without fans is the same in all leagues.
  2. Soccer outcomes (goals, yellow cards) can be modeled well using linear regression.
- Are these reasonable assumptions?

Existing Approaches

Paper	Leagues	Method	Finding
Sors et al. (2020)	8 (Together)	Correlation	Drop in HA
Leitner and Richlan (2020b)	8 (Together)	Correlations	Drop in HA
Endrich and Gesche (2020)	2 (Together)	Linear Regression	Drop in HA
Fischer and Haucap (2020b)	3 (Separate)	Linear Regression	Mixed
Dilger and Vischer (2020)	1 (NA)	Linear Regression, Correlations	Drop in HA
Krawczyk et al. (2020)	4 (Separate)	Linear Regression	Mixed
Ferraresi et al. (2020)	5 (Together)	Linear Regression	Drop in HA
Reade et al. (2020)	7 (Together)	Linear Regression	Drop in HA
Jiménez Sánchez and Lavín (2020)	8 (Separate)	Linear Regression, Correlations	Mixed
Scoppa (2020)	10 (Together)	Linear Regression	Drop in HA
Cueva (2020)	41 (Together)	Linear Regression	Drop in HA
McCarrick et al. (2020)	15 (Together)	Linear, Poisson Regression	Drop in HA
Bryson et al. (2020)	17 (Together)	Linear, Poisson Regression	Mixed
Benz and Lopez (this manuscript)	17 (Separate)	Bivariate Poisson Regression	Mixed

Data

League	Country	Tier	Restart Date	Pre-Covid Games	Post-Covid Games	# of Team-Seasons
German Bundesliga	Germany	1	2020-05-16	1448	82	90
German 2. Bundesliga	Germany	2	2020-05-16	1449	81	90
Danish Superliga	Denmark	1	2020-05-31	1108	74	68
Austrian Bundesliga	Austria	1	2020-06-02	867	63	54
Portuguese Liga	Portugal	1	2020-06-03	1440	90	90
Greek Super League	Greece	1	2020-06-06	1168	58	78
Spanish La Liga 2	Spain	2	2020-06-10	2233	129	110
Spanish La Liga	Spain	1	2020-06-11	1790	110	100
Turkish Super Lig	Turkey	1	2020-06-13	1460	70	90
Swedish Allsvenskan	Sweden	1	2020-06-14	960	198	80
Norwegian Eliteserien	Norway	1	2020-06-16	960	175	80
English Premier League	England	1	2020-06-17	1808	92	100
Italy Serie B	Italy	2	2020-06-17	2046	111	105
Swiss Super League	Switzerland	1	2020-06-19	836	65	50
Russian Premier Liga	Russia	1	2020-06-19	1136	64	80
English League Championship	England	2	2020-06-20	2673	113	120
Italy Serie A	Italy	1	2020-06-20	1776	124	100

17 European leagues in 13 countries between 2015/16 - 2019/20, scraped from fbref.com

Bivariate Poisson Model

$$(Y_{Hi}, Y_{Ai}) = BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$$

$$\log(\lambda_{1i}) = \mu_{ks} + T_k + \alpha_{H_i ks} + \delta_{A_i ks},$$
$$\log(\lambda_{2i}) = \mu_{ks} + \alpha_{A_i ks} + \delta_{H_i ks},$$
$$\log(\lambda_{3i}) = \gamma_k.$$

- Goals for Home (***H***) and Away (***A***) teams in game ***i.***
- $\boldsymbol{\lambda}_{1i} + \boldsymbol{\lambda}_{3i}$ = goal expectation for $\boldsymbol{Y}_{Hi}$ and $\boldsymbol{\lambda}_{2i} + \boldsymbol{\lambda}_{3i}$ = goal expectation for $\boldsymbol{Y}_{Ai}$ with $\boldsymbol{\lambda}_{3i}$ representing the covariance between Y_{Hi} and Y_{Ai}
- Intercept term for expected goals in season ***s*** in league ***k***
- Attacking (***α***) and defensive (***δ***) team strengths
- Home advantage for league ***k***

A “Common” Linear Model

- Many papers fit a linear regression model that roughly follows the structure below

