COVID-19 countermeasures, Major League Baseball, and the home field advantage: Simulating the 2020 season using logit regression and a neural network [version 1; peer review: awaiting peer review]

Justin Ehrlich
Shankar Ghimire

1Department of Sport Analytics, Syracuse University, Syracuse, NY, 13244, USA
2Department of Economics and Decision Sciences, Western Illinois University, Macomb, IL, 61455, USA

Abstract

Background: In the wake of COVID-19, almost all major league sports have been either cancelled or postponed. The sports industry suffered a major blow with the uncertainty of sporting events being held in the near future. Various scenarios of how and when sports might recommence have been discussed. This paper examines various scenarios of how Major League Baseball team performance is going to be impacted by the presence of fans, or the lack thereof, in the context of physical distancing and other COVID-19 countermeasures.

Methods: The paper simulates, using a neural network and a logit regression model, the win-loss probabilities for various scenarios under consideration and also estimates the home effect for each team using data for the 2017-2019 seasons.

Results: The model demonstrates that individual team home effect is symmetric between home and away and teams will not necessarily have a win or loss of any additional games in neutral stadiums, as teams with a high home field effect will lose more neutral games that would have been at home but will win more neutral games that would have been away. However, the result of individual games will be different since home effect is asymmetric between teams. Our simulation demonstrates that these individual game differences may lead to a slight difference in Play-Off Berths between a full season, a half season, or a full season without fans.

Conclusions: Without fans, any advantage (or disadvantage) from home field advantage is removed. Our models and simulation demonstrate that this will reduce the variance. This stabilizes the outcome based upon true team talent, which we estimate will cause a larger divide between the best and worst teams. This estimation helps decision makers understand how individual team performance will be impacted as they prepare for the 2020 season under the new circumstances.

Keywords
MLB, Baseball, COVID-19, Neural Network, Logit
**Introduction**

The 2019–2020 pandemic from the novel coronavirus (COVID-19) has brought unprecedented countermeasures to every sector of the economy, including individuals, groups, institutions, and industries. The sports industry took one of the biggest hits, with all major leagues in the U.S. cancelling or halting their events. While these actions were necessary to address the public health concern, each segment is now floating various proposals to resume operations and give some relief to the significant portion of the economy that the sports industry comprises.

Major League Baseball (MLB) is likely to be the first American professional sporting league to resume, probably in May or June (Passan, 2020). Players are willing (although not all players agree to the method), the League is willing, the Arizona government is willing, and health professionals have approved a plan to move forward, known as the Arizona Plan. This plan calls for players, coaches, and staff to be quarantined in hotels around the Phoenix area, and to play in empty ballparks that include the ten Cactus League Spring Training parks, Chase Field, and other Phoenix ballparks. One interesting aspect of this arrangement is that the stadium size will not matter because there will not be any fans. This will be an opportunity for MLB to get back into the spotlight and accumulate massive television viewership that MLB has not seen in decades. The experience will be completely optimized for TV viewing, and so the league will finally be able to experiment with proposed rule changes, including removing mound visits to make the game go faster, adding a Robo Umpire, which has already been successfully tested last season via a partnership with the independent Atlantic League (Bogage, 2019), and an expanded roster giving players more rest due to the extremely hot temperatures of Phoenix. While all of this will alter predictions on who’s going to the playoffs, probably the biggest impact that this plan will have on the games is the lack of the home field advantage (HFA): the advantage that the home team has over the visiting team due to the home team having fans, the familiarity of the home team to their own ball park, and the away team having to travel.

Baseball has been shown in previous studies to be less susceptible to the HFA effect than other professional sports (Edwards & Archambault, 1979; Gómez et al., 2011; Pollard et al., 2017). Despite this, there is a measurable home field advantage in baseball, as shown by Jones (2015); Jones (2018). Building on this, we extend the analysis for the MLB under uncertainty of which scenario the League will be following for the 2020 season. In particular, we simulate the win-loss probabilities for three different scenarios as well as estimate the home advantage for each team using the past three seasons’ data. This estimation helps us understand how individual team performance is going to be impacted as they prepare for the 2020 season in the new circumstances.

**Methods**

**Data sources**

We use the MLB 2017–2019 season data for the 30 teams represented in the league. The data were obtained from the MLB Advanced Media’s Baseball Savant Website using the Python package PyBaseball 1.0.4 (LeDoux, 2017/2020). The data shows that out of the 7,290 home games played during the 2017–2019 seasons, 3,881 (53.237%) resulted in wins and the remaining 3,409 (46.763%) resulted in a loss. Next, we seek to quantify the HFA’s role in this difference.

