Predicting goals and results in the Premier League from the perspective of Tottenham Hotspur

Kasper Nikolaj Michelsen

School of Communication and Culture, University of Aarhus

Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

AU ID: au612877

Lecturer: Chris Mathys

June 2nd, 2023

Abstract

Researchers have, through time, proposed many different approaches in an attempt to forecast and predict future sports events. In this project, forecasts are produced using data from Premier League games with a specific focus on the matches played by Tottenham Hotspurs (Spurs) through the seasons from the year 2008 until 2016. Two multivariate time series regression models are constructed, estimated, and evaluated in order to try to forecast the goal difference between Spurs and opposing teams and the total number of goals scored in future games. The forecasts are evaluated and it is concluded that the error measures of the models indicate too big uncertainty in the generated predictions. While the author did not find the optimal combinations of predictors, the methodological approach, discussion of variables, and its results can still be used as valuable knowledge and inspiration for future studies within sports forecasting research.

Keywords: Forecasting, Time series regression models, Premier League, Sport, Prediction

Introduction

Football (also referred to as soccer) is the most popular sport in the world and is watched daily across the globe. In particular, the best league in England, the Premier League, has been growing in interest and according to Wikipedia, the league was broadcast to more than 643 million homes and potentially 4.7 billion viewers in 2021 ("Premier League," 2023). Forecasting of football has been of great interest both to scientists, betting companies, and dedicated football fans with different objectives in mind. Whether it is to beat bookmakers or investigate social patterns in team sports, various methods have been explored in an attempt to predict future outcomes of football matches with relatively great success. Yiannis and colleagues (2006) utilized multivariate ARIMA modeling to predict match outcomes of three different teams in the Premier League season 1997/1998, Schumaker and colleagues (2016) constructed a model from sentiment analysis of relevant tweets to predict wins in the Premier League, and Baboota & Kaur (2019) utilized machine learning methods to predict outcomes in the Premier League. With inspiration from previous research and the author's great interest in the Premier League and, in particular, the team Tottenham Hotspur (Spurs), it will be attempted to construct a forecasting model with the purpose of predicting outcomes of future Spurs matches. Previous research has had a great focus on the match result (Baboota & Kaur, 2019; Joseph et al., 2006; Koopman & Lit, 2015; Owramipur et al., 2013; Schumaker et al., 2016; Yiannakis et al., 2006), but to the best of the author's knowledge, the total number of goals scored has not been shown the same interest in the literature. This project complements existing literature by exploring the forecasting possibilities of Premier League matches by using multivariate time series regression models estimated on data across many seasons. It will be attempted to construct two forecasting models. One with the purpose of predicting the goal difference between Spurs and the opposing team in a Premier League match, and the second model with the purpose of predicting the total number of goals in a Spurs match. Before the construction of the models, a brief motivation and description of the proposed predictor variables are given below.

Home/away

Football teams and sports teams in general playing at home or away have been thoroughly investigated, and the home advantage is a commonly accepted phenomenon (Baboota & Kaur, 2019; Smith, 2003; Yiannakis et al., 2006). Various reasons, such as the absence of travel fatigue and the local crowd cheering for the home team, have earlier been proposed as part of the explanation (Smith, 2003; Yiannakis et al., 2006). Thus, the home/away predictor seems to be important to include in a model when predicting the outcome of a football match.

Form and differences in level

Naturally, the level of Spurs and the opposing team differs from game to game and season to season. Thus, another key variable when trying to predict match outcomes and the number of goals scored is to determine the differences in the overall level and, in particular, the defensive and offensive strength of the teams. Furthermore, in a study by Baboota & Kaur (2019), they constructed a *form* variable that reflected the outcomes of previous matches. A similar dynamic, though a bit simpler, is incorporated into a variable for this project (also named *form*). The specific dynamics of the variable are explained in the methods section.

Hypotheses and the purpose of this project

The research questions (or hypotheses) in this project are, to a large extent, exploratory. The point of setting up these hypotheses is to explore whether this approach and proposed variables appear promising. To test the *H1,* the goal difference in games is forecasted and is subsequently used to determine the outcome of the match (win, draw, loss). To test the *H2,* total goals scored are forecasted and the values are subsequently used to determine whether a match will have above or below 2.5 goals. The accepted thresholds for the forecast models are to be better than chance level.