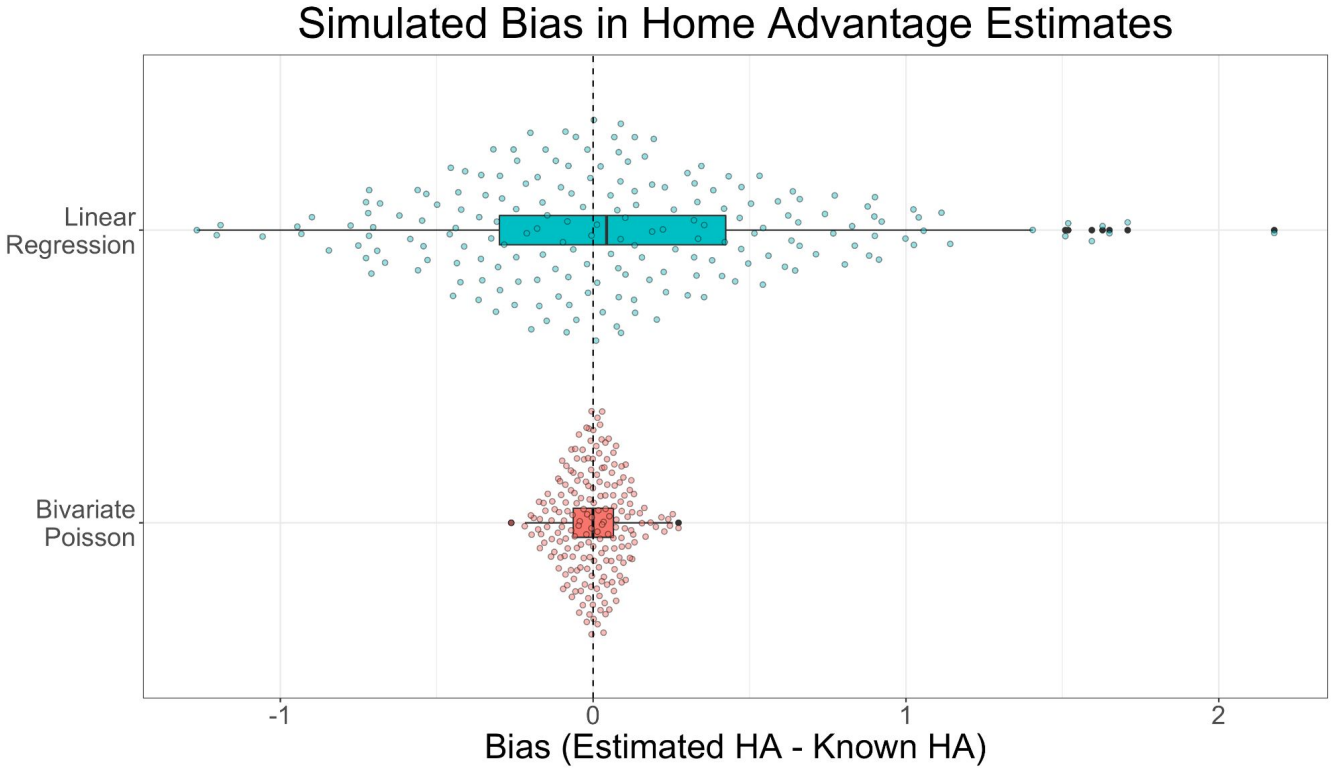
$$GD_i = \alpha + home_{H_i} + away_{A_i} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- Goal differential (home goals - away goals) for game *i*
- Fixed effects for home and away team respectively
- Home advantage

Simulations

- Generate 200 simulated seasons using Bivariate Poisson
  - 20 teams play double round robin (home and away)
  - Attacking (α) and defensive (δ) team strengths randomly drawn from independent $\mathbf{N(0, 0.35^2)}$
  - Home advantage (T) drawn from $Unif(0, log(2))$
  - Use above to generate λ_{1i} and λ_{2i}
  - Draw scores (Y_{1i}, Y_{2i}) from $BVP(\lambda_{1i}, \lambda_{2i}, 0)$

- Fit Bivariate Poisson Model and Linear Regression Model, compare HA estimates to “true” HA for each simulated season



Model	Mean Absolute Bias (Goals)	Mean Bias (Goals)
Bivariate Poisson	0.075	0.006
Linear Regression	0.45	0.09

