ssdnm
środowiskowe
studia doktoranckie
z nauk matematycznych

Damian Brzyski

Uniwersytet Jagielloński

The estimation of the overall number of points in the Polish football league

Praca semestralna nr 1

(semestr letni 2011/12)

Opiekun pracy: Jerzy Ombach

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The estimation of the overall number of points in the Polish football league

Damian Brzyski

Abstract

In this paper we shall discuss the issue of estimating the number of points gained at the end of the season by teams from the Polish football league. We shall present and compare two models based on various approaches and show how their appropriate mixture improves the prediction ability. We shall use mean absolute error to measure the models fit.

Keywords: Football (soccer), Statistical prediction, Time dependent model, Polish league

1. Introduction

Statistical Football prediction is a method used to predict the outcome of Association football matches by means of statistical tools. For this purpose, over the years, various approaches and types of models were tested. The issue is important especially in establishing the odds on all possible outcomes of a match by bookmakers. One of the papers written on the topic is the article by Mark J. Dixton and Stuart G. Coles [1]. The authors aim is to exploit potencial inefficiencies in the association footbal betting market. Following M. J. Maher [2] they were derived a model for football scores in a match between specific teams and used technique based on a Poisson regression. In contrast Hill [3] analysed football teams places in the final league table. In 1974 he applied a comparison test and showed significant positive correlation between final league placings and forecasts. Clarke and Norman [4] investigated the English league matches and used least squares method to conclude the adventage of playing at home. Another view on the issue was proposed in [5] by Fahrmeir and Tutz who used time-dependent ordered paired comparision systems. Their work is based on German football data.

The aim of this paper is to present and compare various author's approaches for estimating number of points gained by each team from Polish football top division at the end of the season. The paper is based on author's Master Thesis, [6],

defended in September 2011. All required data originates from [7] and Microsoft Office Excel 2007 was used for calculations. Evaluation of the final number of points by each of presented models consists in determination of the average number of points gained by the team in all matches. Firstly, in Section 3, we will discuss the *rank's model* which introduces the term *rank of team* and generates forecasts based on there values. The *weight model*, presented in Section 4 uses the exponential function to reflect time dependency that causes the newest matches to have the biggest impact on prognosis. Function parameter finding is described in Section 5 whereas the Section 6 is dedicated to the *mixed model* which is a combination of two previous approaches. The next chapter provides the results obtained by all considered models in seasons $2008/2009$, $2009/2010$ and $2010/2011$.

2. Preliminaries

The names of divisions in Polish football league system was changed by years for sponsorship reasons. In this paper we will always use the term „First League" with respect to the Polish top division and the term „Second League" if we refer to the second level of the Polish football league system[1]. Since the season 2005/2006, the First League consists of 16 teams. According to current regulations each team during the season plays 30 games, last two teams from the First League get relegated to the Second League and the first two teams from the Second League get promoted to the top division.

By S_i we will denote the set of teams which played in the season $i/i+1$ and by n_i its cardinality. Each team from the set S_i has been indexed alphabetically, thus we can write: $S_i = \{s_{ij} : j = 1, \ldots, n_i\}$. The position taken by the team s_{ij} in season $k/k+1$ in some division (not necessarily in First League) will be denoted as $POS_k(s_{ij})$.

In this paper we present different Methods for Estimating the Number of Points (MENP) gained by every First League's team at the end of the season. To describe the algorithms which we will use it is helpful to define the vector $NP_i := (np_{i,1}, \ldots, np_{i,n_i})$ denoting the actual number of points gained respectively by $s_{i,1}, \ldots, s_{i,n_1}$ at the end of the season $i/i+1$. Vector of the analogous values estimated by the given algorithm will be denoted by $\widehat{NP_i} := (\widehat{np_{i,1}}, \ldots, \widehat{np_{i,n_i}})$. We need a tool that lets us evaluate the fit of each model to the empirical data and

[1] at the moment this paper has been written top division in Poland is called "Ekstraklasa" and lower one is called First League

allows us to compare the results. For that purpose we will use the mean absolute error which in this case has interpretive meaning in contrast to commonly used the mean square error.

Definition 2.1. *For a given season $i/i+1$ and vector $\widehat{NP_i}$ constructed using the given MENP we consider the value*

$$FE := \frac{1}{n_i}\sum_{j=1}^{n_i} |np_{i,j} - \widehat{np_{i,j}}|$$

which we call the fit error.

