Analysis of the Effect of Home Ground on Wins in EPL

I, Shourya Sankha Mishra, affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

SHOURYA SANKHA MISHRA

Roll no: 472

Supervisor: Prof. Ayan Chandra

Table of Contents

1. The Problem **2**

Probable reasons for home-advantage 3

2. Methodology **5**

2.1 Descriptions of the tests used 5

2.1.1 Shapiro-Wilk test 5

2.1.2 Median test 6

2.1.3 Paired t-test 9

2.2 Procedure 10

2.3 Analyses 10

3. Results and Discussion **12**

Tests 12

Plots 15

Conclusion **17**

References & Bibliography **19**

Acknowledgements **20**

With its gradual development and increasing demand over the years, statistics has found its way into various application-based fields in sports. For example, statistical modelling has totally changed the face of player scouting (the process of evaluating the talent of players with an objective of signing them on a professional contract for their respective teams). A common notion exists in sports which states that in any match contested by two opponents, be it individual or team performance, the party contesting in their home enjoys significant advantages in comparison to the visitors. So much so, that tournaments have been designed to provide the benefit of doubt to the team performing better in away games in case of ties. Over the past 20 years, the home advantage has emerged as an important statistical research paradigm in organised sport. The home advantage is a robust phenomenon that has been consistently demonstrated in both individual and team sport competitions. We seek to test that whether such a concept holds valid in football, arguably the most popular form of sport currently played all over the world.

1. The problem

In majority of the football tournaments, be it at club level or international level, the tournaments are designed in such a way that every team has equal number of home and away fixtures. This is ensured by the thumb rule that each team plays each of the other teams twice, once in their own ground and again in the opponents' ground. The team hosting the match i.e. playing in their own ground is said to be the 'home team', playing in their 'home ground', whereas the other team are called visitors or more commonly the 'away team' playing on an 'away ground'. The common belief which exists is that the home team enjoys certain advantages over the away team in a home fixture, and hence is more likely to win the match. This, in a nutshell, is what is known as home advantage. A study by Courneya and Carron, (1992) indicated that the number of home game wins usually exceeds the number of away game wins over a balanced home and away competition. Carron et al. (2005) define home advantage as, "the consistent finding that home teams in sports competitions win over 50% of the games played under a balanced home and away schedule." The notion of home advantage hypothesises that the game location factors feed into the psychological states of the players, coaches and officials so as to impact the behaviour of these individuals resulting in advantage for the home team. Now we speculate as to what might lead to the existence of such a notion. From a large number of existing researches, the probable reasons can be listed as follows---

1)*Influence of the crowd*- The presence of supporting fans in large numbers cheering for the home team at the top of their voice provides a psychological boost in the morale of the home team players. Also, players are less likely to shirk in front of their home fans because they do not want to lose their approval. Shirking is when a player purposefully does not perform to the best of his/her abilities. Fans express their approval or disapproval by not attending games, cheering or booing at games, or buying the player's jersey. Attendance and merchandise sales are a large part of a player's salary, so a player is going to make sure he performs especially well at home to keep the fans happy and his salary high.

2)*Travel effects*- Observe that the home team is playing on their own ground and hence is well suited to their conditions whereas the away team has to travel to a different ground with possibly different geographical conditions such as temperature, altitude etc.

Other than this, the journey which the away team has to make to reach the venue leads to travel fatigue and has a negative effect on the away team's performance.

3)*Referee bias*- An interesting concept regarding home advantage is referee bias. It can be explained through the psychological theory of human nature that people want to be liked and to be confirmed in their judgements. If the home crowd is loud and clamorous, it is likely that 50-50 decisions will go to the home side (resulting in a bias towards the home team), thus resulting in cheers in favour of the referee from the home crowd.

4)*Location familiarity*- As mentioned earlier, each team plays half of the matches in a tournament in their home ground. It is expected that they are well versed with the details of the playing ground. Also, the home team has the independence of manipulating the grounds within small margins so as to suit their tactical needs.

The home advantage is generally thought to occur due to the aforementioned four causes. Based on the above theoretical background, we wish to analyse the effect of home advantage in producing wins in football matches in this study. In other words, we wish to check the significance of the factor 'home advantage' on the result of the match. However, other potential factors may also affect the extent of home advantage.