H1: Using the best-performing combination of proposed key variables, it is possible to estimate a time series regression model that is able to predict the outcome of future Spurs games above chance level

H2: Using the best-performing combination of proposed key variables, it is possible to estimate a time series regression model that is able to predict if more than two goals will be scored in a future game above chance level

Methods

The data preprocessing and analysis were conducted in R in Rstudio (R Core Team, 2023; Rstudio Team, 2021) using the Tidyverse library (Wickham et al., 2019). The full analysis can be accessed here: https://github.com/KasperNM/DS_Exam

Data

The dataset used for this project is gathered from the 'The European soccer database' (*European Soccer Database*, n.d.), which is publicly available at Kaggle.com. The data consists of data from more than 25.000 football matches across 11 countries from seasons between 2008 and 2016. This project attempts to investigate and forecast outcomes of games where Spurs is included. Thus, matches played by Tottenham Hotspurs were extracted for all the seasons available, which resulted in data from eight seasons (8*38=304 matches). The dataset consists of information on home- and away-team goals and information regarding odds probabilities set by the bookmakers from various betting companies. For this project, only data concerning home- and away-team goals and date/season information were kept of the initial variables.

Computed variables

Home/away

As mentioned earlier, playing on home ground or away from home has shown to be an influential predictor in match forecasting (Baboota & Kaur, 2019; Smith, 2003; Yiannakis et al., 2006). The variable is modeled as a dummy variable of zeros and ones (one if Tottenham is playing at home).

Differences in level between teams

In an attempt to capture the difference in strength between teams, the FIFA ratings from the 'team stat database - FIFA Index' (*Team Stats Database - FIFA 09 - FIFA Index*, n.d.) starting from FIFA 09 (released around the same time as the beginning of the 2008/2009 Premier League Season) and each year until the season 2015/2016. The FIFA index includes an overall rating of the individual team and ratings of their offensive and defensive strength. These three metrics were manually encoded in Google Sheets (*Google Sheets*, n.d.) into the dataset for each team throughout the eight seasons. The FIFA ratings have previously been used as a representation of team strength in forecasting research (Baboota & Kaur, 2019). In this project, the ratings were used to construct three variables. The rating difference between Spurs and the opponent team for every match (*rate_diff*), the difference between Spurs' offensive rating and the opponent team's defensive rating (*off_adv*), and the difference between Spurs' defensive rating and the opponent team's offensive rating (*def_adv*).

Form

As previously mentioned, this variable is inspired by work done by Baboota & Kaur (2019). However, the variable is not computed in the same manner. The *form* variable is based on the performance of previous matches and is an integer from -inf to +inf. The variable starts off at zero and depending on the outcome of the following match, the variable gets updated. Wins against stronger teams suggest better current form, thus adding more to the *form* compared to beating a weaker team. The individual outcomes can be seen in Table 1. After a full season of games (38), the form variable is reset to zero.

Stronger team			equal strength			Weaker team		
Win	Draw	Loss	Win	Draw	Loss	Win	Draw	Loss
3	1	-1	2	0	-2	1	-1	-3

Table 1: The coding scheme for the form variable

Analysis

As previously mentioned, the analyses attempt to investigate if the constructed variables can be used to forecast the goal difference and amount of goals in future Spurs matches. The data was divided into a training set for the model estimation and a test set to evaluate the generated forecasts. The training set consists of data for all of the first seven seasons and half of the matches in the final season of the data set (n=285). The test set was the remaining games of the last season (n=19). Before constructing and estimating any models the data is plotted over time, where a match is used as the index for a step in time.

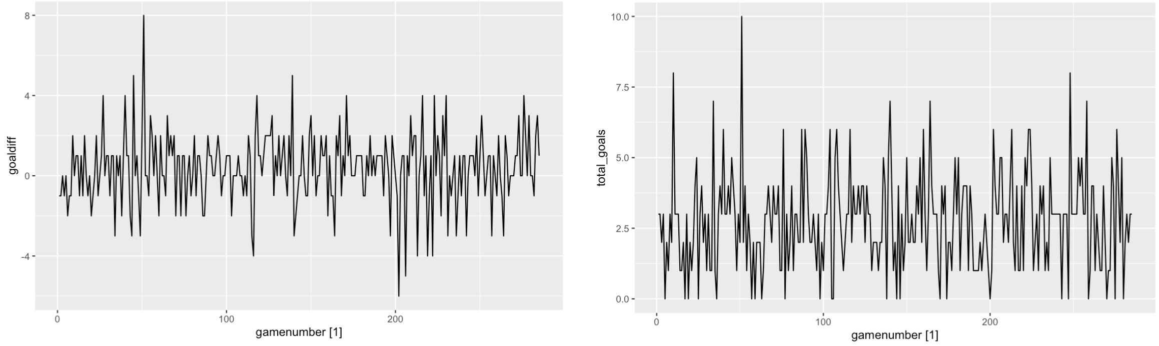


Figure 1: left) The goal difference in matches between Spurs and the opposing team across the 285 games. Other than a few abnormal values, no particular patterns are identified. Right) The total number of goals in a Spurs match across 285 games. Again no particular patterns are identified.

