GDE ASE Project 2023

Daniel Harris

Can height predict playing position in football?

Abstract

Height is a key characteristic in the physical composition of the association football player. The aims of this study were to explore the relationship between height and playing position, and to examine whether height could be a predictor for playing position. Player heights of English Premier League players were taken from an online database along with their corresponding positions and an initial analysis was performed. These data were then used to build a predictive model using probability mass functions. The model was tested against a reference model, outperforming it by more than a factor of two.

Programs used: Excel, Python

Word Count: 2507

1. Introduction

In football, a player's physical attributes play an important role in determining his or her suitability for different positions on the pitch. One such physical characteristic is height. Height offers advantages in certain aspects of the game — for example, taller outfield players may have a greater ability to win aerial duels, which can provide a competitive edge, especially in set-piece situations such as corners and free-kicks. Height may also be associated with longer strides and a greater reach, which could be advantageous in tackling and intercepting passes. For goalkeepers, being tall may help in covering more of the goal, and dominating the penalty area. On the other hand, taller players are sometimes perceived to have reduced agility and manoeuvrability, and therefore to be less suited to positions which require lots of fast changes in direction and the ability to navigate through tight spaces on the pitch. Their larger frames and higher centre of gravity may negatively impact their endurance and balance respectively. Whilst this study does not aim to draw conclusions with regard to the aptitude of players of different heights to play in different positions, the aim is, rather, to investigate the relationship between height and position, and to answer the research question: *Can height predict playing position in football?*

This project uses cross-sectional data on all players currently registered in the English Premier League (EPL), as per the online football database *Transfermarkt (2023)*. It consists of two main parts. Firstly, I present an analysis of the relationship between height and playing position from the EPL data. Secondly, I present a model using probability mass functions which attempts to predict which position a player is likely to end up playing in, given a certain height. The structure of the project is as follows: Section 2 provides background information on football, the importance of player positions, and relevant literature on the role of height in predicting these positions. Section 3 describes the dataset used for the analyses. Section 4 describes the analysis methods and the results in detail. Section 5 gives a summary and concluding thoughts, with suggestions for future research.

2. Background

2.1. Playing Positions

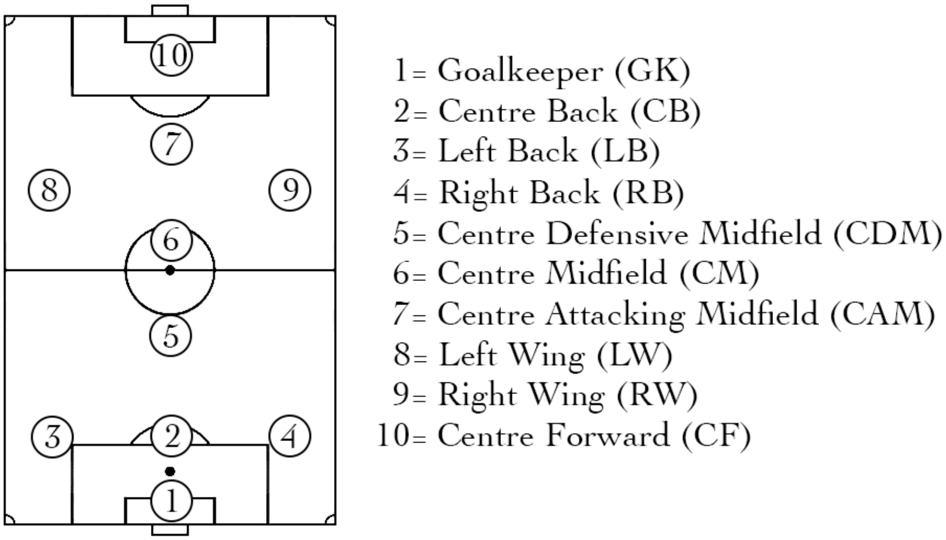


Figure 1: Overview of football positions (attacking towards the goal at the top of the image). A team will typically select a lineup of 11 through some combination of these positions.

To provide context, Figure 1 illustrates the different playing positions that are used throughout this positional analysis. Teams will typically select some combination of the 10 illustrated positions to assemble their 11-player lineup on the pitch (for example, teams will often play two centre backs, two centre midfielders, etc.). In terms of outfield players, in this study the mention of ‘central’ players refers to positions 2, 5, 6, 7, and 10. ‘Wide’ players occupy positions 3, 4, 8, and 9.