The first potential factor affecting home advantage is athlete and team ability. One may hypothesise that teams with low ability might exhibit greater home advantage than teams with higher ability. Since teams with low ability win seldomly i.e. less frequently, the home advantage becomes a crucial factor in their ability to win games. Thus, in this study, we also seek to explore whether the extent of home advantage differs in teams of varying ability.

Secondly, there is evidence from both European and English football competitions that the home advantage has been in decline since the 60's. This has generally been attributed to increased professionalism and the development of a market culture in the game. This study seeks to explore further, using recent data, whether the home advantage has continued to decline during the modern era.

2. Methodology:

2.1 Descriptions of tests used in the study.

- ### 2.1.1 *Shapiro-Wilk Test for normality*

Let $X_1, X_2, \dots, X_n$ be a random sample of size n drawn from the distribution of the random variable X. We wish to test H_o:Sample is drawn from a normal population against H_1: not H_o.

We arrange the data in ascending order and obtain the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Let $Z_1, Z_2, \dots, Z_n$ be a random sample of size n from $N(0,1)$ distribution and define their order statistics as $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$.

Let $E\left(Z_{(i)}\right) = m_i \; \forall \; i = 1(1)n$ for a given n, as clearly the expectations will depend on n. Let us define $\left(E\left(Z_{(1)}\right), E\left(Z_{(2)}\right), \dots, E\left(Z_{(n)}\right)\right) = (m_1, m_2, \dots, m_n)$. The idea behind the Shapiro-Wilk test is to consider the correlation of $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ with $(m_1, m_2, \dots, m_n)$. In other words, we check how well the order statistics $X_{(i)}$ are correlated with expected standard normal order statistics. A correlation close to 1 would suggest a good fit to normality, whereas a correlation much less than 1 would suggest non-normality. The Shapiro-Wilk statistic uses the means and the covariances of the standard normal order statistics $Z_{(i)}$ (as $Z_{(i)}$ is strongly correlated with the adjacent $Z_{(i-1)}$ and $Z_{(i+1)}$ due to the ordering $Z_{(i-1)} \leq Z_{(i)} \leq Z_{(i+1)}$) to give a test statistic for normality. Let the covariance matrix of $Z_{(i)}$ be given by $V^{n\times n}$, $V_{ij} = E\left[\left(Z_{(i)} - m_i\right)\left(Z_{(j)} - m_j\right)\right]$ $i = 1(1)n, j = 1(1)n$. We define $m = (m_1, m_2, \dots, m_n)'$, where m is an $n \times 1$ column vector. Consider the vector $m'V^{-1}$, let the length of this vector be given by $C = \parallel m'V^{-1} \parallel = (m'V^{-1}V^{-1}m)^{\frac{1}{2}}$. Now we define $a' = (a_1, a_2, \dots, a_n) = \frac{m'V^{-1}}{C}$, which is a unit row vector consisting of the ith expected value of normalised order statistics. Then the Shapiro-Wilk test statistic is given by

$$W = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

For i.i.d observations, the values of a_i can be obtained from the table presented by Shapiro and Wilk (1965) for sample sizes up to 50. W can be expressed as a square of the correlation coefficient between a_i and $X_{(i)}$.So, $0 < W \leq 1$. For our test purposes, we shall obtain the p-values of the tests.

Rejection rule: Reject H_o in favour of H_1 at α level of significance iff, $p < \alpha$ and conclude that the sample is not drawn from a normal distribution.

• 2.1.2 *Median test*

Let X and Y be two independent random variables having absolutely continuous distributions given by the distribution functions F_X and F_Y respectively. Here we assume that the two distributions may differ only with respect to their location i.e. we may have

$F_Y(x) = F_X(x - \theta)\ \forall\, x$ and some $\theta \neq 0$.Under this assumption, the test $H_o: F_Y(x) = F_X(x)\ \forall\, x$ against H_1: not H_o is equivalent to the test $H_o: \theta = 0$ against H_1: not H_o.