**Calculating home advantage**

There are various techniques to calculate the home advantage depending on the sport, gender, league, and the nature of scoring (Jones, 2015). Pollard et al. (2017) use a general linear model to fit the home advantage. However, because we have a categorical variable of win or lose, we need to follow a non-linear approach. To test the hypothesis that teams have home-field advantage, we apply a logit regression model to predict the probability of winning as a function of home game dummy, team fixed effects, opponent fixed effects, and the win-loss records. We estimate the following regression equation:

\[ \text{Win}_i = \alpha_1 + \alpha_2 \text{Home}_i + \alpha_3 \text{Team}_i + \alpha_4 \text{Home} \times \text{Team}_i + \alpha_5 \text{Home} \times \text{Opp}_i + \eta Z_i + \epsilon_i \]  (1)

where \( \text{Win}_i \) is a dummy variable that takes the value 1 if the recorded game resulted in a win for the team-opponent pair, and zero otherwise; \( \text{Home}_i \) accounts for home game; and \( \text{Team}_i \) controls for the individual team fixed effects, and \( \text{Opp}_i \) controls for the opponent fixed effects, \( Z_i \) represents the win-loss percentage for the team as well as the opponent, \( \epsilon_i \) stands for the error term. The HFA is calculated accounting for the team fixed effects as well as the opponent fixed effects by interacting home with team and opponent separately. We run the logit model on all the data, with \( \text{Home}_i=1 \) for the home team and equal to zero for the travelling team. Doing so separates the team fixed effects and home field advantage. The model in Equation 1 is used to estimate both the win probability and the HFA per team. The HFA is obtained by calculating the marginal effect (ME) of \( \text{Home}_i \) on the win probability for each team separately.

**Development of the neural network model**

A neural network model was also created to act as a robustness check for the logit win prediction model. The software to train the model is hosted on GitHub (Ehrlich, 2020a). We used the R package caret 6.0–86 (Kühn et al., 2020). We developed a simulator to estimate what might happen if: 1) The full 2020 season continued on in a parallel universe devoid of COVID-19; 2) MLB waits and is able to return and play half a season to packed stadiums around the All Star Break, which is assuming an extremely optimistic timeline of a return to normal life; 3) a full season is played without fans, which is likely the only way they will be able to play this season (i.e., the Arizona Plan). The simulation was executed 100 times, and the logit win prediction model was used as the basis for predicting each win. A random number between 0 and 1 was generated and checked against the win probability provided by the model. If the random number was below the probability, then the team won, otherwise the team lost.
Results
The summary statistics of the training data is contained in Table 1. The logit results from Equation 1, without the fixed effects, are reported in Table 2. Both log odds ratios and the MEs are reported in this table. The results show that the individual regressors included in the model show plausible impacts. Looking at the log-odds ratios, home games and the home team’s previous win-loss percentage (WL%) are more likely to result in a win but the opponent’s WL% is less likely to result a loss for the home team. These results support the presence of the HFA. The right half of the table shows the MEs for each variable. We are mainly interested in the MEs for the Home variable, which is 0.064. This means, the marginal probability of winning a game at home versus away field goes up by 6.4%. This is the average HFA for all of the teams as a whole. The HFA for each team is presented in Figure 1. In our sample, PHI seems to have the highest home advantage and HOU seems to have the lowest (negative, in fact) home advantage.

The model was trained using the 2017–2019 MLB regular season games. The schedule for the 2020 season was estimated using the schedule from the 2019 season. While the dates will be off slightly, the team pairings will be nearly the same. The wins and losses of the 100 simulations were added to form the result of the 2020 season. The overall results are visualized in Figure 2, while the divisional results shown in Table 3. Table 4 provides statistics calculated during each season and averaged. This includes the correlation between the full season and the half and no-fan seasons using both the overall rankings and the win-loss percent (WL%). The full seasons rank correlations are higher with the no-fans seasons (0.825) than the half seasons (0.735). The correlations using WL% is similar. The standard deviation of the predicted win probabilities is lower for the no-fans seasons (0.073) than the full (0.085) and half seasons (0.085). The home effect was correlated with the win probabilities’ standard deviations and is negative for the no fans seasons (-0.221). In other words, the higher the home effect, the lower the variance.