Bivariate Poisson Model COVID Version (Goals)

$$(Y_{Hi}, Y_{Ai}) = BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$

$$\log(\lambda_{1i}) = \mu_{ks} + T_k \times (I_{pre-Covid}) + T'_k \times (I_{post-Covid}) + \alpha_{H_i ks} + \delta_{A_i ks},$$

$$\log(\lambda_{2i}) = \mu_{ks} + \alpha_{A_i ks} + \delta_{H_i ks},$$

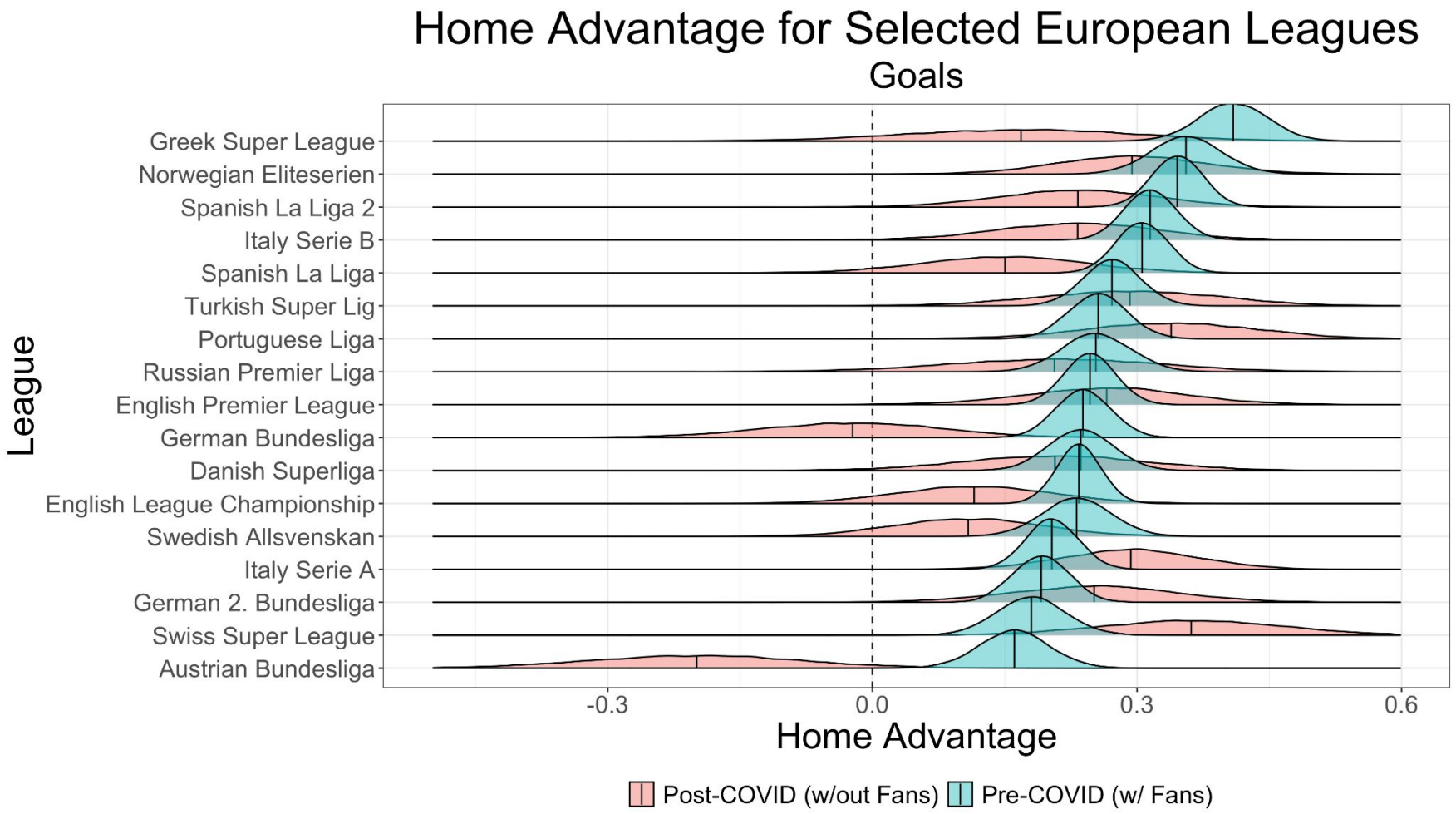
$$\log(\lambda_{3i}) = \gamma_k,$$