Let ($X_1, X_2, \ldots, X_{n_1}$) and $(Y_1, Y_2, \ldots, Y_{n_2})$ be two random samples drawn independently from the distributions of X and Y respectively. We combine the observations from the two samples and arrange them in ascending order of magnitude. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, $(n = n_1 + n_2)$ denote the ordered arrangement of these n observations in the combined sample. Let X_{me} denote the median of the combined sample. If n is odd, the median is the $\frac{n+1}{2}$ th observation in the ordered arrangement. If n is even, the median is any value between the $\frac{n}{2}$ th and $(\frac{n}{2} + 1)$th observation in the ordered arrangement. Let V denote the number of first sample observations which are $< X_{me}$. V is our test statistic. Under H_o, the n observations may be regarded as a single random sample drawn from a single continuous population. Hence the event that first sample observation $< X_{me}$ and the event that second sample observation $< X_{me}$ are equally likely under H_o. We are interested to test the alternative hypothesis $H_1: F_Y(x) = F_X(x - \theta)\ , \theta < 0$.

Under $H_1: F_Y(x) = F_X(x - \theta)\ , \theta < 0, X$ is stochastically larger than Y. Hence, first sample observations will tend to be larger, consequently the number of first sample observations $< X_{me}$ will tend to be smaller. Thus, V will tend to be small, and a too small value of V will indicate departure from H_o. We will reject H_o against H_1 at α level of significance iff $V_{obs} \leq v_\alpha$ where v_α is $\ni P_{H_o}[V \leq v_\alpha] \leq \alpha$.

If n is even, the median is any value between the $\frac{n}{2}$ th and $(\frac{n}{2} + 1)$th observation in the ordered arrangement. Hence the number of observations in the combined sample which are $< X_{me}$ will be $\frac{n}{2}$. If n is odd, the $\frac{n+1}{2}$th observation is the median. Hence the number of observations in the combined sample which are $< X_{me}$ will be $\frac{n-1}{2}$ $(= \frac{n+1}{2} - 1)$. Hence in general there are exactly $\left[\frac{n}{2}\right]$=t (say) observations in the combined sample which are $< X_{me}$, where $\left[\frac{n}{2}\right]$ is the largest integer not exceeding $\frac{n}{2}$. Under H_o, the $n = n_1 + n_2$ observations may be considered to be a simple random sample drawn from a single continuous distribution. Thus, under H_o, total number of equally likely ways in which these t observations can be chosen out of n values is $\binom{n}{t}$. To find out the total number of favourable cases, we note that V takes value v if exactly v of the n_1 values of X_i and hence $t - v$ of the n_2 values of Y_j are less than the combined sample median X_{me}, and hence, total number of favourable cases will be $\binom{n_1}{v}\binom{n_2}{t-v}$, v=0,1, … , min(n$_1$,t). Hence under H_o, V~Hyp(n, n_1, t). Thus

$$E_{H_o}(V)= \frac{n_1}{n}t$$

$$V_{H_o}(V)= \frac{n-t}{n-1} \cdot t \cdot \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right) = \frac{tn_In_2(n-t)}{n^2(n-1)}$$

Let us define $Z = \frac{V - \frac{n_1 t}{n}}{\sqrt{\frac{tn_1n_2(n-t)}{n^2(n-1)}}}$. For large values of n, under H_o, $Z \sim N(0,1)$ asymptotically.

To test $H_o\colon \theta = 0$ against $H_1\colon \theta < 0$.

Rejection rule: We reject H_o in favour of H_1 at α level of significance iff $Z_{obs} < -\tau_\alpha$, where τ_α is the upper α point of a standard normal distribution.

- 2.1.3 *Paired t-test*:

Let X and Y be two sample observations which are paired together. The pair of observations (x_i, y_i), $i = 1(1)n$ corresponds to the same ith sample unit. Our aim is to test whether the sample means differ significantly or not. Here we consider the differences $d_i = x_i - y_i$, $i = 1(1)n$. Now, we know that, if $Y_i \sim N(\mu_i, \sigma_i^2)\ \forall i$ then $\sum_i l_i Y_i \sim N(\sum_i l_i \mu_i, \sum_i l_i \sigma_i^2)$. In other words, a linear combination of normal random variables also follows a normal distribution. It can be easily observed that d_i's follow a normal distribution. Let $d_i \sim N(\mu, \sigma^2)$ identically and independently. The mean of the differences of the sample is given by $\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$.