We will use the FE value to evaluate each model in the last three complete seasons (at the time of writing this paper): 2008/2009, 2009/2010 and 2010/2011. It is obvious from the definition that the smaller the FE value, the better fit of the model to the data.

The models presented in this paper are based on estimating the probabilities of winning, drawing and losing each teams of S_i in matches played during the season $i/i+1$. Therefore we will need the following definitions.

Definition 2.2. *For teams s_{ij} and s_{il} from the set S_i and given MENP we will consider the 3-dimensional vector whose coordinates are respectively: the estimated probability that s_{ij} will win the match over s_{il}, estimated probability that s_{ij} will draw a match with s_{il} and estimated probability that s_{ij} will lose a match against s_{il}. We will call this vector the result's probability vector (s_{ij} with s_{il}) and denote by $\widehat{P^i_{jl}} := (\widehat{w^i_{jl}}, \widehat{d^i_{jl}}, \widehat{l^i_{jl}})$. Additionally for each $j \in \{1, \ldots, n_i\}$ we will define $\widehat{P^i_{jj}}$ as a 3-dimensional zero vector.*

We will identify the planned matches in the season $i/i+1$ with the elements of set $M_i := \{(s_{ij}, s_{il}) : j, l = 1, \ldots, n_i, j \neq k\}$ such that the pair $(s_{ij}, s_{il}) \in M_i$ will mean a planned match between teams s_{ij} and s_{il} in season $i/i+1$ in which s_{ij} will play at home. It is clear that this identification is unique due to fact that in every season there are two matches scheduled for each pair of teams at both stadiums. The number of points gained at the end of season by team s_{ij} is equal to the sum of points gained in matches played at home and the same number of matches played away. Therefore it was important only to predict the sum of points gained by team s_{ij} in two matches played with each of the teams from S_i. The resulting probability vectors for the pair (s_{ij}, s_{il}) and the pair (s_{il}, s_{ij}) were determined based on results of the same matches of those teams no matter where

they were played. This procedure has allowed us to not take into account the advantage of playing at home which greatly simplifies the models.

Now we define three functions that will help us to provide further formulas.

Definition 2.3. *We define functions* $W : M_i \to \{0,1\}$, $D : M_i \to \{0,1\}$ *and* $L : M_i \to \{0,1\}$ *as follows:*

$$W((s_{ij}, s_{il})) := \begin{cases} 1 & \text{, if } s_{ij} \text{ won over } s_{il} \text{ in season } i/i+1 \\ 0 & \text{, in other cases} \end{cases}$$

$$D((s_{ij}, s_{il})) := \begin{cases} 1 & \text{, if } s_{ij} \text{ drew with } s_{il} \text{ in season } i/i+1 \\ 0 & \text{, in other cases} \end{cases}$$

$$L((s_{ij}, s_{il})) := \begin{cases} 1 & \text{, if } s_{ij} \text{ lost from } s_{il} \text{ in season } i/i+1 \\ 0 & \text{, in other cases} \end{cases}$$

In the definitions above, the victory or the defeat granted by walkover are included in „other cases".

3. Rank's model

Rank's model consists in assigning each team s_{ij} (or almost each which will be described later) of set S_i a natural number intended to illustrate the potential team's position in set S_i at the start of the season $i/i+1$. For this purpose we will define some subset of S_i for which rank assignment will be considered.

Definition 3.1. *We will call the remaining rank teams (for the set* S_i*) the set* $REM_i := \{s_{ij} \in S_i : s_{ij} \in S_{i-1}$ *and* $POS_{i-1}(s_{ij}) \in \{1, \ldots, 14\}, j \in \{1, \ldots, n_i\}\}$ *while the set* $PROM_i := \{s_{ij} \in S_i : s_{ij} \notin S_{i-1}$ *and* $POS_{i-1}(s_{ij}) \in \{1, 2\}, j \in \{1, \ldots, n_i\}\}$ *will be called the promoted rank teams (for the set* S_i*). The union of sets* REM_i *and* $PROM_i$ *we will simply call the rank teams (for the set* S_i*) and denote by* $\overline{S_i}$*.*

In some situations the cardinality of set $\overline{S_i}$ could be less than 16. It is possible that some First League teams which have remained out of the relegation zone can possibly be demoted to a lower division for example due to corruption. It also happened that in the First League less than 14 football clubs remained because of the rules which assume a smaller number of teams in the top division or allowing additional falls/promotions due to the existence the play-off system.