Both the time series visualizations in Figure 1 and the autocorrelation plots (See appendix A) do not suggest any particular trends, seasonality, or cyclicality. Thus, it is decided not to include any function or variable accounting for any of these patterns in the regression models. In the attempt to forecast goal differences in future Spurs games, several time series linear regression models (TSLM) with different combinations of the three proposed key variables are estimated and evaluated on several model performance metrics. The key variables are *home, rate_diff*, and *form.* The glance() function from the fable package (O'Hara-Wild et al., 2023) allows for efficient evaluation of the optimal predictor combination by including the calculation of the Mean-Squared-Error (MSE) of time series cross-validation (CV), adjusted r-squared, and corrected Akaike information criterion (AICc). The model estimations can be seen in Table 2.

Model syntax:	**adj R^2**	**CV**	**AIC**	**AICc**	**BIC**
GD ~ home + rate_diff + form	0.123	2.714	287.1	287.3	305.4
GD ~ home	0.044	2.944	309.8	309.8	320.7
GD ~ home + rate_diff	0.126	2.700	285.1	285.3	299.7
GD ~ form	-0.003	3.087	323.6	323.7	334.6
GD ~ rate_diff + form	0.077	2.851	300.9	301	315.5
GD ~ rate_diff	0.080	2.835	298.9	298.9	309.8

Table 2: Evaluation of combinations of model predictors. The model with the syntax: GD ~ home + rate_diff has the highest adj. R^2 and lowest CV and AICc values.

According to the model tests, the model that explains the data best when taking overfitting into account (lowest CV and AICc values and highest adj. R^2 value) seems to be the model with the following syntax (referred to as GD-model):

GD ~ home + rate_diff

To check assumptions, the gg_tsresiduals() from the feasts package (O'Hara-Wild et al., 2023) is applied to the model to generate time plot, autocorrelation plot, and histogram of the residuals.