We wish to test the null hypothesis $H_o: \overline{d} = 0$ against $H_1: \overline{d} > 0$. The estimate of the variance is given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$. We observe, $\overline{d} \sim N(\mu, \frac{\sigma^2}{n})$. The null hypothesis suggests that increments are due to fluctuations of sampling only. Under H_o, $\mu = \mu_o = 0$. σ^2 being unknown, we estimate σ^2 by $\widehat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2$. We observe that, $\frac{\bar{d}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)s^2}{\sigma^2} \sim {\chi^2}_{n-1}$.

An appropriate test statistic for the above test is then given by $t = \frac{\frac{\bar{d}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} = \frac{\bar{d}-\mu}{s/\sqrt{n}}$.

Under H_o, $t = \frac{\bar{d}-\mu_o}{s/\sqrt{n}} = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$. ($\because$ numerator is a standard normal variate and denominator is a χ^2 variate divided by its degrees of freedom).

Rejection rule: We reject H_o in favour of H_1 iff $t_{obs} > t_{\alpha,n-1}$ where $t_{\alpha,n-1}$ is the upper α point of the distribution with $n - 1$ degrees of freedom.

2.2 Procedure

Data has been collected on last 10 seasons of the English Premier League from 2008-09 to 2017-18. A total of 36 teams have played in 3800 matches over these 10 seasons under a balanced home-away schedule. Each season has 20 teams contesting throughout the year. During this period every team plays each of the other 19 teams twice, both at their home and away from home. Thus, a season consists of $\binom{20}{2} \times 2 = 380$ matches being played. A team gets 3 points for winning, 1 point for a draw and doesn't get any point in the event of a loss. The data collected consists of paired data in the form of $(home, away)$ for the teams participating in the respective seasons, where $home$ denotes the number of points scored by the team in their home ground and $away$ denotes the number of points scored by the teams away from home. The sum of these two variables thus give the total points scored by the respective teams in given season. This data serves as the samples of analysis. For the 10 seasons, 10 sets of paired data on home and away points have been obtained.

2.3 Analyses

We wish to test whether the home points are significantly higher than the away points. We begin by testing every sample for normality. If the normality assumption holds, we can say that the sample is derived from a bivariate normal distribution and proceed to perform paired t-test for the mean of the differences. If home advantage does truly exist, the home points will be much higher than the away points and if not, then the pair of data should be evenly distributed. Hence, we test for the hypothesised mean of differences equal to zero against the alternative hypothesis that the mean of differences is greater than zero. If the normality assumption does not hold true, we proceed to perform the non-parametric median test among the pairs of the samples to test for the difference of their locations.

As for a measure of the team ability, the finishing positions of the teams in the league table will be considered. For the measure of the extent of home advantage, we calculate the percentage of $\frac{hom}{home+away}$ for the teams. Let us define these percentages as Home-Advantage index. Let us define a matrix $X^{n\times p} = ((x_{ij}))$, where x_{ij} denotes the index for the jth ranked team in the ith season, $i = 1(1)n, j = 1(1)p$. Here we have under study 10 seasons with 20 teams in each season. Therefore, $n = 10, p = 20$.

Clearly, $\overline{x_{\iota o}} = \sum_j x_{ij} \quad \forall i$ gives the average of the indices for the ith season which can be considered as the overall home-advantage index for the ith season.

Again, $\overline{x_{o\jmath}} = \sum_i x_{ij} \quad \forall j$ gives the average of the indices for the jth position over the 10 seasons, which can be considered as the average home-advantage index for the teams in jth position over all seasons.

We shall plot the $\overline{x_{\iota o}}'$s with the subsequent seasons to observe how the effect of home advantage behaves over time. To find out how home advantage affects teams of different abilities, we shall plot the $\overline{x_{o\jmath}}'$s against finishing positions (1 to 20).

3. Results & Discussion

We begin by performing Shapiro-Wilk test on the sample data to check the normality assumptions. The results obtained are summarised in table 1.

