

MULTIVARIATE STATISTICAL MODELS TO FORECAST THE RESULTS OF EURO2016 QUALIFIERS

Bence Jámbo, Máté Jámbo, Dávid Szabó



ABSTRACT

The main idea of the paper was whether the actual performance of players can explain the future match results better than national team's previous match results. For the statistical method, we observed 58 matches in the European qualifiers with arbitrary sampling. In every occasion, we examined every player's actual performance from their club team's matches, before their actual national match. Our model has cca. 10 000 data.

We forecasted the European qualifiers' results of Hungarian national team by two multivariate regression models, based on the parameters of the players measured in their previous club matches. The forecast from our second model was more efficient than from the first one, the outcome of the observed matches was predicted right in 62.1%, while the number of goals scored was correct in one-third of the cases.

KEYWORDS

Statistical modelling. Multivariate regression. Factor analysis. Predicting football matches.

JEL CLASSIFICATION: C31, C38, C51

INTRODUCTION

There are a lot of ways to predict the outcome and result of a football game. At the beginning of our paper, we raised two hypotheses to help us predicting the possible outcomes. During the test of our first hypothesis, we studied the actual form of all the players in the selected WorldCup qualifier matches in both teams – based on the matches played by these players in their clubs right before (but at most one month before) the actual national game –

considering many points of view, so basically our model is based on cca. 10 000 data. During the test of our second hypothesis we corrected the actual form with the players' basic skills. While gathering all the data, we managed to get essential help from InStat, a football data base available on the internet, and from EA Sports video games FIFA 2014 and 2015.

The prognosis was done by the help of SPSS and we compressed the defending players' five defending, and the attacking players' six attacking qualities in three-three factors, thus creating defending and attacking factors for teams, which have enough explanatory power.

As a result, we received, that there is a significant connection between the players' actual form in the club football, and the outcome of the matches played by their national teams. Our multivariate regression model was able to predict the real outcome of the examined games with a rate of 56.9%, while it predicted the exact number of goals scored by the teams in 33.6% of the cases.

We gratefully acknowledge the help of István Kovách for providing us the InStat football database.

1. OVERVIEW OF THE LITERATURE

We searched for domestic and international articles, regarding the points of view, which must be considered, when we want to predict the efficiency of national teams.

Lago-Peñas (2009) examined *the effects of a tight match-schedule on the performance of a football team*. He generally found that Spanish teams did not underachieve on weekend matches, even when they had more during the week. What is more, participants of the Champions League sometimes played even better. The risk of underachievement did not grow in the first 15 weeks, even though they had to play more and more matches per weeks. This leads to the conclusion, that a first class team will not likely perform poorly even with a very tight schedule.

Marek M. Kaminski (2014) points the following concerning the 'Host Paradox' in the FIFA ranking:

- Many times the ranks do not show, what might be obvious or fair to most of the people, in general.
- The points received for a match does not depend on the place of a game, whether it's played home, or in the stadium of the opposite team.
- It also does not broadcast reality, when a team receives more points for defeating Qatar, then for playing a tie with Brazil.

- Many people would find it also obvious and fair, if a number of points for a team would increase in linear relationship with the number of goals it scored.
- Another possible contradiction: let us assume a rank of teams A and B, where A is on first place. In case of the current Fifa ranking it is possible that A plays a match with B, A will get behind B in the ranking, even if A wins.

Soccer Power Index

- The Soccer Power Index (SPI) is the daily refreshable assessment system of the ESPN TV channel, which can predict the possible result of a match from data occurred in the past. The algorithm uses multiple years of data, such as scored and received goals, line-up of the beginning team, and the location of the match. Beside of this, SPI gives more credit to the recent matches and it also takes the importance of a game into account (this way a World Cup match is counted much more important, than a friendly game).

The all-seeing software, which only cannot play football (InStat)

Valerij Lobanovskij, the Ukrainian trainer legend began to write in his copybook, how famous players like Platini, Pelé or Maradona did their tricks, where did they pass, from where to where were they moving on a certain match. This forms the base of the InStat software, continuously developed for over eight years, which contains the data of every player in a 2-3 year time-scale. Nowadays giant football clubs are using the software, teams like Chelsea, Valencia, Roma, Lazio, and the biggest Russian clubs.