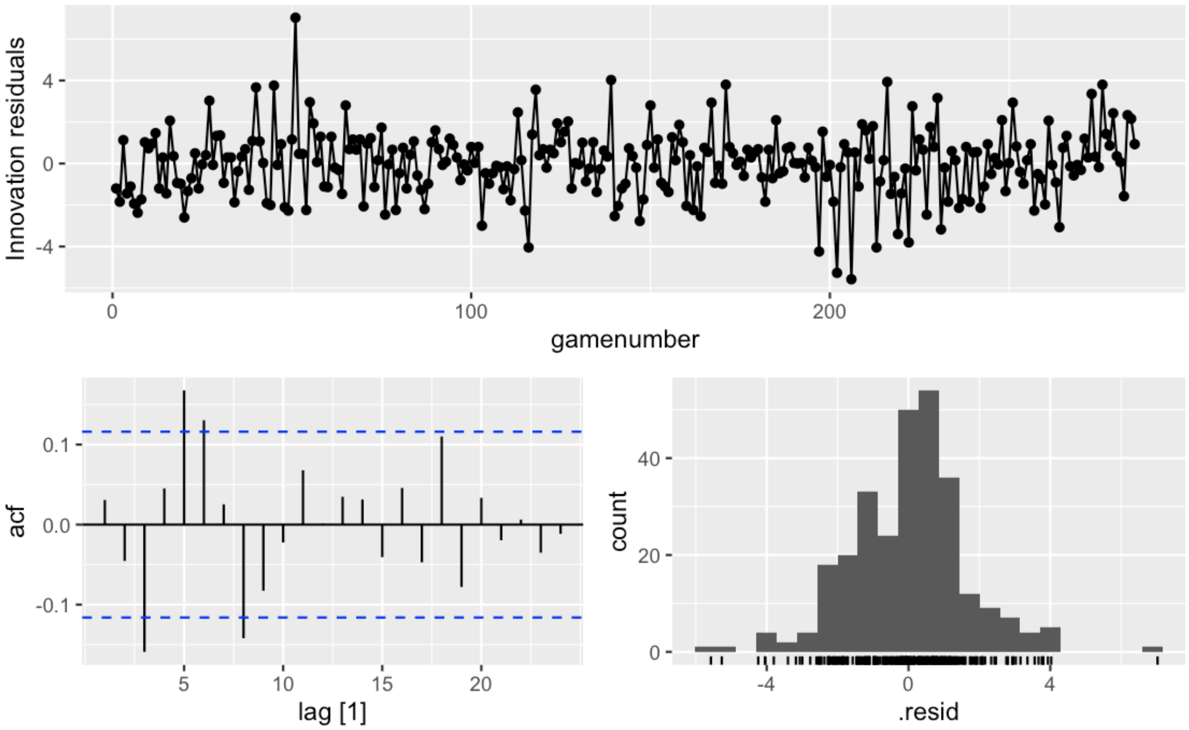


Figure 2: Top) Time series plot of residuals of the GD-model. Bottom left) Autocorrelation plot of the residuals of the GD-model. Bottom right) Histogram of the distribution of residuals.

As can be seen in Figure 2, the residuals are relatively white noisy, with a mean around zero and constant variance. There seems to be some sort of autocorrelation within the residuals, which is not optimal and indicates that model might be inefficient (Hyndman & Athanasopoulos, 2021, Chapter 7). However, for this project, this model is deemed sufficient for the further steps of the analysis.

In the attempt to forecast total goals scored, TSLM models with different combinations of the variables *off_adv, home, form, and def_adv* were estimated, and evaluation metrics were acquired using the glance() function (O'Hara-Wild et al., 2023) (See Table 3).

Model syntax	**adj R^2**	**CV**	**AIC**	**AICc**	**BIC**
GS ~ off_adv + home + form + def_adv	0.009	2.984	312.8	313.1	334.7
GS ~ off_adv + def_adv + home	0.01	2.968	311.5	311.7	329.7
GS ~ off_adv + home	0.011	2.952	310.2	310.3	324.8
GS ~ def_adv + home	0.013	2.945	309.5	309.7	324.1

GS ~ form + home	0.006	2.966	311.6	311.8	326.3
GS ~ off_adv + form	0.003	2.978	312.3	312.5	326.9
GS ~ def_adv + form	0.005	2.973	311.9	312	326.5
GS ~ def_adv + home + form	0.012	2.962	310.8	311	329.1

Table 3: Evaluation of combinations of model predictors. The model with the syntax TG ~ def_adv + home seems to have the best fit on the data (highest adj. R^2, lowest CV and AICc).

As Table 3 suggests, the model including the predictors *def_adv* and *home* seems to be the best fit. Model syntax (referred to as GS-model):

GS ~ def_adv + home

As Figure 3 indicates, residuals have constant variance and are centered around zero. The autocorrelation plot shows almost no signs of autocorrelation.