This neural network was also used as the win predictor in 100 simulations and the results are very similar to the logit win prediction model, which shows robustness in the simulation results. Table 4 shows the statistical results of both models. The correlation and standard deviation differences are approximately the same between the two models.

The results of the simulations are available as Extended data (Ehrlich, 2020b) and the code necessary for replicating the results, including training the models, are hosted on GitHub (Ehrlich, 2020a).

Table 1. Summary statistics used for training the model.

| Variable  | Mean | SD  | Min  | Median | Max  |
|-----------|------|-----|------|--------|------|
| Home      | 0.500| 0.500| 0.000| 0.500  | 1.000|
| Prev WL % | 0.500| 0.114| 0.000| 0.500  | 1.000|
| PrevWL% Opp| 0.500| 0.114| 0.000| 0.500  | 1.000|
| Season    | 2018.000| 0.816| 2017.000| 2018.000| 2019.000|

Table 2. Results of regression analysis.

| Predictors             | Logit win prediction model | Logit win prediction marginal effects |
|------------------------|----------------------------|--------------------------------------|
|                        | Odds ratios | Std. error | p | AME | Std. error | p   |
| (Intercept)            | 0.876       | 0.174       | 0.447 |     |            |     |
| Home                   | 1.304       | 0.034       | <0.001 | 0.064 | 0.008       | <0.001 |
| PrevWLPerC             | 2.307       | 0.167       | <0.001 | 0.203 | 0.040       | <0.001 |
| PrevWLPerC Opp         | 0.433       | 0.167       | <0.001 | -0.203 | 0.040       | <0.001 |
| WLPerC Season Prev     | 8.251       | 0.245       | <0.001 | 0.513 | 0.060       | <0.001 |
| WLPerC Season Pre Opp  | 0.121       | 0.245       | <0.001 | -0.512 | 0.060       | <0.001 |
| Observations           | 14580       |            |     |     |            |     |
| R² Tjur                | 0.029       |            |     |     |            |     |
Figure 1. MLB home field advantage effect of individual teams.

Figure 2. MLB season 2020 change in simulated rank after 100 simulations.
Table 3. Results of the simulation using the logit win prediction model.

| Division | Tm   | Rank | Berth | WL%  | Rank | Berth | WL%  | Rank | Berth | WL%  | Home effect |
|----------|------|------|-------|------|------|-------|------|------|-------|------|-------------|
| AL Central | CLE  | 1    | y     | 0.593 | 1    | y     | 0.587 | 1    | y     | 0.584 | 0.034       |
| AL Central | MIN  | 2    | w     | 0.573 | 2    |       | 0.565 | 2    | w     | 0.572 | 0.033       |
| AL Central | CHW  | 3    |       | 0.419 | 3    |       | 0.414 | 3    |       | 0.425 | 0.065       |
| AL Central | KCR  | 4    |       | 0.390 | 4    |       | 0.397 | 4    |       | 0.402 | 0.062       |
| AL Central | DET  | 5    |       | 0.337 | 5    |       | 0.340 | 5    |       | 0.341 | 0.050       |
| AL East   | NYY  | 1    | y     | 0.623 | 1    | y     | 0.614 | 1    | y     | 0.607 | 0.115       |
| AL East   | BOS  | 2    |       | 0.569 | 2    |       | 0.573 | 3    |       | 0.568 | 0.003       |
| AL East   | TBR  | 3    |       | 0.567 | 3    |       | 0.564 | 2    | w     | 0.582 | 0.069       |
| AL East   | TOR  | 4    |       | 0.438 | 4    |       | 0.435 | 4    |       | 0.408 | 0.072       |
| AL East   | BAL  | 5    |       | 0.347 | 5    |       | 0.349 | 5    |       | 0.357 | 0.092       |
| AL West   | HOU  | 1    | y     | 0.653 | 1    | y     | 0.650 | 1    | y     | 0.657 | -0.014      |
| AL West   | OAK  | 2    | w     | 0.580 | 2    | w     | 0.576 | 2    |       | 0.566 | 0.115       |
| AL West   | SEA  | 3    |       | 0.473 | 3    |       | 0.468 | 3    |       | 0.491 | 0.016       |
| AL West   | LAA  | 4    |       | 0.469 | 4    |       | 0.467 | 4    |       | 0.462 | 0.056       |
| AL West   | TEX  | 5    |       | 0.463 | 5    |       | 0.454 | 5    |       | 0.436 | 0.070       |
| NL Central | MIL  | 1    | y     | 0.565 | 1    | y     | 0.555 | 1    | y     | 0.553 | 0.071       |
| NL Central | CHC  | 2    | w     | 0.544 | 2    | w     | 0.545 | 2    |       | 0.550 | 0.118       |
| NL Central | STL  | 3    |       | 0.534 | 3    |       | 0.542 | 3    |       | 0.544 | 0.056       |
| NL Central | PIT  | 4    |       | 0.450 | 4    |       | 0.451 | 4    |       | 0.464 | 0.084       |
| NL Central | CIN  | 5    |       | 0.437 | 5    |       | 0.445 | 5    |       | 0.440 | 0.099       |
| NL East   | WSN  | 1    | y     | 0.569 | 1    | y     | 0.565 | 2    | w     | 0.556 | 0.016       |
| NL East   | ATL  | 2    |       | 0.559 | 2    |       | 0.559 | 1    | y     | 0.558 | -0.005      |
| NL East   | NYM  | 3    |       | 0.494 | 3    |       | 0.485 | 3    |       | 0.505 | 0.045       |
| NL East   | PHI  | 4    |       | 0.475 | 4    |       | 0.477 | 4    |       | 0.479 | 0.163       |
| NL East   | MIA  | 5    |       | 0.394 | 5    |       | 0.393 | 5    |       | 0.392 | 0.107       |
| NL West   | LAD  | 1    | y     | 0.632 | 1    | y     | 0.632 | 1    | y     | 0.634 | 0.081       |
| NL West   | ARI  | 2    |       | 0.540 | 2    |       | 0.539 | 2    | w     | 0.553 | 0.049       |
| NL West   | COL  | 3    |       | 0.500 | 3    |       | 0.499 | 3    |       | 0.488 | 0.092       |
| NL West   | SFG  | 4    |       | 0.437 | 4    |       | 0.441 | 4    |       | 0.437 | 0.071       |
| NL West   | SDP  | 5    |       | 0.432 | 5    |       | 0.423 | 5    |       | 0.402 | 0.050       |
Table 4. Key summary statistics of the simulation using the logit win prediction model.