Advantage of the home court

The rule of goal scored on opponents' field did not exist in the football world until 1965. But due to the very few winnings in the opponent teams' home on the cup matches of 1964 (only 16% of the teams was able to win away), the rule has been introduced. The main reason of the lack of these successes far from home was that they had to travel a lot to the stadium of the opponent teams and they were in a hostile environment.

We did not refute the change of trend that occurred in the past one or two decades (which trend is: more and more guest-victories are happening on the fields), but the result was however – based on our model being introduced later –, that a team playing in its own stadium is more likely to score a goal, if the form of its attackers and the form of the defenders of the guest team is considered constant.

As we could see from the mentioned articles, there are many aspects from which one can give a forecast on the form of football teams. On the other hand, our aim was to build a model, which can give more explanation.

2. THE DEVELOPMENT OF THE MODEL PREDICTING THE PERFORMANCE OF NATIONAL TEAMS

2.1. Method of gathering data

With arbitrary sampling, we observed 58 qualifying matches of World Cup 2014, European zone, to help us build our statistical model. In each national match, we studied the actual form of all players in both teams – based on their club matches right before (or at most one month before) the actual national game – considering several personal efficiency measures, thus our model is based on cca. 10 000 data.

The basic concept was the following: we divided the 11 players of the national teams into two groups, based on whether they are more in attacker, or in defensive role. We examined the efficiency of the roles of the players according to Figure 1.

Beside these variables, the chart of every examined match contained the number of goals and assists of the players achieved on their former club game, as well as the InStat index of the players of their former club game (see explanation of InStat index later), and the FIFA index of the players.

We observed furthermore the FIFA-rank point of the team at the time of the actual national match, and how many percent of possible points at the last five competitive games did the certain team get (we defined this as the trend form of the national team).

We determined the players' form shown in club teams based on indicators gathered from InStat database, which are the following: defensive (save percent of goalkeepers, successful passes, challenges won, aerial challenges won, successful tackles) and attacking (successful passes, challenges won, aerial challenges won, successful dribblings), which can be seen on the figure below.

Basic concept of the forecasting model



Figure 1 Basic concept of the forecasting model

Source: Self-edited.

2.2. Determining the defense factor of national teams

We extracted the five defensive qualities (variables) of defender players in three factors with factor analysis using SPSS (Varimax rotation), creating three new variables. We marked these defending factors DF_1 , DF_2 and DF_3 .

We used the variance-proportion method in determining the number of the factors. In case of factors which hold significant insecurity in themselves, like examining performance of athletes, explanatory power above 60 percent is acceptable.

Table 1 Factors describing defensive form

Total Variance Explained				
Component	Extraction Sums of Squared Loadings	Rotation Sums of Squared Loadings		
	Cumulative %	Total	% of Variance	Cumulative %
1	29.732	1.347	26.933	26.933
2	54.903	1.317	26.341	53.274
3	74.771	1.075	21.497	74.771

Source: Self-calculated.

The eigenvalue of DF_1 is 1.347, DF_2 is a little less and DF_3 is 1.075. The variance-proportions explained by the actual factor in the total variance are the following (in the same order): 26.933%; 26.341% and 21.497%, so the aggregate variance explained by the 3 factors is 74.771% of the five variables.

After using factor-analysis, extracting the 5 variables into 3 factors, then defining the eigenvalues, we pursued exploring the connections between variables

Table 2 – Explanatory variables for factors of defensive form

	Rotated Component Matrix		
	Component		
	1	2	3
save_percent	.758	-.199	-.239
pass_percent	-.006	.002	.949
challenges_won_percent	.394	.720	.230
aerial_challenges_won_pt.	-.225	.852	-.122
tackle_percent	.753	.184	.220

Source: Self-calculated.

The variables being in stronger correlation with factor DF_1 are the “save percent” and the “tackle percent”. The value of their factorweights are 0.758 and 0.753. The variables in stronger correlation with factor DF_2 are “challenges won” and “aerial challenges won”. The value of their factorweights are 0.720 and 0.852. These two variables being in the same factor is not a coincidence, because both in duels and in aerial duels a player is battling against an opponent player. The variable being in stronger correlation with factor DF_3 is “pass percent”. Being the only factor here, the factorweight is very high, 0.949.