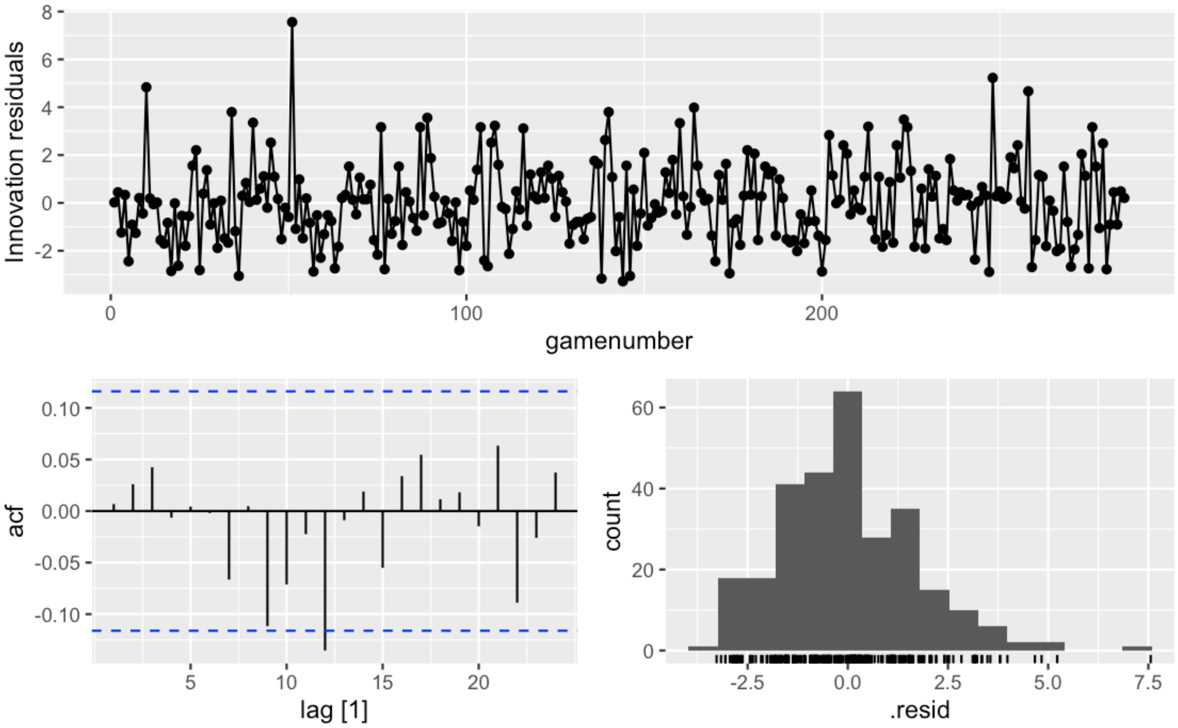


Figure 3: Top) Timeplot of the residuals. Bottom left) Autocorrelation plot of the residuals. Bottom right) histogram of the residuals.

Forecasting was done using the forecast function in the fabletools package (O'Hara-Wild et al., 2023). As mentioned previously, the models were estimated on data from all of Spurs' matches from season 2008/2009 until 2014/2015 and the first 19 games of season 2015/2016 (n=285). Subsequently, the model was set to forecast the remaining 19 games of season 2015/2016. To evaluate the forecasting models, accuracy measures such as RMSE, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are computed, and the forecasts are plotted against the actual values (Hyndman & Athanasopoulos, 2021, Chapter 5).

Additionally, it is assessed how many times the GD-model correctly predicted a game outcome (Win, Draw, Loss) and how many times the GS-model correctly predicts if goals scored in a Spurs game exceed 2 goals. If the GD-model predicts a goal difference above 0.5 it is regarded as a Spurs Win, between 0.5 and -0.5 is regarded as a draw, and below -0.5 it is regarded as a predicted loss. Additionally, if the GS-model predicted above 2.5 it was regarded as 2+ goals, and below 2.5 was regarded as 2 goals or lower.

Results

The forecasting of goal difference in Spurs matches is visualized in Figure 4 and the forecasting of total goals scored is visualized in Figure 5.

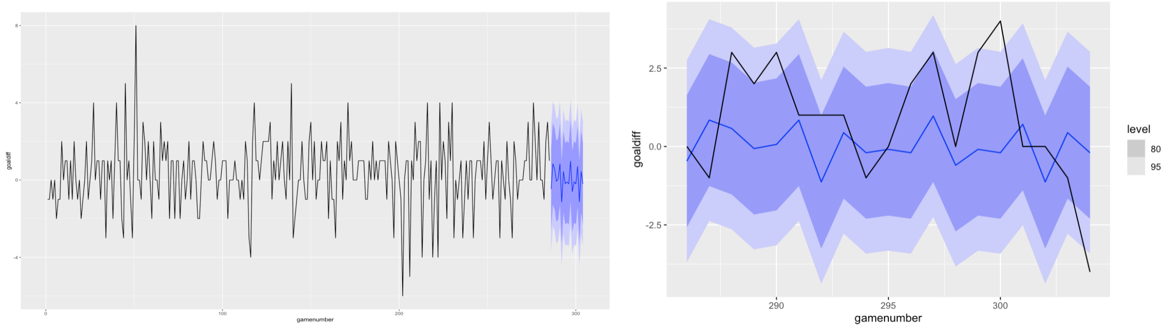


Figure 4: left) Forecast produced by the GD-model of the predicted goal difference in the last 19 matches of the 2015/2016 season. Right) The forecast (blue line) is plotted against the actual goal difference (black line) with 80% prediction intervals (dark blue) and 95% prediction intervals.

