Data and Methods - Myth: Fans beeinflussen den Ausgang von Fußballspielen.

By Johan Brockschmidt and Aishwarya Ganta

Home advantage is a physiological or psychological advantage, where a sporting team tends to perform better in their home stadium than away. This home advantage is affected by various factors such as audience, biased referees, familiarity with the ground, and change in-game tactics. In this report, we statistically prove how the home advantage for various European soccer teams is affected due to corona.

What data is the evaluation based on?

The data for 'Myth: Fans beeinflussen den Ausgang von Fußballspielen' has been extracted from three different websites. The R-package `rvest` (Wickham, 2021) is used for the data scrapping. The data cleaning has been done using the R package `dpylr` (Wickham et al., 2021).

The entire analysis is performed on 21 European soccer league data which includes: England, Sweden, Spain, Slovakia, Portugal, Poland, Netherlands, Malta, Italy, Ireland, Hungary, Greece, Germany, France, Finland, Estonia, Denmark, Cyprus, Croatia, Bulgaria, and Austria. The home advantage of 21 European soccer leagues for years 2018 and 2020 is collected from `https://www.soccerstats.com` (from the respective league pages). The year 2018 data is used instead of 2019 as it was partially affected by corona. The extracted data were further processed by removing the missing values. In the analysis, we work on data related to 'club names', 'home points per game', 'away points per game, and 'home advantage'. The information about the audience percentage in the respective stadium for Premier League 2018 and 2021 has been extracted from `https://www.transfermarkt.com/`. In data processing, the data corresponding to 'stadium name' and 'attendance or spectator percentage' is extracted, and all the missing values are removed from the extracted data. The possession data was extracted manually from the `https://www.kicker.de/`. The data regarding away and average possession for years 2018 and 2020 are extracted for the German soccer league.

How the home advantage is measured?

The points per game (PPG) for home are calculated by dividing the total points made during the home games by the total number of home games. Similarly, we calculate the points per game (PPG) for away by dividing total points made at away games by the total number of away games. Finally, the home advantage is calculated by subtracting points per game for away from a point per game for home (PPG Home - PPG Away). The average home advantage for the respective league is calculated by dividing the total home advantage of all the clubs within the respective league by the total number of clubs.

Methods

The average home advantage of 21 European soccer leagues is calculated for the years 2018 and 2020. The home advantage of the respective soccer league for years 2018 and 2020 represents the home advantage of the respective league without and with the corona effect, respectively. Most of the graphs are produced using the package ggplot2 (Wickham, 2016). The effect of corona on the home advantage of various leagues is represented in a heat map. The heat maps are generated using library `rworldmap` (South, 2011) (`getMap()` function). From Figure 1, we can interpret that most leagues' home advantage decreased during corona. The countries Bulgaria, Finland, Portugal, England, Germany, France, Spain, Estonia, Greece, Ireland, Italy, Netherlands, Poland, Sweden, Slovakia, Hungary, and Croatia have higher home advantage values (darker color) in the year 2018 when compared to the year 2020. It means that the home advantage of these leagues is highly affected due to corona when compared to other leagues. The bar plot in Figure 2 is a detailed representation of the home advantage value of 21 leagues for the years 2018 and 2020. The bar plot is generated using library `ggplot2`. In contrast to many leagues, Austria, Denmark, and Cyprus have a higher home advantage during corona. Further from Figure 2, we can interpret that all 'the big 5' leagues show negative home advantage during corona, and home advantage of England and France leagues are highly affected by corona.

The difference in home advantage for the big 5 and other leagues is plotted for years 2018 and 2020 in Figure 3. This shows that not only the big 5 leagues but most of the other (small) leagues are affected by corona. To statistically confirm that the home advantage of most of the soccer leagues is affected by corona, we are performing one-sided

multiple t-tests for home advantage (league wise) for years 2020 and 2018 with unequal variances at a 0.05 level of significance. The null hypothesis states that the mean home advantage for both the years is equal i.e., $H_0 : \mu_1 = \mu_2$ where μ_1 and μ_2 represent mean home advantage of respective leagues for years 2018 and 2020, respectively. Alternative hypothesis states that home advantage of 2020 is less than 2018 i.e., $H_0 : \mu_1 < \mu_2$. The estimated p-values of the multiple t-tests can be seen in Figure 4. The p-value is a probability metric that is used to determine whether or not an observed test statistic is due to chance. If the p-value is low, the effect is said to be statistically significant as it implies that it did not happen by coincidence. From Figure 4, we can interpret that the p-values of the leagues of England, Portugal, Netherlands, Italy, Greece, and France are less than the significance value of 0.05. This states that the null hypothesis can be rejected for the respective leagues which means that the mean home advantage value of the respective leagues for the year 2020 is significantly less than the year 2018 at a significance level of 0.05.

From the above multiple t-test results, we can see that Premier League (England) is one of the most affected leagues due to corona. The Figure 6, we can interpret that for most of the premier league teams the home advantage has reduced from the years 2018 to 2020. From the figure, we can also interpret that the mean home advantage of the Premier League for the year 2018 (green line) is much higher than 2020 (violet line). In contrast to Premier League, from Figure 5 we can see that the home advantage of Germany didn't change significantly as mean home advantage values for years 2018 and 2020 are very close.

The one main reason for this would be the strictness of corona measures which resulted in less audience in the stadium in the year 2020 when compared to 2018. The scatter plot Figure 7 shows how the home advantage of clubs in the Premier League is affected by audience percentage for the year 2018. In Figure 7, we can see that the audience effect varies from team to team. For example, the home advantage value of Arsenal, Everton, Chelsea, and Brighton & Hove Albion which have a high audience percentage is also high.

Apart from audience percentage, another parameter that might have been affected is possession. Figure 8 shows a difference in possession (away-average) for German teams for years 2018 and 2020. From the graph we can say that possession away did not change during corona and teams still have less possession away.

Conclusion

In the course of analysis, we have found that the home advantage for most European soccer clubs has decreased due to corona that concludes that audience percentage affects home advantage. The home advantage of the big leagues with more fans is influenced more as the home advantage value decreased significantly. The leagues where mental factors might play a vital part, and fans follow the game directly lost more home advantage. The game tactics (possession) did not change during corona. As the Germany home advantage was not affected more, we might conclude that traditional and long-existing teams with loyal fans suffered the most.

Notes for lecturers:

- The code that is used to generate the graphs can be found in code folder which comprises of seven R files. The R files are to be executed in an order which can be found me README file as the code is dependent on each other.
- The R files do not require any data file for execution. In case, the code doesn't work due to some changes in the web link please use data folder to load the required data. The respective information about the data files can be found in the respective README file.
- The required graphs can be found in graphs directory with name: required_graphs.pdf.
- The extra graphs that have been generated can be found in graphs.pdf file. If required, any relevant graphs can be included in the article.

Bibliography

South, A. (2011). rworldmap: A new r package for mapping global data. *The R Journal*, 3(1):35–43.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2021). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 1.0.0.

Wickham, H., François, R., Henry, L., and Müller, K. (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.5.