2.2. Literature Review

There is a body of literature exploring the relationship between footballers’ physical characteristics and the positions in which they play. These studies tend to explore other physiological and anatomical metrics such as speed, body composition, and weight, all of which are beyond the scope of this study, but they lay the groundwork by introducing evidence of a relationship between height and playing position. This section provides a brief synthesis of these studies and critically analyses their findings.

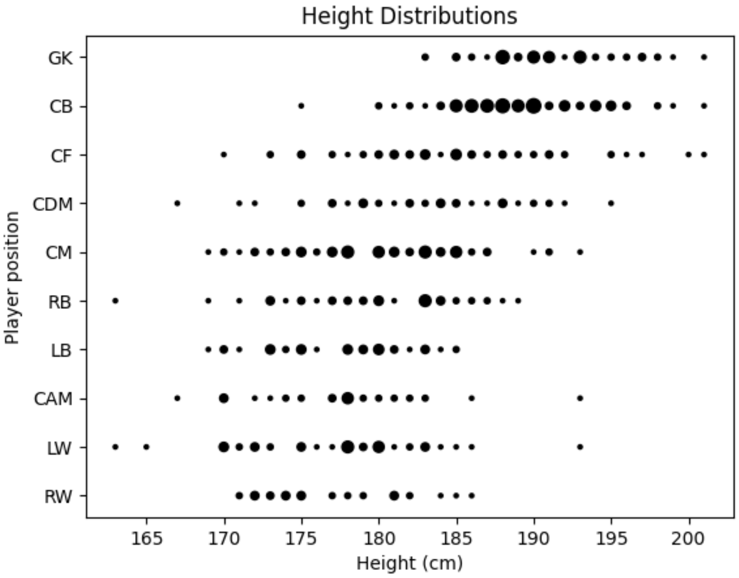
Goalkeepers are often considered to benefit from greater height, as it provides them with an extended reach to dive for saves and intercept crosses. Indeed, several studies have consistently found that goalkeepers are taller than outfield players (Bell & Rhodes, 1975; Matkovic *et al*., 2003; Hazir, 2010; Sutton *et al*., 2009). Bell and Rhodes (1975, p.200) found midfield players to be the shortest (173 cm on average), with strikers and defenders both around 177 cm, although it was concluded that since there were no statistically significant differences it would be “reasonable for playing groups to be considered as being fairly homogeneous”. Sutton *et al*. (2009) also found no significant differences between defenders and forwards, but found that defenders were, however, significantly taller than midfielders.

A major limitation of these studies is their broad categorisation of players as either goalkeepers, defenders, midfielders, or forwards. As Leao *et al*. (2019) also acknowledge, the division of the pitch into four horizontal lines, as is most commonly found in the literature, is an oversimplification which may overlook differences in height among more specific positions within these categories. By assigning players to groups based only upon their roles along the *length* of the pitch, it becomes impossible to differentiate between wide players and central players — a distinction which is likely to be important. With this in mind, the present study aims to offer a more specific and targeted investigation in this regard.

3. Data collection and preparation

Cross-sectional data for each registered EPL player as of March 2023 ($n =$ 533, where n denotes the number of observations) were collected from *Transfermarkt* and subsequently hand-coded into an Excel spreadsheet. The values corresponding to each player were name, height (in cm); main position; and, if applicable, other (second and third) positions — although in the final analysis, only main position was considered.

It should be noted that Transfermarkt makes distinctions between some positions — for example, between wide midfielders and wingers — that are perhaps useful when examining football tactics in a highly nuanced manner, but are negligible in the context of this analysis. Many football followers would argue that wide midfielders and wingers are synonymous with each other. Furthermore, only a very small number of players are listed in these ‘unusual’ positions ($n =$ 4 for left midfielders; $n =$ 2 for right midfielders; and $n =$ 4 for second strikers). To ensure adequate statistics for each position ($n \geq$ 30) I decided to reclassify these players into new positions, based on their ‘other’ positions listed on Transfermarkt. Consolidating these positions in this way reduces the complexity of the analysis and ensures a more consistent and coherent classification of playing positions.