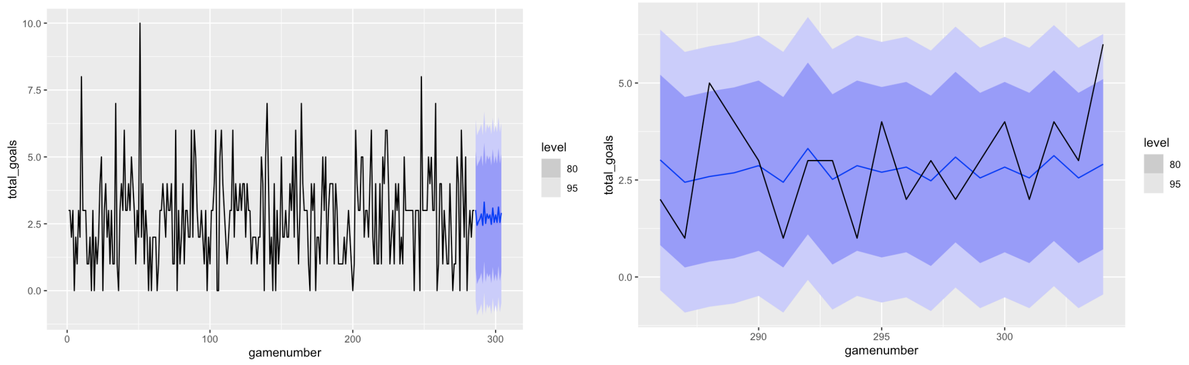


Figure 5: Left) Forecast produced by the GS-model of the predicted goals scored in the last 19 matches of the 2015/2016 season. Right) The forecast (blue line) with 80% prediction intervals (dark blue) and 95% prediction intervals is plotted against the actual goals scored (black line).

	RMSE	***MAE***	***MAPE***	***MASE***
GD-model	*2.086*	*1.719*	*inf*	*0.898*
GS-model	*1.314*	*1.092*	*49.692*	*0.599*

Table 4: Evaluation metrics of the forecasts.

Evaluation metrics of the performance of the model forecast are shown in Table 4

H1: GD-model	***Number of games***	***Correctly predicted***
Wins	*10*	*3*
Draws	*5*	*2*
Losses	*4*	*0*
Total	***19***	***5***
H2: GS-model	***Number of games***	***Correctly predicted***
Above 2.5 goals	*12*	*11*
Below 2.5 goals	*7*	*2*
Total	***19***	***13***

Table 5: The classification of the forecast values when predicting the match outcomes: Win, draw, loss, and above/below 2.5 goals.

In the categorization of the forecasting values, the GD-model correctly predicted 5 out of 19 match outcomes (see Table 5). This means that the GD-model got around 26.3% of predictions correct, which is below the chance level of 33.33%. The GS-model correctly predicted 13 out of 19 match outcomes correctly. The model got 68.4% of predictions correct, which is above the chance level (50%) and above the number of correct predictions one would get by just sticking to one of the categories e.g. ‘above 2.5 goals’ (63%).

Discussion

H1: GD-model

The GD-model's forecast of the last 19 games of the 2015/2016 season had an RMSE of 2.086, suggesting a mean squared error of around 2 goals between the predicted value and the actual values. Additionally, the Mean Absolute Error (MAE) was 1.719. The Mean absolute percentage error (MAPE) was inf due to some of the observations being 0 (Hyndman & Athanasopoulos, 2021, Chapter 5). The visual representation of the forecast in Figure 4 supports the large error measures. Additionally, in the attempt to use the model to predict match outcomes, it was relatively poor (26.3% correctly predicted), which makes sense when the RMSE and MAE were around 2 goals. Two goals can be quite determining for a match outcome. Thus, The exploratory *H1* is not supported. However, the possibility of another representation or combination of the proposed predictors being more appropriate for this forecasting approach cannot be rejected.

H2: GS-model

The GS-model's forecast of the last 19 games of the 2015/2016 season had an RMSE of 1.3 and an MAE of 1.09, indicating an error of around a little more than a goal on average between predicted and actual values. The MAPE was at 49.7%, which is assessed to be quite a high absolute error percentage. In general, these metrics suggest better predictability compared to the GD-model and the model predicted the test data (above/below 2.5 goals) better than chance. This might suggest that playing home or away and the difference between Spurs' defensive strength and the opposing team's offensive strength are appropriate predictors when forecasting the total number of goals scored. However, the large values of MAE and RMSE, and MAPE are still indicating that forecasts generated from this model should be interpreted and used with caution. In this case, the *H2* was supported but one cannot rule out the possibility of other combinations of predictors might be more beneficial for a successful forecasting model.

General discussion, limitations, and future research

Overall the models constructed for the forecasting attempts are not deemed successful. The uncertainty apparent in the error measurements is relatively substantial. From a theoretical point of view, this is not surprising due to the overwhelmingly complex and comprehensive interplay between influential factors that determines a football game. The results might suggest either that the variables were not the optimal predictors, the construction of the variables was not representative enough, or that influential predictors were missing.

It is worth noting that the ratings and the calculated differences between team strengths were based on the FIFA team index. Even though the ratings are supposed to reflect the real world as best as possible, a single rating for each team for an entire season might not capture reality well enough. Furthermore, the form variable assumes that games from the entire season represent the current form of the team. While this might be true in the sense that games from the entire season will reflect the number of points gathered, the players within the team might not still be too influenced by what happened 10 games ago. Additionally, winning and losing streaks do not necessarily affect the team in opposite directions. In the work by Mizruchi (1991), evidence shows that a basketball team that won the previous match was less likely to win the current match. The form variable should for future studies also be computed for the rest of the teams so that differences in form can be calculated. When assessing the evaluation of the models none of the chosen models included the form variable, which could indicate that the variable is not adequately depicting an influential real-world factor.