Table 1: Results obtained for Shapiro-Wilk test of the samples

Season	Samples	p-value of Shapiro-Wilk test	Decision	
2008-09	08-09 home	0.9409	Accept	______
	08-09 away	0.0087	Reject	
2009-10	09-10 home	0.8516	Accept	Bivariate normal sample
	09-10 away	0.7633	Accept	
2010-11	10-11 home	0.3584	Accept	Bivariate normal sample
	10-11 away	0.2979	Accept	
2011-12	11-12 home	0.9277	Accept	Bivariate normal sample
	11-12 away	0.3603	Accept	
2012-13	12-13 home	0.8697	Accept	Bivariate normal sample
	12-13 away	0.0587	Accept	
2013-14	13-14 home	0.0302	Reject	______
	13-14 away	0.0204	Reject	
2014-15	14-15 home	0.1589	Accept	Bivariate normal sample
	14-15 away	0.6972	Accept	
2015-16	15-16 home	0.3788	Accept	Bivariate normal sample
	15-16 away	0.8812	Accept	
2016-17	16-17 home	0.7184	Accept	Bivariate normal sample
	16-17 away	0.0793	Accept	
2017-18	17-18 home	0.0269	Reject	______
	17-18 away	0.0011	Reject	

We observe that the samples corresponding to the seasons 2008-09, 2013-14 and 2017-18 are not normally distributed. For each of these seasons, we perform the non-parametric median test among the variables $home$ and $away$. The results are provided in table 2.

Table 2: Results obtained for the median tests

Season	Home	Away	V_{obs}	Z_{obs}	Critical value	Decision
2008-09	30.8 ± 8.56	21.35 ± 11.83	6	-2.49	-1.64	Reject
2013-14	30.75 ± 10.96	22.35 ± 9.31	8	-1.24	-1.64	Accept
2017-18	30.9 ± 10.1	21.15 ± 10.4	7	-1.87	-1.64	Reject

We find that the test for 2013-14 is accepted whereas the tests for 2008-09 and 2017-18 are rejected.

For the other 7 samples 2009-10, 2010-11, 2011-12, 2012-13, 2014-15, 2015-16, 2016-17 which are bivariate normal samples, paired t-test for difference of means is carried out. The results are summarised in table 3.

Table 3: Results obtained for the Paired t-tests

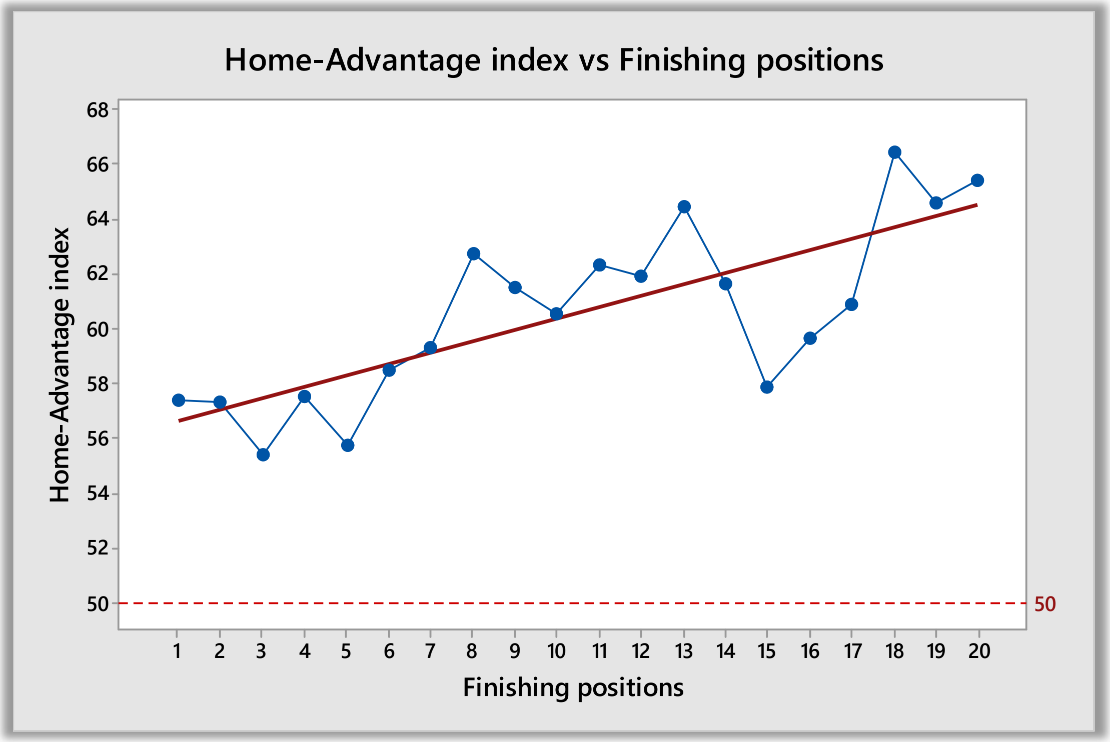
Samples	Home	Away	t_{obs}	Critical value	Decision
2009-10	33.3 ± 10.76	18.0 ± 10.17	10	1.72	Reject
2010-11	32.4 ± 8.74	19.05 ± 5.74	8.02	1.72	Reject
2011-12	30.3 ± 10.96	22.05 ± 7.73	4.95	1.72	Reject
2012-13	30.3 ± 9.44	21.3 ± 9.48	6.31	1.72	Reject
2014-15	30.45 ± 9.14	21.9 ± 8.14	6.7	1.72	Reject
2015-16	28.9 ± 8.28	22.75 ± 8.25	4.67	1.72	Reject
2016-17	32.25 ± 10.55	20.55 ± 11.11	6.03	1.72	Reject