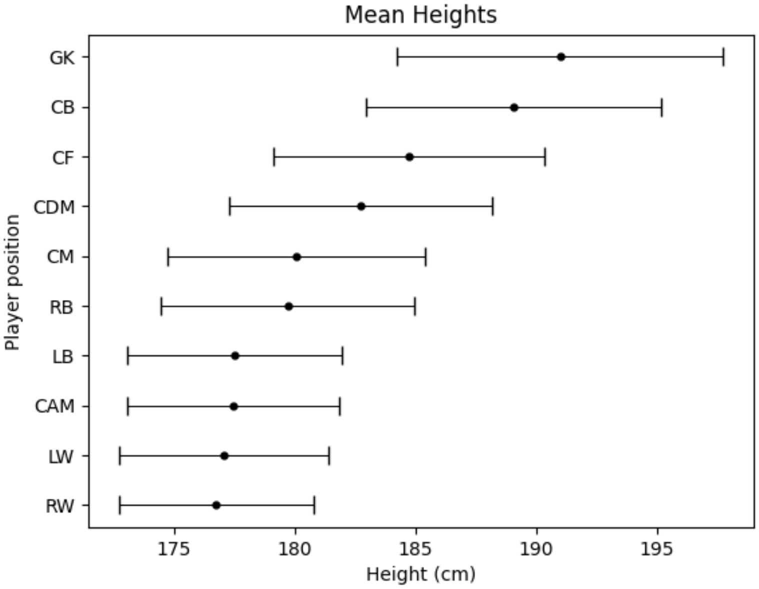


Figure 2: Height distributions by position. Playing positions are ordered by mean height. The size of each 'bubble' is directly proportional to the number of observations at that point.

Figure 3: Mean heights (with standard deviations) by position.

The height measurements provided in the database are given as integer values (in cm), resulting in a discrete distribution and the clustering of data points at some of the more common heights for each position, as shown in Figure 2. Figure 3 presents the mean heights and standard deviations by player position.

4. Analysis

4.1. Positional analysis

4.1.1 Initial assessment

Players were arranged by position, with calculations made for mean height, standard deviation and standard error. They were then ranked in descending order by mean height. The results are shown in Table 1. 95% confidence intervals were computed for the difference in mean height ($h_i - h_j$) for each two adjacently ranked positions. A t-test at a significance level $\alpha = 0.05$ was constructed for the null hypothesis H_0 (no difference in height between positions i and j) and an alternative hypothesis H_1 (mean player height for position i is greater than mean player height for position j). The formula for the t-test used is given below:

$$\tau = \frac{h_i - h_j}{\sigma_M^{ij}},$$

where σ_M^{ij} is the standard error on the difference in mean height,

$$\sigma_M^{ij} = \sqrt{\sigma_M^2(h_i) + \sigma_M^2(h_j)}.$$

Since the number of observations n for each position was greater than 30, the differences in mean height $h_i - h_j$ were assumed to be normally distributed. Results are presented in Table 2.

Player position	No. of entries	Mean height (cm)	Standard deviation (cm)	Standard error (cm)
GK	58	190.98	4.01	0.53
CB	97	189.06	4.44	0.45
CF	56	184.73	6.76	0.90
CDM	41	182.73	6.12	0.96
CM	75	180.05	5.32	0.61
RB	47	179.70	5.43	0.79
LB	44	177.52	4.37	0.66
CAM	33	177.45	5.24	0.91
LW	50	177.06	5.62	0.80
RW	32	176.75	4.36	0.77
Total:	**533**			

Table 1: Descriptive statistics by playing position.

Player position	Mean height (cm)	t-value	CI left	CI right	$p(H_0)$	$p(H_1)$
GK	190.98	2.77	0.56	3.28	0.006	0.997
CB	189.06	4.29	2.35	6.31	1.82E-05	1.000
CF	184.73	1.52	-0.58	4.58	0.128	0.936
CDM	182.73	2.36	0.45	4.90	0.018	0.991
CM	180.05	0.35	-1.61	2.31	0.726	0.637
RB	179.70	2.12	0.16	4.20	0.034	0.983
LB	177.52	0.06	-2.14	2.28	0.952	0.524
CAM	177.45	0.33	-1.98	2.77	0.745	0.628
LW	177.06	0.28	-1.86	2.48	0.779	0.610
RW	176.75					

Table 2: t-values, confidence intervals, and p-values for the difference in mean height between each position and the position in the row below.

Goalkeepers were, as expected, the tallest on average, with a mean height of 190.98 ± 4.01 cm. Goalkeepers differed significantly in height from the next-tallest group, centre-backs ($t = 2.77$). The four tallest outfield positions on average were all down the 'spine' or centre of the pitch — centre backs, centre forwards, centre defensive midfielders, and centre midfielders (189.061 ± 4.44 cm; 184.73 ± 6.76 cm; 182.73 ± 6.12 cm; and 180.05 ± 5.32 cm respectively). The tallest 'wide' position on average was right back (179.70 ± 5.43 cm). Interestingly, the test for a significant difference between right back and left back returned a t-test score of 2.12 (statistically significant at the 5% level). The average heights of left wingers and right wingers were not significantly different from each other ($t = 0.37$).