It is also important to note that a lot of other factors that are not included in the model might be influential for a football game. Factors like the number of rest days between games, some teams (Spurs included) might have played additional games some seasons due to European competitions, and key players missing games due to injury.

To sum up, the produced forecasts from the models yielded high error measures and were not convincing enough in the forecast evaluation. This suggests that the optimal combination of predictors was not found or that key predictors were missing. In other words, the project did not find compelling new evidence, but the results, the methodological considerations, and the considerations of predictor selection still offer relevant and useful perspectives for future research within this field.

Bibliography

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

European Soccer Database. (n.d.). Retrieved May 26, 2023, from https://www.kaggle.com/datasets/hugomathien/soccer

Google Sheets. (n.d.). Retrieved May 31, 2023, from https://docs.google.com/spreadsheets/u/0/

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on May, 31, 2023.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553. https://doi.org/10.1016/j.knosys.2006.04.011

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *178*(1), 167–186.

Mizruchi, M. S. (1991). Urgency, Motivation, and Group Performance: The Effect of Prior Success on Current Success Among Professional Basketball Teams. *Social Psychology Quarterly*, *54*(2), 181–189. https://doi.org/10.2307/2786935

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering*, 812–815. https://doi.org/10.7763/IJCTE.2013.V5.802

Premier League. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Premier_League&oldid=1156718935

Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, *88*, 76–84. https://doi.org/10.1016/j.dss.2016.05.010

Smith, D. R. (2003). The Home Advantage Revisited: Winning and Crowd Support in an Era of National Publics. *Journal of Sport and Social Issues*, *27*(4), 346–371. https://doi.org/10.1177/0193732503258637

Team Stats Database—FIFA 09—FIFA Index. (n.d.). Retrieved May 26, 2023, from https://www.fifaindex.com/teams/fifa09_5/?league=13&order=desc

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yiannakis, A., Selby, M. J. P., Douvis, J., & Han, J. Y. (2006). Forecasting in Sport: The Power of Social Context — A Time Series Analysis with English Premier League Soccer. *International Review for the Sociology of Sport*, *41*(1), 89–115. https://doi.org/10.1177/1012690206063508

Software and packages

O'Hara-Wild M, Hyndman R, Wang E (2023). _fable: Forecasting Models for Tidy Time Series_. R package version 0.3.3, <https://CRAN.R-project.org/package=fable>.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

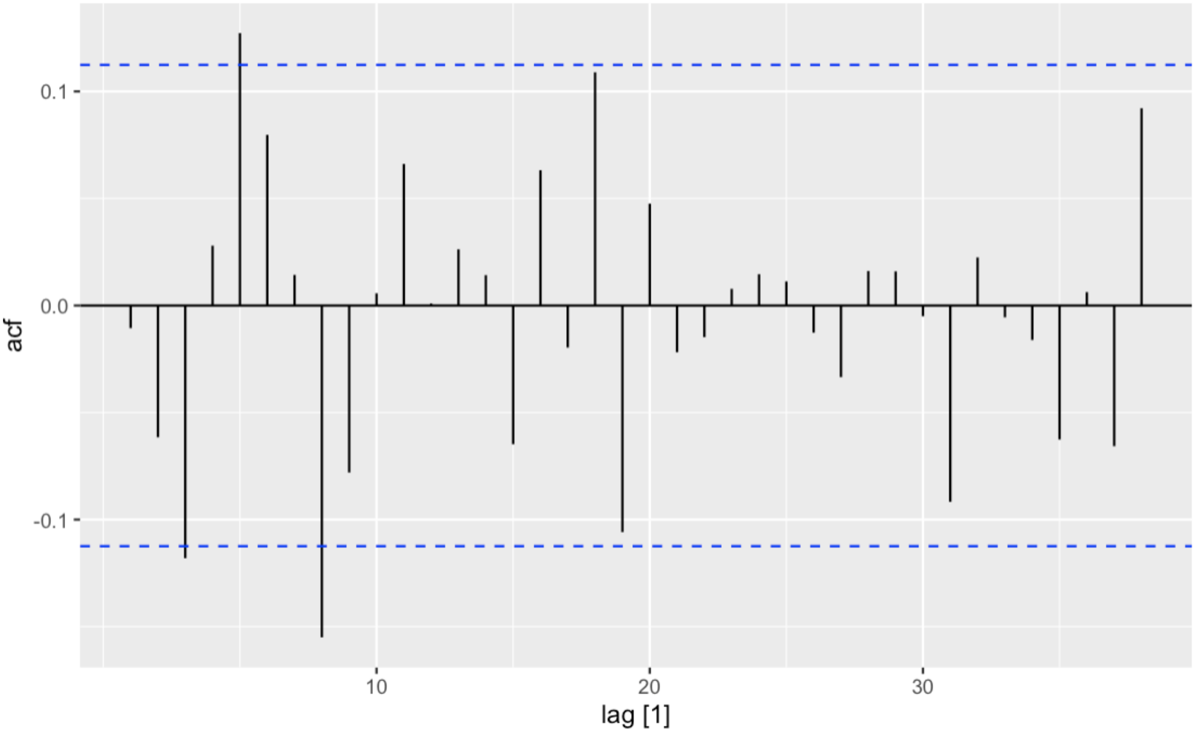
O'Hara-Wild M, Hyndman R, Wang E (2023). _fabletools: Core Tools for Packages in the 'fable' Framework_. R package version 0.3.3, <https://CRAN.R-project.org/package=fabletools>.