Definition 3.2. *For any s_{ij} from the set $\overline{S_i}$ we determine the value $R_i(s_{ij}) \in \{1, \ldots, 16\}$ as follows:*

$$R_i(s_{ij}) := \begin{cases} POS_{i-1}(s_{ij}) & \text{, if } s_{ij} \in REM_i \\ 14 + POS_{i-1}(s_{ij}) & \text{, if } s_{ij} \in PROM_i \end{cases}$$

and call it the rank of team s_{ij}. Additionally we define:
$R_i(s_{ij}, s_{il}) := (R_i(s_{ij}), R_i(s_{il}))$.

Let's take teams s_{ij} and s_{il} from S_i. We assume that we know the results of all matches played since the season $i-h/i-h+1$ to season $i-1/i$, where h belong to natural numbers is a given range of data. Before we present formulas for the vector $\widehat{P_{jl}^i}$ coordinates, we will denote by A_{ij} the number of all matches played between teams with similar ranks as s_{ij} and s_{il} in the considered data range:

$$\begin{aligned} A_{ij} :=& \big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{ka}, s_{kb}) = R_i(s_{ij}, s_{il}), (W+D+L)((s_{ka}, s_{kb})) = 1\big\}\big| + \\ & \big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{kb}, s_{ka}) = R_i(s_{ij}, s_{il}), (W+D+L)((s_{ka}, s_{kb})) = 1\big\}\big|, \\ & \text{where } k \text{ takes each values from the set } \{i-h, \ldots, i-1\}. \end{aligned}$$

Now we define coordinates $\widehat{w_{jl}^i}$, $\widehat{d_{jl}^i}$ and $\widehat{l_{jl}^i}$ as follows:

$$\begin{aligned} \widehat{w_{jl}^i} := \frac{1}{A_{ij}}\Big(&\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{ka}, s_{kb}) = R_i(s_{ij}, s_{il}), W((s_{ka}, s_{kb})) = 1\big\}\big| + \\ &\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{kb}, s_{ka}) = R_i(s_{ij}, s_{il}), L((s_{ka}, s_{kb})) = 1\big\}\big|\Big), \end{aligned}$$

$$\begin{aligned} \widehat{d_{jl}^i} := \frac{1}{A_{ij}}\Big(&\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{ka}, s_{kb}) = R_i(s_{ij}, s_{il}), D((s_{ka}, s_{kb})) = 1\big\}\big| + \\ &\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{kb}, s_{ka}) = R_i(s_{ij}, s_{il}), D((s_{ka}, s_{kb})) = 1\big\}\big|\Big), \end{aligned}$$

$$\begin{aligned} \widehat{l_{jl}^i} := \frac{1}{A_{ij}}\Big(&\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{ka}, s_{kb}) = R_i(s_{ij}, s_{il}), L((s_{ka}, s_{kb})) = 1\big\}\big| + \\ &\big|\big\{(s_{ka}, s_{kb}) \in M_k : R_k(s_{kb}, s_{ka}) = R_i(s_{ij}, s_{il}), W((s_{ka}, s_{kb})) = 1\big\}\big|\Big), \end{aligned}$$

where k takes values from the set $\{i-h, \ldots, i-1\}$.

Resulting probability vectors designated in this way were used to estimate

the number of points gained by each First League's team at the end of the seasons $2008/2009$, $2009/2010$ and $2010/2011$ which were compared with empirical data using the FE value. The results (presented below in the paper) show the tendency of increase in the fit of the model with increasing the range of data.

4. Weight model

The algorithm described in this chapter uses only matches played between the teams s_{ij} and s_{il} (in contrast to rank's model) in order to estimate the vector $\widehat{P^i_{jl}}$. This is connected with smaller probability samples and therefore, data was used from matches played between these teams in Second League too (if such matches were played). It was considered that the balance between the desire to increase probability samples and to avoid taking into account obsolete data is to analyse match results from six previous seasons. Therefore we define the following auxiliary six dimensional vectors $WINS^i(jl) := (w^i(jl)_0, \cdots, w^i(jl)_5)$, $DRAWS^i(jl) := (d^i(jl)_0, \cdots, d^i(jl)_5)$, $LOSSES^i(jl) := (l^i(jl)_0, \cdots, l^i(jl)_5)$ and $MATCHES^i(jl) := (m^i(jl)_0, \cdots, m^i(jl)_5)$ which respectively means wins, draws, losses and all matches of team s_{ij} with team s_{il} in the First or Second League in seasons $i-6/i-5, \ldots, i-1/i$ respectively. In the above definitions we do not take into account walkovers so the coordinates of the vectors $WINS^i(jl)$, $DRAWS^i(jl)$, $LOSSES^i(jl)$ and $MATCHES^i(jl)$ can take all values from the set $\{0, 1, 2\}$. It is obvious that:

$$WINS^i(jl) + DRAWS^i(jl) + LOSSES^i(jl) = MATCHES^i(jl) \tag{1}$$

$$WINS^i(lj) = LOSSES^i(jl), \tag{2}$$

$$DRAWS^i(lj) = DRAWS^i(jl), \tag{3}$$

$$LOSSES^i(lj) = WINS^i(jl). \tag{4}$$

The exponential function was used to reflect the impact of time on the estimating vector $\widehat{P^i_{jl}}$. Coordinates $\widehat{w^i_{jl}}$, $\widehat{d^i_{jl}}$ and $\widehat{l^i_{jl}}$ (dependent on the parameter a) were estimated as follows:

$$\widehat{w^i_{jl}} := \sum_{k=0}^{5} e^{ak} w^i(jl)_k \left(\sum_{k=0}^{5} e^{ak} m^i(jl)_k \right)^{-1}, \tag{5}$$

$$\widehat{d^i_{jl}} := \sum_{k=0}^{5} e^{ak} d^i(jl)_k \left(\sum_{k=0}^{5} e^{ak} m^i(jl)_k \right)^{-1}, \tag{6}$$

$$\widehat{l^i_{jl}} := \sum_{k=0}^{5} e^{ak} l^i(jl)_k \left(\sum_{k=0}^{5} e^{ak} m^i(jl)_k \right)^{-1}. \tag{7}$$

If the parameter dependence of the vector $\widehat{P^i_{jl}}$ is important (for example when we consider the issue of parameter a estimation), there will be used marking $\widehat{P^i_{jl}}(a) = (\widehat{w^i_{jl}}(a), \widehat{d^i_{jl}}(a), \widehat{l^i_{jl}}(a))$. When parameter a is greater than zero, the expected property of a lesser impact of older matches results on the probability shall occurs.

Using the vector $\lambda_a := (1, e^a, e^{2a}, e^{3a}, e^{4a}, e^{5a})$ and standard scalar product, formulas (5), (6) and (7) can be expressed also in the form:

$$\widehat{w^i_{jl}} := \frac{WINS^i(jl) \cdot \lambda_a}{MATCHES^i(jl) \cdot \lambda_a}, \tag{8}$$

$$\widehat{d^i_{jl}} := \frac{DRAWS^i(jl) \cdot \lambda_a}{MATCHES^i(jl) \cdot \lambda_a}, \tag{9}$$

$$\widehat{l^i_{jl}} := \frac{LOSSES^i(jl) \cdot \lambda_a}{MATCHES^i(jl) \cdot \lambda_a}. \tag{10}$$

It is obvious from (8), (9), (10) and (1) that:

$$\widehat{w^i_{jl}} + \widehat{d^i_{jl}} + \widehat{l^i_{jl}} = 1,$$

hence the vector $\widehat{P^i_{jl}}$ is a probability vector. Moreover from (1)-(4) we have:

$$\widehat{P^i_{jl}} = (\widehat{l^i_{lj}}, \widehat{d^i_{lj}}, \widehat{w^i_{lj}}). \tag{11}$$

The above formulas make sense if $MATCHES^i(jl) \neq (0,0,0,0,0,0)$. If there is a situation that for some teams s_{ia} and s_{ib}, we have $MATCHES^i(ab) = (0,0,0,0,0,0)$, then we put $\widehat{P^i_{ab}} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Vectors $\widehat{P^i_{jl}}$ were used to estimate the number of points gained by every First League's team at the end of seasons $2008/2009$, $2009/2010$ and $2010/2011$ which were compared with empirical data using the FE value.

5. Weight model parameter estimation

The data from season $2010/2011$ has been used in order to estimate parameter a. Therefore for all formulas present in this section we will fix $i = 2010$. In estimation were used the results of matches teams s_{ij} and s_{il} from S_i that meet the following criteria (called *the selection criteria*):

- teams s_{ij} and s_{il} played together in the season $i-1/i$,
- in seasons $i-6/i-5,\ldots,i-1/i$ at least six times teams s_{ij} and s_{il} played matches in First or Second League.

The second condition means that (if there were no walkovers) during the six previous seasons teams s_{ij} and s_{il} played at least for three in the same division (First or Second League).