After calculating the DF_i factor values for each team, from these factors we determined one single defense factor (DF) using the rotation sums of squared loadings as weights. Thus the DF describes the actual defense form of a national team at a certain date based on the club form of the defending players.

2.3. Determining the attacking factor of national teams

We extracted six attacking qualities (variables) in three factors with factor analysis, creating three factors describing attacking qualities, marked by AF_i . The eigenvalue of AF_1 is 1.683, AF_2 is a little less 1.623, and AF_3 is 1.136. Their explanatory power of the total variance is (in the previous order): 28.052%; 27.043% and 18.926%, which are 74.021% in total.

After using factor-analysis, extracting the 6 variables into 3 factors, then defining the eigenvalues, we pursued exploring the connections between variables.

Table 3 Factors describing attack forms

Total Variance Explained				
Component	Extraction Sums of Squared Loadings	Rotation Sums of Squared Loadings		
	Cumulative %	Total	% of Variance	Cumulative %
1	28.838	1.683	28.052	28.052
2	56.016	1.623	27.043	55.095
3	74.021	1.136	18.926	74.021

Source: Self-calculated.

The related variables, which are in stronger correlation with factor AF_1 are “challenges won”, “aerial challenges won” and “dribble percent”. The variables belonging to AF_2 are “assist percent” and “shot percent”, while the variables of AF_3 are “aerial challenges won”, “dribble percent”, and “pass percent” attack qualities.

Table 4 Explanatory variables for factors of attacking form

	Rotated Component Matrix		
	Component		
	1	2	3
challenges_won_percent	.871	.096	.085
aerial_challenges_won_pt.	.695	-.002	-.492
assist_percent	.075	.871	.168
shot_percent	-.095	.888	-.032
dribble_percent	.648	-.222	.404
pass_percent	.077	.134	.832

Source: Self-calculated.

After calculating the AF_i factor values for each team, from these factors we determined one single attack factor (AF) using the rotation sums of squared loadings as weights. Thus the AF describes the actual attack form of a national team at a certain date based on the club form of the attacking players.

2.4. Our first multiple regression model

We determined the optimal multiple linear regression function using backward elimination method. The first set of variables contained the following:

- difference between current form of attacker players and current form of opponent defender players,
- difference between current InStat index of attacker players and current InStat index of opponent defender players,
- difference between FIFA index of attacker players and opponent defender players,
- number of scored goals of attacker players in their previous club game,
- number of assists of attacker players in their previous club game,
- difference between the trends of their previous five matches of the two national teams,
- difference between the Fifa world rank score of the the two national teams,
- dummy variable of home field.

The independent variable was the number of goals scored and at the end, 5 significant variables remained in the model, which helped us building our first model. The multiple correlation coefficient of our first model is $R=0.717$, so the explanatory variables are in strong connection with the number of scored goals. In the case of coefficient of multiple determination (R^2) the 4 variables together explain the scored goals in 51.5%. If we eliminate the effect of the number of the explanatory variables, then the number of scored goals are explained in 49.3%.

Table 5 The explanatory power of the first regression model

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.717	.515	.493	1.36956

Source: Self-calculated.

In Table 6 we can find the coefficients of the linear regression function and the t-values of the variables. The first explanatory variable is the difference between the current form of the attackers and the defenders of the opponent team. If the attacker form is better/worse with 10 percentage points than the defender form of the opponent team, then the attacker team will score an avarage more/less by 0.3 goals. In this case the t-value of the variable is 4.097, which

is significant at any level of significance used in practice.

The second explanatory variable is the difference between the InStat values of attackers and defenders of the opponent team. If the attacker InStat value is 100 points more/less than the defender InStat value of the opponent team, then the attacker team will score an average more/less by 0.5 goals. In this case the t-value of the variable is 2.028, which is significant on a 4.5% or higher levels of significance.

The “own fifa attacker vs opponent fifa defender” variable combines the player-values of Fifa Football 2014 and 2015. This is a static index, by which we graded the players for the matches of 2012 based on their values in Fifa 2014 and for the matches of 2013 based on their values in Fifa 2015. (Fifa 2014 contains the statistically calculated player-values based on the games played between September 2012 and June 2013 while Fifa 2015 is based on the games played between September 2013 and June 2014.) The player values may vary from 1 to 100, so if the value of the attackers is an average 10 points more/less than the defenders’ value of the opponent team, then the attacker team will score an average more/less by 0.92 goals. In this case the t-value of the variable is 2.836, which is significant on a 0.5% or higher levels of significance.