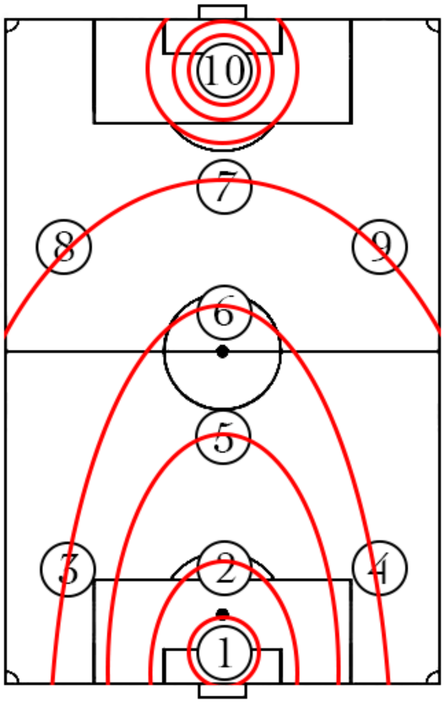


Figure 3: Contour map showing qualitatively how the heights are distributed across the playing positions.

The general pattern, illustrated qualitatively in Figure 3, is one of falling heights along the centre line, from the goalkeeper to the attacking midfielder, with the exception of the centre forward. There is also an independent trend of falling heights from the centre to the wide positions.

4.1.2 Regrouping positions

Based upon the results above, some positions with similar mean heights were then grouped together with the aim to achieve statistically significant separation between the groups:

- Goalkeepers (GK)
- Centre backs (CB)
- Centre forwards and centre defensive midfielders (CEN)
- Left backs, right backs, and centre midfielders (MNB)
- Left wingers, right wingers, and centre attacking midfielders (ATT)

It is important to recall that the main justification for these categories is the differences in mean height between positions rather than proximity on the pitch. Goalkeepers and centre backs kept their original categories (GK) and (CB) respectively, as they already differed significantly from neighbouring groups. The analysis was then repeated using these new groups, with the methodology and significance levels unchanged. The results are shown in Table 3.

Player position/ group	No. of entries	Mean height (cm)	Standard error (cm)	t-value	CI left	CI right	$p(H_0)$	$p(H_1)$
GK	58	190.98	0.53	2.77	0.56	3.28	5.58E-03	0.9972
CB	97	189.06	0.45	6.45	3.60	6.75	1.15E-10	1.0000
CEN	97	183.89	0.66	5.92	3.08	6.13	3.14E-09	1.0000
MNB	166	179.28	0.40	3.50	0.97	3.43	4.63E-04	0.9998
ATT	115	177.09	0.48					
Total:	**533**							

Table 3: Values for the regrouped positional categories.

The difference in mean height is now significant between each group and the next tallest group, with t-statistics ranging from 2.77 to 6.45. The prediction model described in Section 4.2 uses these updated positional categories.

4.2. Position prediction model

The position prediction model uses probability mass functions built from the EPL data to predict player positions: given a specific height (in cm) as input, the model will return the numerical probability that a player of that height will play in any one of the five positional groups defined in Section 4.1.2.

Height from (cm)	Height to (cm)	p (GK)	p (CB)	p (CEN)	p (MNB)	p (ATT)
163	170	0	0	0.087	0.391	0.522
171	172	0	0	0.095	0.286	0.619
173	174	0	0	0.065	0.548	0.387
175	176	0	0.029	0.147	0.500	0.324
177	178	0	0	0.130	0.444	0.426
179	180	0	0.037	0.204	0.463	0.296
181	182	0	0.075	0.275	0.325	0.325
183	184	0.036	0.073	0.218	0.527	0.145
185	186	0.091	0.309	0.236	0.273	0.091
187	188	0.234	0.426	0.213	0.128	0
189	190	0.275	0.500	0.175	0.050	0
191	192	0.296	0.333	0.296	0.074	0
193	194	0.455	0.409	0	0.045	0.091
195	196	0.250	0.500	0.250	0	0
197	201	0.500	0.286	0.214	0	0

Table 4: Probabilities. The highest probability for each height bin is highlighted in red, with the second-highest highlighted in blue.

Some heights, particularly those on the fringes of the dataset, have relatively few data points, and so the data were re-binned in order to ensure sufficient statistics in each bin. Heights between 163 cm and 170 cm were consolidated into one bin, and 197 cm to 201 cm likewise.