| Model simulated               | Logit       | NN          |
|-------------------------------|-------------|-------------|
| Simulated Seasons             | 100         | 100         |
| Full-NoFans Rank Correlation  | 0.825       | 0.814       |
| Full-Half Rank Correlation    | 0.735       | 0.719       |
| Full-NoFans WL% Correlation   | 0.823       | 0.817       |
| Full-Half WL% Correlation     | 0.734       | 0.718       |
| NoFans WL% SD                 | 0.073       | 0.073       |
| Full WL% SD                   | 0.085       | 0.086       |
| Half WL% SD                   | 0.085       | 0.086       |
| Full WinProb SD-HomeEffect Correlation | 0.361 | 0.251 |
| NoFans WinProb SD – HomeEffect Correlation | -0.221 | -0.268 |

Note: These statistics are calculated for each season and averaged. WL% is predicted based upon the Win Prediction.

Discussion
Based on the above results, since the team-home effect is symmetric between home and away, teams will not necessarily win or lose any additional games in neutral stadiums as teams with a high home field effect will lose more neutral games that would have been at home but will win more neutral games that would have been away. The greater the home-team ME, the less variance there will in the predicted win probabilities. To verify this assumption, we calculated the correlation of HomeEffects and the standard deviation (SD) of win probabilities between a full (0.361) and no-fan season (-0.221). Since the home effect is symmetric for each team (the away field disadvantage = -the home field advantage), decreasing the variance does not affect the overall expected WL% for each team. However, the result of individual games will be different since home effect is asymmetric between teams. For example, if the Cubs (highest home effect in the NL Central) plays the Cardinals (lowest home effect in the NL Central), the Cubs will have a larger advantage playing at home then the Cardinals will have playing at home (besides team fixed effects). These differences are removed with the No-Fan scenario and the outcome will be solely based upon the talent of the teams. However, on average there only a slight change of overall WL% (or playoff berth), just the SD of the results (see Table 3). Without fans, any advantage (or disadvantage) from home field advantage, which cause higher levels of variance, is removed. This stabilizes the outcome based upon true team talent. As fewer games have been played, the half-season will have more upsets, but the SD is close to the same as the full season.