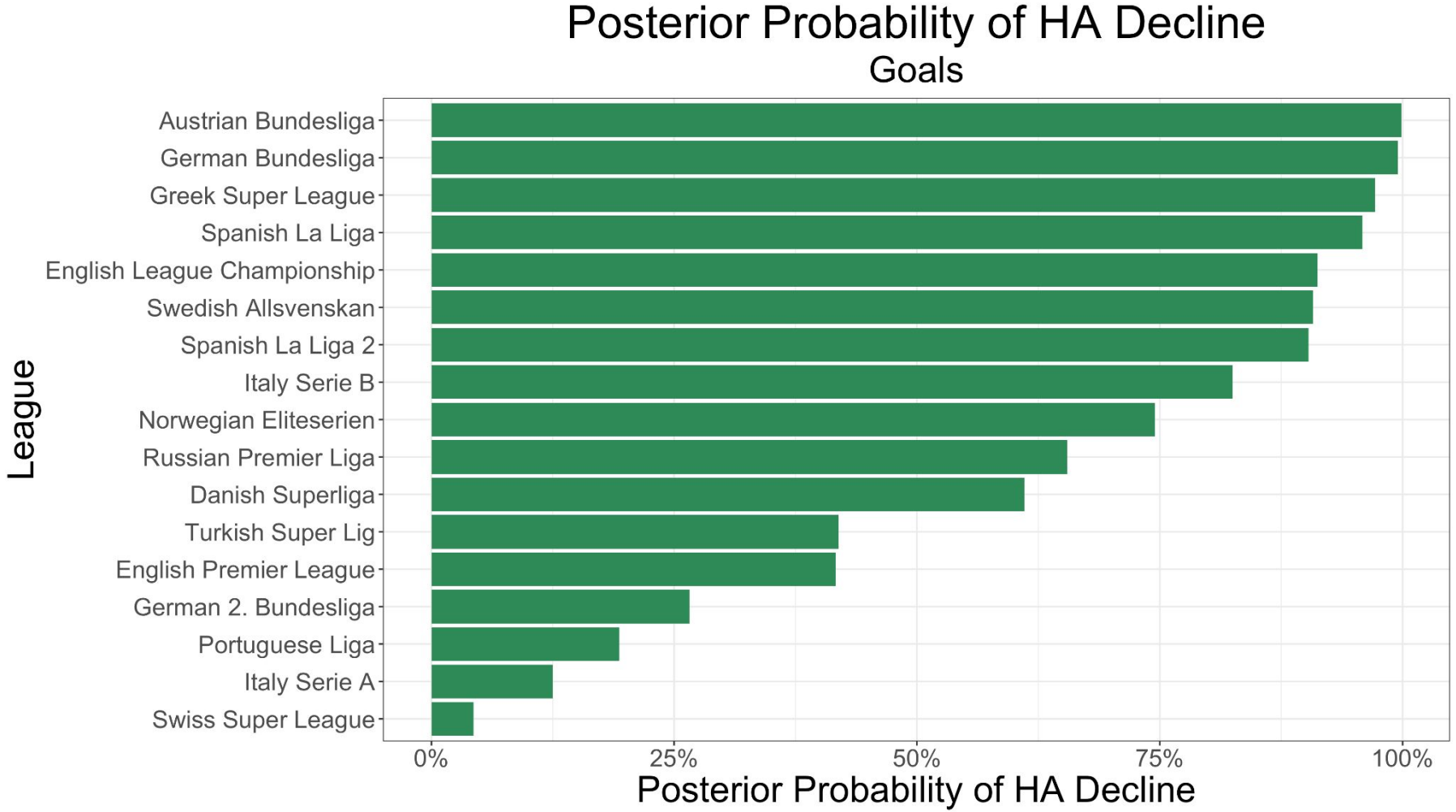
- Home advantage terms pre- and post-COVID restart
- Take $\boldsymbol{\lambda}_{3i} = 0$ based on empirical evidence and theoretical considerations
  - Draw rate lower than data used in [Karlis and Ntzoufras, 2003] since switch to 3/1/0 point system vs. 2/1/0 system for win/draw/loss
  - Empirical correlations between home/away goals range between -0.16 and 0.07