The bins in the middle were merged into 2 cm intervals, as opposed to the initial 1 cm intervals. Then, for each height bin, I computed the frequencies of each positional group, relative to the total number of players in that height bin. The relative frequencies were used as estimators for the corresponding probabilities. The result is a lookup table giving the numerical probabilities associated with each positional group (GK, CB, CEN, MNB, and ATT), for any height in the range 163-201 cm, shown in Table 4. The most probable position for each height bin shows a correlation between height and position that is broadly consistent with the results of the position analysis by group in Section 4.1.2.

The prediction model was tested using an independent data sample from the EFL Championship, the division below the EPL. I made sure to use data from another English league in order to minimise potential biases arising from ethnic height differences or variations in playing styles across countries that could affect the distribution of heights across positions. The sample consisted of 55 players, drawn from the two teams that happened to be listed first in Transfermarkt's Championship database (Burnley and Watford). The lookup table was employed to predict each player's most likely positional group, and these predictions were then compared to the players' main playing positions, using the same groupings. The model successfully predicted the positional group for 49.1% of the players (27 out of 55).

The model's performance was measured by comparing the result to a uniform prediction model, which assumes equal probabilities for all *individual* playing positions and does not take into account any association with player height (that is, assuming a team plays with two centre backs and one of each of the other positions, for the sake of example: p(GK) = p(CB$_1$) = p(CB$_2$) = p(CF) = p(CDM) = p(CM) = p(RB) = p(LB) = p(CAM) = p(LW) = p(RW) = 1/11, which, when grouped, translates to: p(GK) = 1/11, p(CB) = p(CEN) = 2/11, p(MNB) = p(ATT) = 3/11). This resulted in a success rate of 21.7%, suggesting that the height-based position prediction model holds comparatively superior predictive power. The expected success rates were then also computed for each of the two models. The height distribution of the EPL dataset was used to estimate the population distribution for the expectation based on Table 4, whereas a uniform population distribution (p = 1/11 for each *individual* position) was used with the reference model. The resulting expected success rates were 45.2% and 22.3%, respectively. The two comparisons demonstrate that the performance of the height-based position prediction model exceeds that of the reference model by more than a factor of 2.

5. Summary and conclusions

This study investigated the relationship between height and playing position in football. Players were initially sorted into specific positions, presenting a more targeted approach than the ‘four horizontal lines’ seen in the existing literature. Amongst the central players there was a trend of descending height, moving from the goalkeeper up the pitch, with the exception of the centre forward which was the third-tallest position overall by mean height. Wide players were generally shorter than central players, with left wingers and right wingers being the two shortest positions by mean height. Based on the statistical significance of the differences in mean height, some positions were then clustered together to make five new categories that differed significantly from each other in mean height. Using the EPL data, a height-based position prediction model was developed using probability mass functions, and tested against a reference model which gives equal probabilities to each individual position and takes no account of player height. The height-based model outperformed the reference model by over a factor of two, both in expectation and when tested with real data samples. Future iterations of this study could build upon this analysis by assigning unequal weights to players’ second and third positions in order to incorporate them into the model. This would give smoother probability mass distributions. The study could also be repeated with a larger dataset, including former EPL players from previous seasons, or it could be repeated for non-English leagues.

6. Bibliography

- BELL, W., & RHODES, G. (1975). The Morphological Characteristics of the Association Football Player. *British Journal of Sports Medicine*, 9(4), 196–200.
- MATKOVIC, B. *et al*. (2003). Morphological differences of elite Croatian soccer players according to the team position. *Collegium Antropologicum*, 27(1), 167–74
- HAZIR, T. (2010). Physical Characteristics and Somatotype of Soccer Players According to Playing Level and Position. *Journal of Human Kinetics*, 26, 83–95
- SUTTON, L. *et al*. (2009). Body composition of English Premier League soccer players: Influence of playing position, international status, and ethnicity. *Journal of Sports Sciences*, 27(10), 1019–1026
- TRANSFERMARKT (2023). *Premier League* [viewed 15 March 2023]. Available from: https://www.transfermarkt.co.uk/premier-league/startseite/wettbewerb/GB1
- TRANSFERMARKT (2023). *Championship* [viewed 4 April 2023]. Available from: https://www.transfermarkt.co.uk/championship/startseite/wettbewerb/GB2
- COWAN, G. (1998). *Statistical Data Analysis*. Oxford: Oxford University Press
- LIPSCHUTZ, S., & SCHILLER, J. (1998). *Introduction to Probability and Statistics*. New York: McGraw-Hill