Observe that all of the tests are rejected.

In the above tables, the columns Home and Away give the $mean \pm sd$ of the corresponding variables for the respective seasons. We observe that the sd's remain more or less same in all the samples irrespective of venue or season. But the $home$ means are visibly larger than the $away$ means for all of the ten seasons. In overall, of the 10 seasons included under this study, 9 of them show significant difference between the points scored by teams in their home and away fixtures.

The plot of $\overline{x_{oJ}}'$s against finishing positions is obtained in figure 1.

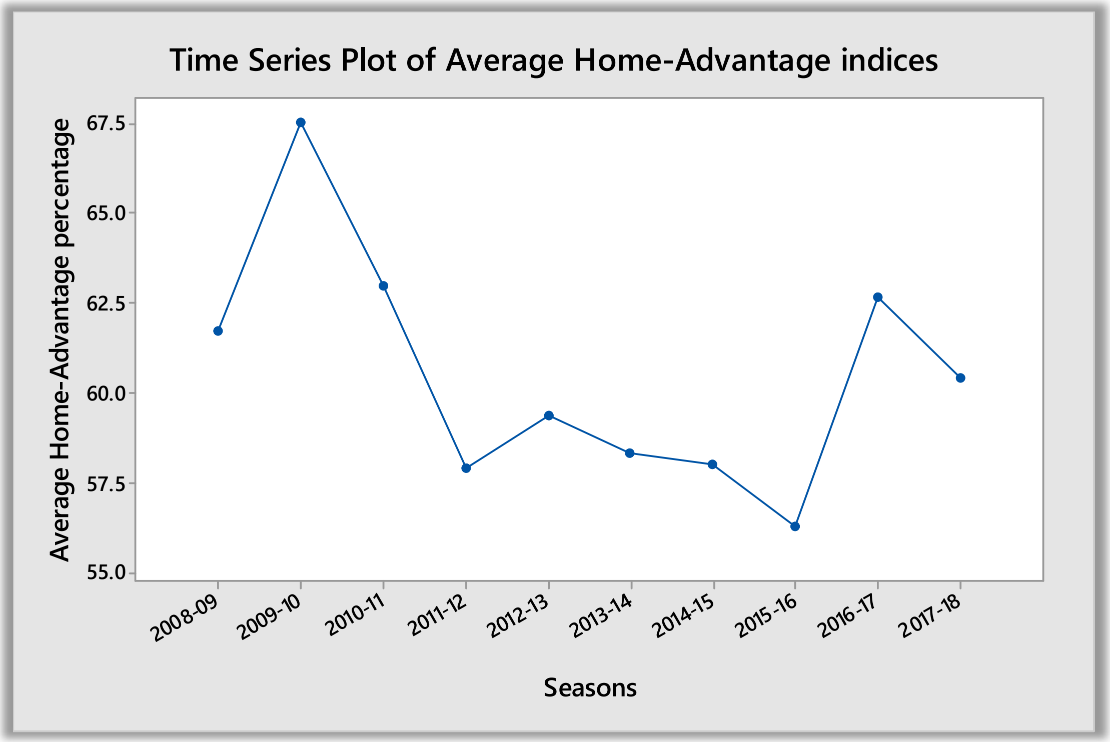
Figure 1: Plot of home-advantage indices for each of the finishing positions