Model Fit Details

- Fit Bayesian bivariate poisson model in STAN
- Major benefit of Bayes is P(HA Decline) for each league
- Weakly informative priors
- 3 chains, 7000 iterations, and a burn in of 2000 draws
- Assess convergence using R-Hat [Gelman et al., 2013]

$$\mu_{ks} \sim N(0, 5),$$
$$\alpha_{ks} \sim N(0, \sigma_{att,k}),$$
$$\delta_{ks} \sim N(0, \sigma_{def,k}),$$
$$\sigma_{att,k} \sim \text{Inverse-Gamma}(1, 1),$$
$$\sigma_{def,k} \sim \text{Inverse-Gamma}(1, 1),$$
$$T_k \sim N(0, 5),$$
$$T'_k \sim N(0, 5)$$

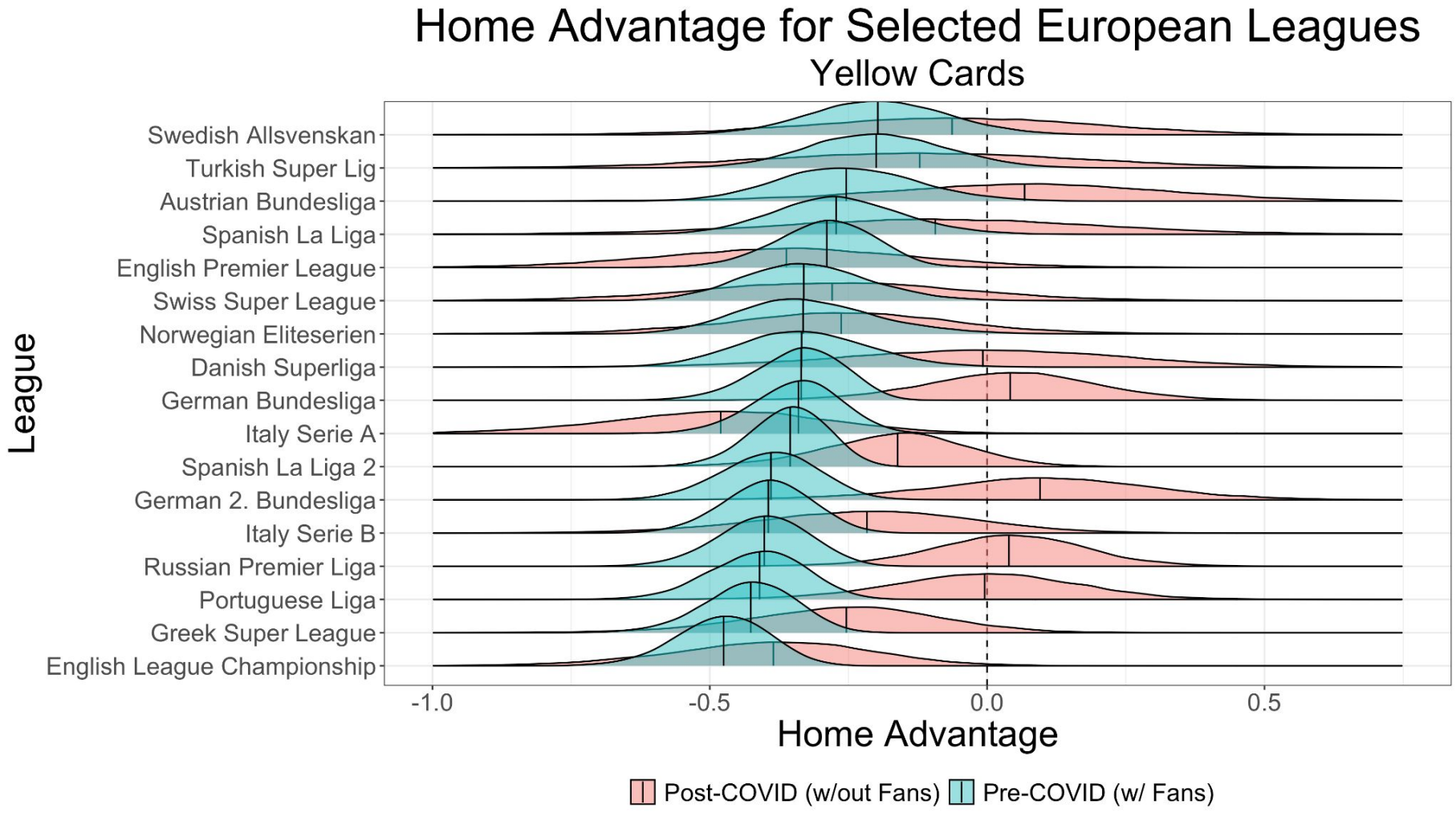

Home Advantage for Selected European Leagues
Goals
League
Greek Super League
Norwegian Eliteserien
Spanish La Liga 2
Italy Serie B
Spanish La Liga
Turkish Super Lig
Portuguese Liga
Russian Premier Liga
English Premier League
German Bundesliga
Danish Superliga
English League Championship
Swedish Allsvenskan
Italy Serie A
German 2. Bundesliga
Swiss Super League
Austrian Bundesliga
-0.3
0.0
0.3
0.6
Home Advantage
Post-COVID (w/out Fans)
Pre-COVID (w/ Fans)