The fourth variable is a dummy, the advantage of home field. The point of this index is that if two teams with equal player parameters meet on home field or guest field, then the team playing on home field scores an average by 0.675 goals more than playing on guest field. The t-value of the variable is 3.052, which is significant on a 0.3% or higher levels of significance.

Table 6 The first regression model

Model	Unstandardized Coefficients		t	Sig
	B	Std. Error		
own_attacker_vs_opponent defender	.030	.007	4.097	.000
own_attacker_vs_opponent_defender_In Stat	.005	.003	2.028	.045
own_fifa_attacker_vs_opponent_fifa_defender	.092	.032	2.836	.005
own_fifa_worldrank_vs_opponent_fifa_worldrank	-.001	.000	-1.680	.096
own_homefield	.675	.221	3.052	.003

Source: Self-calculated.

2.5. The second model

In case of this statistical model, we assumed that we can predict the outcome of games between national teams based upon the static basic qualities of players – based on FIFA 2014 and 2015 programs – corrected with their actual form.

The optimal version of the second multivariate regression model, created by backward elimination, consists of 3 explanatory variables. The multiple correlation coefficient for this second model is 0.678. The value of the coefficient of multiple determination is 0.459 which means that the 3 variables together explain the number of the scored goals in 45.9%.

Table 7 Explanatory power of the second regression model

Model Summary				
Model	R	R Square ^b	Adjusted R Square	Std. Error of the Estimate
1	.678 ^a	.459	.440	1.43936

Source: Self-calculated.

We can see in Table 8 that if the static abilities of the attackers corrected with the actual form are better with 1 percentage point than the similar value of the opponent defenders, then the team will score average by 0.033 goals more. At such a deviation the difference is not important, but if the difference is 20 percentage points, that means 0.66 goal already. The t-value of the explanatory variable is 3.337, which is significant at any level of significance used in practice.

The second variable is the difference between the InStat values of attackers and the defenders of the opponent team. Its t-value is 4.653, which is also significant at any level of significance. Note, that InStat index has no top limit and this coefficient is based on a complicated statistical calculation.

Table 8 The second regression model

Model	Unstandardized Coefficients		t	Sig.
	B	Std. Error		
attacker_vs_defender_based_on_fifa_form	.033	.010	3.337	.001
own_attacker_vs_opponent_defender_InStat	.011	.002	4.653	.000
own_fifa_worldrank_vs_opponent_fifa_worldrank	.000	.000	.879	.381
own_homefield	.850	.227	3.754	.000

Source: Self-calculated.

The third variable in this model is home field, t-value is 3.754, which is significant at any level of significance, as well. We can claim, based on the calculation of the multivariate regression model, that if two teams with equal player parameters meet on home field or guest field, then the team playing on home field will score an average by 0.85 more goals, then playing on guest field. So it is not all the same, if a team plays a game on home field, or on guest field.

2.6. Testing the results of the models

In case of the first model, the number of scored goals was exactly predicted 39 times out of 116 cases, which is essentially a one-third rate (33.6%). From the outcomes of 58 matches our model predicted 33 right, which is a 56,9% rate.

In case of the second model the prediction of the number of scored goals was a little bit worse (36 right cases out of 116, which is 31.0%), however the rate of right predictions concerning the outcomes is significantly better (36 matches out of 58, which is 62.1%).

We think that taking a sport, which depends on countless unpredictable factors, like football, this can be considered a pretty good rate, which can predict the remaining qualifier matches of the 2016 European Championship with a relatively confident result.

2.7. Prediction based on the two models

We were curious, what kind of rank can the Hungarian national team achieve in the qualifier group of the European Championship. Therefore we predicted the remaining matches from our second model. We gathered information of the form of the players in their club teams for the time period of 7-8-9 November 2014 (closing date of the paper). The outcomes of the games can be seen in Figure 2 below. We marked the already played games with golden, and the outcome of the games estimated by our model with grey color.