We observe that an increase in the extent of home advantage with the decrease of team ability i.e. with lower finishing positions is demonstrated by figure 1.

The time series plot of the seasonal average indices over seasons is given in figure 2.

Figure 2: Plot of the average home-advantage indices for the first 10 seasons of the EPL



The home advantage indices appear to be stationary, thus indicating that the extent of home advantage has remained unchanged over time. It is also to be noted that the indices for each of the seasons are greater than 50% without a single exception. This fact inclines our decision in favour of presence of home-advantage.

4. Conclusion

The aim of this study was to analyse the effect of home advantage in producing wins in football matches played in the English Premier League based on the hypothesis that the game location factors feed into the psychological states of the players, coaches and officials so as to impact the behaviour of these individuals resulting in advantage for the home team. Home-Advantage may arise due to multiple factors mentioned previously or as a result of the interaction of these factors. The results indicated that venue had a significant effect in determining the result of the match. For each of the 10 seasons of the league, home advantage indices were obtained as 61% (2008-09), 67% (2009-10), 62% (2010-11), 57% (2011-12), 59% (2012-13), 58% (2013-14), 57% (2014-15), 56% (2015-16), 62% (2016-17) and 60% (2017-18), (shown in figure 2). An overall advantage of 60% is observed over all the seasons from 2008-09 to 2017-18. It is observed that out of the ten tests for difference of location conducted, nine of them have been rejected, implying that the points gathered by teams in their home games is significantly larger than that in away games. Thus, we arrive at the conclusion that teams are more successful in home fixtures than in away fixtures.

As for the potential factors affecting home advantage, from figure 1 we conclude that teams of lower ability show home advantage to a greater extent than the teams of higher ability.

Finally, from figure 2 we may conclude that the home advantage in successive seasons do not exhibit any decreasing time trend and thus has remained relatively consistent in recent years.

Further research work can include other leagues from different parts of the world aiming to establish home advantage as a global phenomenon. Attempts can be made to quantify the hypothesised home factor through suitable predicting factors which may be thought to influence home advantage by fitting appropriate regression models.

References

Allen, M. S., & Jones, M. V. (2014). The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends. *International Journal of Sport and Exercise Psychology, 12:1*, 10-18.

Courneya, K., & Carron, A. (1992). The home-field advantagein sport competitions:A literature review. *Journal of Sport and Exercise Psychology, 14*, 28-39.

Koteki, J. (2014). Estimating the Effect of Home Court Advantage on wins in the NBA. *The Park Place Economist, 22*, Article-13.

Vaz, L., Carreras, D., & Kraak, W. (2012). Analysis of the effect of alternating home and away field advantage during the Six Nations Rugby Championship. *International journal of Performance Analysis in Sport, 12*, 594-608.

Bibliography

Gibbons, J. D., & Chakraborti, S. (2010). *Nonparametric Statististical Inference.* CRC Press.

Gun, A. M., Gupta, M. K., & Dasgupta, B. (2014). *Fundamentals of Statistics,Volume 1.* Kolkata: The World Press Pvt. Ltd.

All necessary data for this analysis has been collected from http://www.footstats.co.uk/

Acknowledgements

I would like to extend my heartfelt thanks to our respected professors of the Department of Statistics, St. Xavier's College Kolkata for providing me the privilege of doing this project work. I would especially like to express my sincerest gratitude to my supervisor, Prof. Ayan Chandra, who has guided me, provided assistance and invaluable suggestions without which this project would not have been possible. I am also immensely grateful to my parents and my classmates for their tireless efforts and help throughout the progress of the project.