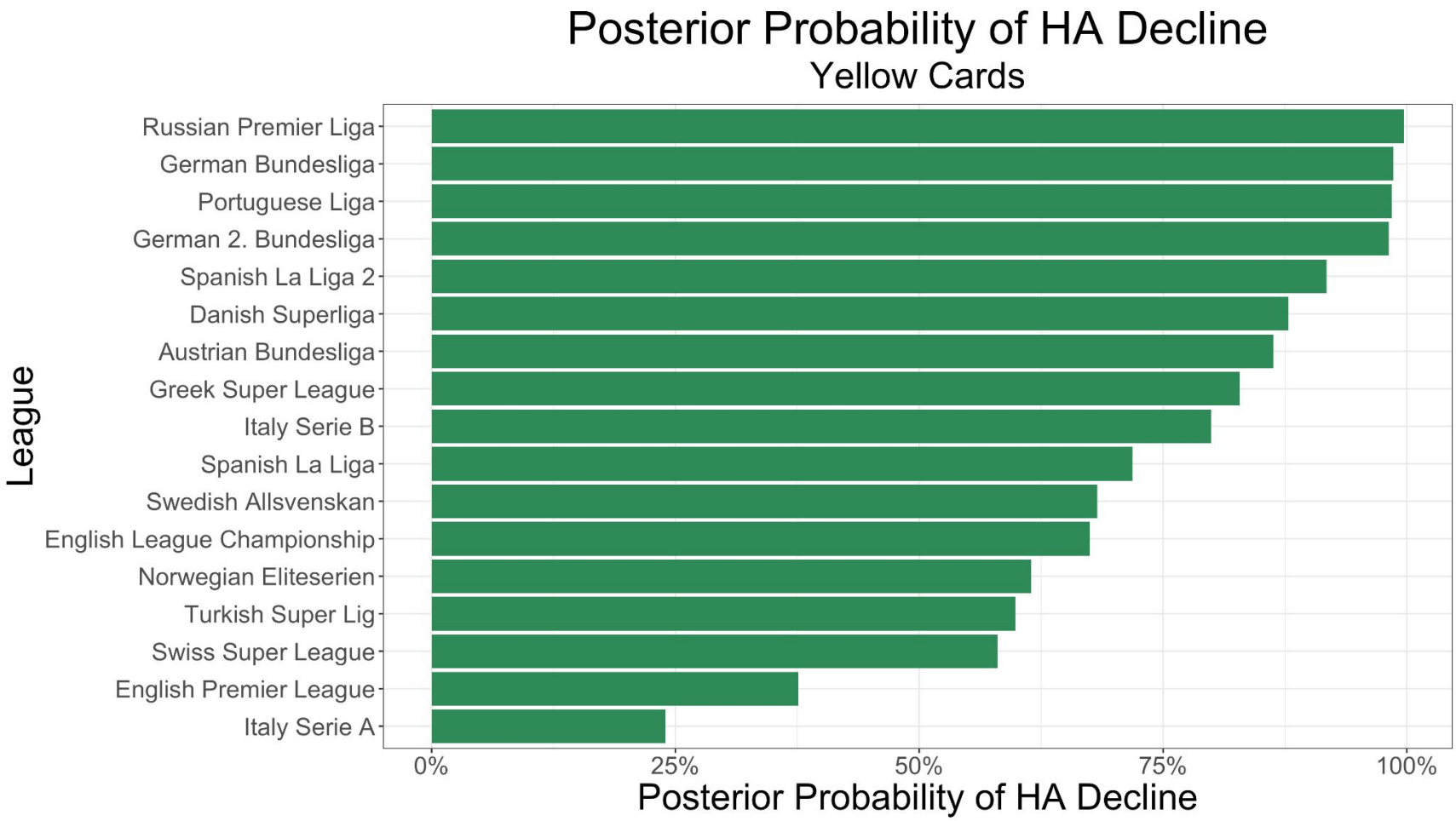

Posterior Probability of HA Decline
Goals
Austrian Bundesliga
German Bundesliga
Greek Super League
Spanish La Liga
English League Championship
Swedish Allsvenskan
Spanish La Liga 2
Italy Serie B
Norwegian Eliteserien
Russian Premier Liga