Conclusion
This paper analyzes the previous season MLB data to estimate the win-loss probabilities for the 2020 season for each of the 30 teams in the League using logit regressions and a neural network. The Arizona Plan’s neutralization of HFA would not significantly affect the overall outcome of the season. In fact, our model predicts that the Arizona Plan season will produce season results that are based more on the true talent of the teams. Further, our simulation demonstrates that there will be less variance in the win probability between any two teams, which we estimate will cause a larger divide between the best and worst teams. In conclusion, we believe that the results of the Arizona Plan will be similar to a regular season with fans, and that the teams’ standings at the end of the regular season will be more predictable than a normal season.

Data availability
Source data
Zenodo: Syracuse-University-Sport-Analytics/MLBCovid19.
https://doi.org/10.5281/zenodo.3775959 (Ehrlich, 2020a).

This project contains the following source data files:
- data/2008_2019Games.csv. (Input data scraped using the scrapingMLB.ipynb.)
- data/divisions.csv. (Input team division data for grouping by division.)
- data/mlbTeamColors.csv. (Input team colors for the visualizations.)

Source data are also available on GitHub: https://github.com/Syracuse-University-Sport-Analytics/MLBCovid19.

Extended data
Harvard Dataverse: Replication Data for: COVID-19 Containment Measures, Major League Baseball, and the Home Field Advantage. https://doi.org/10.7910/DVN/OOMWSD (Ehrlich, 2020b).

This project contains the following extended data files:
- divisionRankings. (Results of simulation using the logit model.)
- divisionRankingsNN. (Results of simulation using the neural network model.)
- homeEffectLogit. (Team home effects using the logit model.)
- homeEffectNN. (Team home effects using the neural network model.)
- modelCorrelationsSummaryWithNNResults. (Simulation statistics from both the logit and neural network models.)

Zenodo: Syracuse-University-Sport-Analytics/MLBCovid19.
https://doi.org/10.5281/zenodo.3775959 (Ehrlich, 2020a).

This project contains the following source files:
- pythonStatcastScraper/scrapingMLB.ipynb. (Python Jupiter Notebook code for scraping Statcast.)
- halfSeasonPrediction.Rmd. (R Markdown Notebook code for developing the logit and neural network models. Also contains the code for running the simulations.)
- All the other data is intermediate output from the simulations. The important output files are located in the above Harvard Dataverse repository.

Source code is also available on GitHub: https://github.com/Syracuse-University-Sport-Analytics/MLBCovid19.
Mixed data and code hosted on GitHub and Zenodo are available under the terms of the GNU General Public License v3.0.

Data hosted on Harvard Dataverse are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

## References

Bogage J: National Baseball Hall of Fame accepts Atlantic League ‘robo ump’. items. Washington Post. 2019.
Reference Source

Edwards J, Archambault D: The home field advantage. Sports, games, and play: Social and psychological viewpoints. 1979; 459–438.
Reference Source

Ehrlich J: Syracuse-University-Sport-Analytics/MLBCovid19: First Release (Version v1.0.0). Zenodo. 2020a.
http://www.doi.org/10.5281/zenodo.377959

Ehrlich J: Replication Data for: COVID-19 Countermeasures, Major League Baseball, and the Home Field Advantage. Harvard Dataverse, V1, UNF-6: LiHptTK15GERswc31Bx9p== [fileUNF]. 2020b.
http://www.doi.org/10.7910/DVN/OOMWSD

Gómez MA, Pollard R, Luis-Pascual JC: Comparison of the home advantage in nine different professional team sports in Spain. Perceptual and motor skills. 2011; 113(1): 150–156.
PubMed Abstract | Publisher Full Text

Jones MB: The home advantage in major league baseball. Percept Mot Skills. 2015; 121(3): 791–804.
PubMed Abstract | Publisher Full Text

Jones MB: Differences in home advantage between sports. Psychol Sport Exerc. 2018; 34: 61–69.
Publisher Full Text

Kuhn M, Wing J, Weston S, et al.: caret: Classification and Regression Training (Version 6.0-86) [Computer software]. 2020.
Reference Source

LeDoux J: Jidbc/pybaseball [Python]. 2020; (Original work published 2017).
Reference Source

Passan J: Sources: MLB, players eye May return in Arizona. ESPN.Com. 2020.
Reference Source

Pollard R, Prieto J, Gómez MA: Global differences in home advantage by country, sport and sex. Int J Perform Anal Sport. 2017; 17(4): 586–599.
Publisher Full Text

Ripley B, Venables W: nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models (Version 7.3-14) [Computer software]. 2020.
Reference Source
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com