	HUN	ROM	GRE	NIR	FIN
HUN	-	0-0	0-0	1-2	1-0
ROM	1-1	-	1-0	2-0	1-0
GRE	2-0	0-1	-	0-2	1-0
NIR	1-0	1-1	1-1	-	1-1
FIN	2-0	0-2	1-1	2-0	-

Already played game



Predicted outcome



Figure 2 Results of the played matches and predicted results of the second model

Source: Self-calculated.

At the moment of writing this paper, the national teams played 4 rounds. The Hungarian team with its 7 points was at the third place of the group, to which we added the points of the predicted matches.

According to the prediction of the second statistical model, based on the **form of players at november 2014** (as we can see on the figure below), the Romanian national team is going to get first place of the group with 24 points. Northern-Ireland will get second place, overtaking Greece. (Due to the lack of data concerning the players of Feroer Island we assumed that each team gets 6 points against them, except Greece which gets only 3 points since they lost the qualifier in October 2014.)

rank	team	match	win	draw	loss	scored	received	point
1.	ROM	10	7	3	0	11	2	24
2.	NIR	10	5	3	2	10	8	18
3.	GRE	10	4	2	4	8	7	14
4.	FIN	10	3	3	4	6	7	12
5.	HUN	10	3	3	4	5	8	12

Qualify for 2016 E.C. 

Post-qualifying round 

Figure 3 – Result of the “Hungarian group” based on the prediction of the second model and results so far

Source: Self-edited.

CONCLUSION

In our research we examined, whether the outcome of the matches between national teams – based on the form of the players they showed in club matches before the national matches – can be predicted or not. We gathered data about the current form of players with the InStat software. We distinguished six defending and five attacking type players in one team. We took five defending, and six attacking qualities into consideration. Using factor analysis, we classified the five, and six variables into three-three factors, than after defining the explanatory power of the factors, we explored the connections between the variables.

As next step, we created two, multivariate regression models. In both cases we showed that there is a significant connection between the form of players shown at their club matches, and the outcome of their national matches (which variable, introduced by ourself, was even more accurate then the variable of InStat index). The first multivariate regression model predicted the final outcome of the examined matches with a 56.9%, and the number of scored goals with a 33.6% accuracy. The second model predicted the outcomes with a 62.1% rate, and the number of scored goals with a 31% rate.

The main question of our paper was whether the Hungarian national team can make it to the 2016 European Championship? Considering the form of the players in November 2014, based on either the first or the second model, the Hungarian team will not make it to any of the first three places, which would mean going forward. Nevertheless, seeing the results, we can determine, how much the current form can influence the result of the matches. This way,

we can only trust to make it to the 2016 European Championship, if the players of the national team show a significantly better form in their clubs in spring 2015.

REFERENCES

ESPN: Soccer Power Index explained. <http://www.espnfc.com/fifa-world-cup/story/1873765/soccer-power-index-explained> (downloaded 2014.10.17.)

KAKAS Péter: Csak focizni nem tud helyettünk a mindent látó szoftver. <http://www.origo.hu/sport/magyarfoci/20131010-az-InStat-elemzorendszer-bemutatasa.html> (downloaded 2014.11.03.)

KAMINSKI M. M. (2014): How Strong are Soccer Teams? 'Host Paradox' in FIFA's ranking. <http://publicchoicesociety.org/content/papers/marekkaminski-845-2014-846.pdf> (downloaded 2014.10.17.)

LAGO-PENAS C. (2009): Consequences of a busy soccer match schedule on team performance: Empirical evidence from Spain http://www.ismj.com/files/311417173/ismj%20pdfs/Vol_10_no_2_2009/Consequences-busy-soccer-match-schedule-performance-final.pdf (downloaded 2014.10.17.)

CONTACTS

Bence JÁMBOR, MA
Budapest Business School
College of International Management
and Business
bencu23@gmail.com
Diósy L. str. 22-24, 1165 – Budapest

Máté JÁMBOR, MA
Budapest Business School
College of International Management
and Business
mate_jambor@hotmail.com
Diósy L. str. 22-24, 1165 – Budapest

Dávid SZABÓ, MA
Budapest Business School
College of International Management and Business
Diósy L. str. 22-24, 1165 – Budapest
szabodavid93@gmail.com