Danish Superliga
Turkish Super Lig
English Premier League
German 2. Bundesliga
Portuguese Liga
Italy Serie A
Swiss Super League
League
0%
25%
50%
75%
100%
Posterior Probability of HA Decline

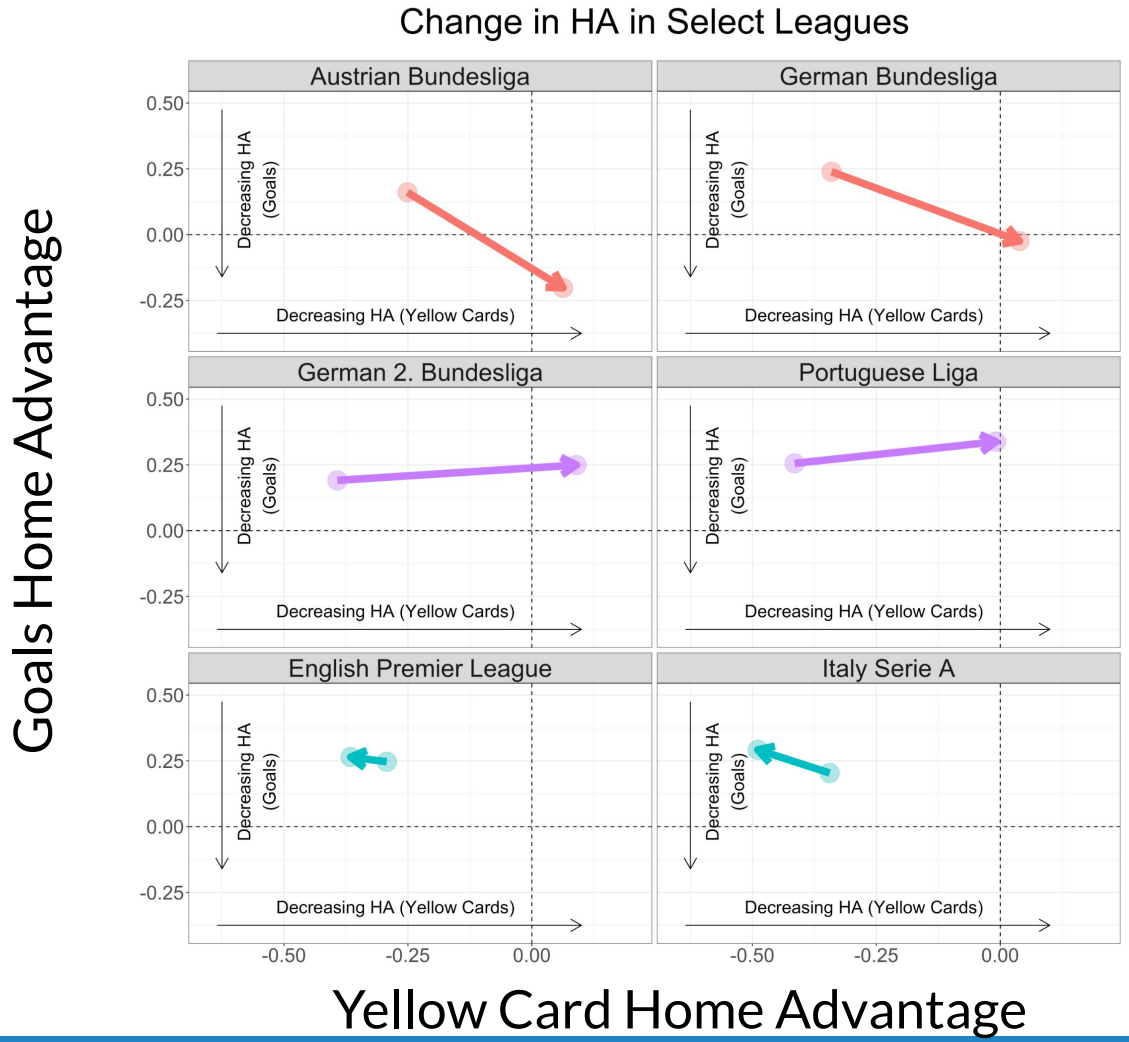
Bivariate Poisson Model COVID Version (YC)