O'Hara-Wild M, Hyndman R, Wang E (2023). _feasts: Feature Extraction and Statistics for Time Series_. R package version 0.3.1,

<https://CRAN.R-project.org/package=feasts>.

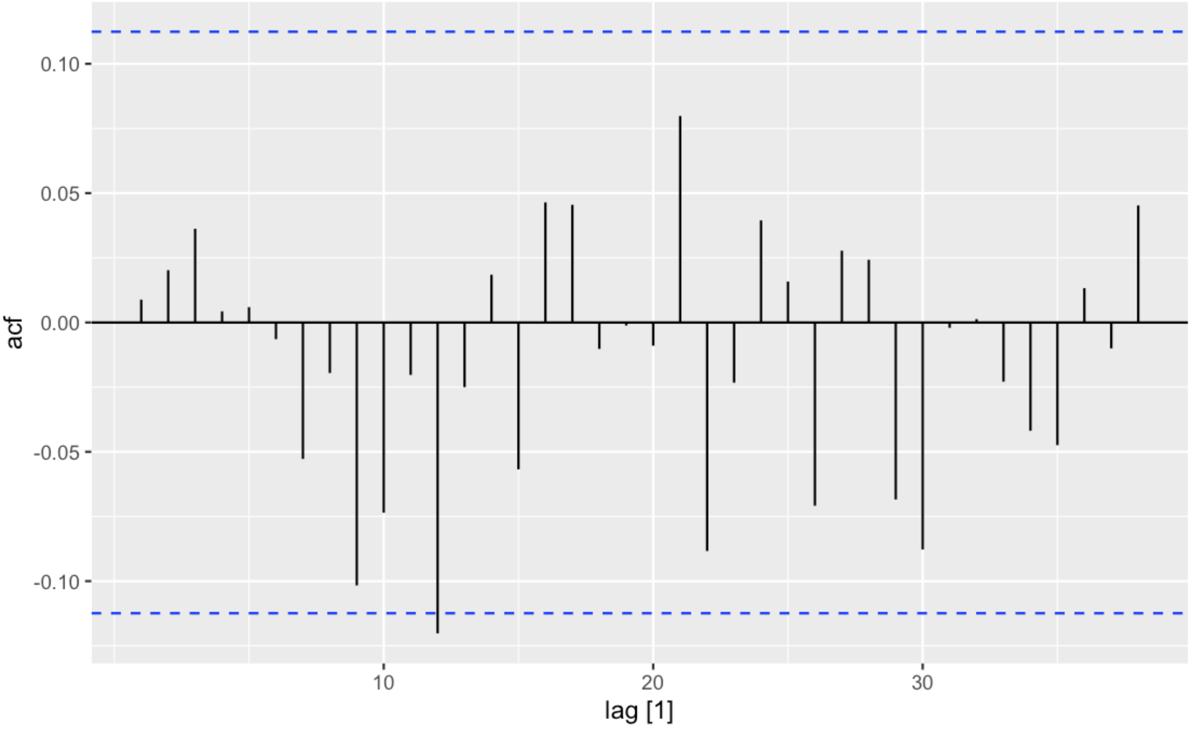
Appendix A

Goal difference autocorrelation



Autocorrelation plot of the goal difference time series. The data show significant signs of autocorrelation in some of the early lags.

Total goals scored



Autocorrelation plot of total number of goals scored time series. Only one lag shows signs of autocorrelation.