$$(Z_{Hi}, Z_{Ai}) = BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$
$$\log(\lambda_{1i}) = \mu_{ks} + T_k \times (I_{pre-Covid}) + T'_k \times (I_{post-Covid}) + \tau_{H_i ks},$$
$$\log(\lambda_{2i}) = \mu_{ks} + \tau_{A_i ks},$$
$$\log(\lambda_{3i}) = \gamma_k,$$

- Yellow card team random effects. Notice we don't have 2 per team as with goals
- Allow $\boldsymbol{\lambda}_{3i} > 0$ based on empirical evidence and theoretical considerations
  - Empirical correlations between home/away goals range between 0.10 and 0.22


Home Advantage for Selected European Leagues
Yellow Cards
League
Swedish Allsvenskan
Turkish Super Lig
Austrian Bundesliga
Spanish La Liga
English Premier League
Swiss Super League
Norwegian Eliteserien
Danish Superliga
German Bundesliga
Italy Serie A
Spanish La Liga 2
German 2. Bundesliga
Italy Serie B
Russian Premier Liga
Portuguese Liga
Greek Super League
English League Championship
-1.0
-0.5
0.0
0.5
Home Advantage
Post-COVID (w/out Fans)
Pre-COVID (w/ Fans)


Posterior Probability of HA Decline
Yellow Cards
League
Russian Premier Liga
German Bundesliga
Portuguese Liga
German 2. Bundesliga
Spanish La Liga 2
Danish Superliga
Austrian Bundesliga
Greek Super League
Italy Serie B
Spanish La Liga
Swedish Allsvenskan
English League Championship
Norwegian Eliteserien
Turkish Super Lig
Swiss Super League
English Premier League
Italy Serie A
0%
25%
50%
75%
100%
Posterior Probability of HA Decline


Change in HA in Select Leagues
Goals Home Advantage
Yellow Card Home Advantage
Austrian Bundesliga
German Bundesliga
German 2. Bundesliga
Portuguese Liga
English Premier League
Italy Serie A
Decreasing HA (Goals)
Decreasing HA (Yellow Cards)
0.50
0.25
0.00
-0.25
-0.50

Discussion

- Did HA decline? In many leagues yes, in others no!
- Not always the case that changes in yellow card HA are linked to changes to goal HA.
    - Not just ‘less referee bias’ but (also) may be difference in player behavior?
- Estimates looking at the impact of HA post-Covid are less of a statement about the cause and effect from a lack of fans, as much as they are about changes due to **both a lack of fans** and changes to training due to Covid-19.

"It's horrible to play without fans. It's not a nice feeling. Not seeing anyone in the stadium makes it like training, and it takes a lot to get into the game at the beginning."

- Lionel Messi (Reuters, 2020)

Summary

- Bivariate Poisson model better suited to model soccer outcomes estimate home advantage than linear regression.
- HA was not the same in each country before the pandemic, and did not affected different in every league post-COVID return.
    - Should not be combining leagues to analyze together.
- HA declined in many, but not all leagues, due to factors including but not limited to lack of fans in the stands.

Links

- Paper: https://arxiv.org/abs/2012.14949
- Code: https://github.com/lbenz730/soccer_ha_covid
- Slides: https://lukebenz.com/slides/harvard_soccer_covid.pdf
- Twitter:
    - Luke: @recspecs730
    - Mike: @StatsByLopez

References

- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian Data Anal-ysis, 3rd edn. CRC Press, Boca Raton, FL.
- Karlis D, Ntzoufras I (2003) Analysis of sports data by using bivariate poisson models. Journal of the Royal Statistical Society: Series D (The Statistician) 52(3):381–393.
- Reuters (2020) Lionel Messi Says Playing Without Fans is 'Horrible and Ugly'. URL https://www.eurosport.com/football/liga/2020-2021/lionel-messi-says-playing-without-fans-is-horrible-and-ugly-after-barcelona-star-collects-pichichisto8042397/story.shtml.