Definition 5.1. *The weight subset of M_i we will call set W_i defined as follows*

$$W_i := \{(s_{ij}, s_{il}) \in M_i \text{ where } s_{ij} \text{ and } s_{il} \text{ fulfill the selection criteria}\}.$$

It is obvious from this definition that:

$$(s_{ij}, s_{il}) \in W_i \Longleftrightarrow (s_{il}, s_{ij}) \in W_i. \tag{12}$$

Now for each team $s_{ij} \in S_i$ we can consider the forecasted (using the weight model) number of points gained in season $i/i+1$ only with the teams from S_i which together with s_{ij} fulfill the selection criteria. Denoting this value by $\widehat{\xi_j^i}(a)$, using vector $\rho := (3,1,0)$ and standard scalar product we can write following formula:

$$\widehat{\xi_j^i}(a) = \sum_{l:(s_{ij},s_{il})\in W_i} \rho \cdot \widehat{P_{jl}^i}(a) + \sum_{l:(s_{il},s_{ij})\in W_i} \rho \cdot (\widehat{l_{lj}^i}(a), \widehat{d_{lj}^i}(a), \widehat{w_{lj}^i}(a)).$$

From (11) and (12) we simply get:

$$\widehat{\xi_j^i}(a) = 2 \sum_{l:(s_{ij},s_{il})\in W_i} \rho \cdot \widehat{P_{jl}^i}(a). \tag{13}$$

Now if we denote by ξ_j^i the actual numbers of points gained in season $i/i+1$ by s_{ij} only with the teams from S_i which together with s_{ij} fulfill the selection criteria, we can consider the function (dependent on parameter a) which gives the mean absolute error of predicted values:

$$\psi(a) := \frac{1}{n_i} \sum_{j=1}^{n_i} \left| \xi_j^i - \widehat{\xi_j^i}(a) \right|.$$

The Solver (an add-in for Microsoft Excel) was used for finding parameter a realizing the global minimum of function ψ which means that the model's fit to the empirical data is the best. This parameter (called *the weight parameter*) was used in earlier presented calculations connected with the weight model. Values of estimated parameter a and $\psi(a)$ are shown in the table below.

Table 1: Estimating of the weight parameter

Outcome	a	$\psi(a)$
Value	0.17	3.17

6. mixed model

The disadvantage of the *weight model* presented in the previous sections is in many cases a small number of matches (even if we take into account games played in Second League) used to estimate a resulting probability vectors. On the other hand omission of the conclusions from previous common matches of two given teams (the rank's model) could be associated with less effective predictions. These observations underlie the idea of combine MENPs presented in the paper.

Established for that purpose the *mixed model* uses *the selection criteria* to split the set M_i as follows:

$$M_i = W_i \cup (M_i \backslash W_i)$$

The resulting probability vector for teams s_{ij} and s_{il} was created based on the weight model if teams s_{ij} and s_{il} meet the selection criteria ($(s_{ij}, s_{il}) \in W_i$) or with using the rank's model otherwise ($(s_{ij}, s_{il}) \in M_i \backslash W_i$).

Table 2 shows the size of the *weight subset* and its percentage in the whole set M_i (denoted by PER_i) in each of the three considered seasons.

Table 2: Weight subsets

Season's index i	$\lvert M_i \rvert$	$\lvert W_i \rvert$	PER_i
2008	240	67	27.92%
2009	240	108	45.00%
2010	240	144	60.00%

Results obtained by the *mixed model* for the seasons $2008/2009$, $2009/2010$ and $2010/2011$ were compared with results given by other MENPs.

7. Results

For each MENP coordinates of the vector $\widehat{NP_i}$, given by a prognosed number of points gained by every First League's team at the end of season $i/i+1$, were

created similarly as in the formula 13:

$$\widehat{np_{i,j}} := 2\sum_{l=1}^{n_i} \rho \cdot \widehat{P_{jl}^i}$$

Presented below are the results given by the rank's model (RM), the weight model (WM) and the mixed model (MM) against the actual number of points (ANP) gained by each football team taking part in the First League competition in seasons from $2008/2009$ to $2010/2011$.

Table 3: Results for season $2008/2009$

Team from S_{2008}	ANP	Results:		
		RM	WM	MM
Arka Gdynia	30	34.42	41.84	34.02
Cracovia	30	40.19	46.10	41.68
GKS Bełchatów	54	40.07	46.72	42.48
Górnik Zabrze	29	41.14	39.31	37.82
Jagiellonia Białystok	34	33.07	33.38	31.06
Lech Poznań	59	44.74	45.58	45.43
Lechia Gdańsk	32	37.10	37.58	36.87
Legia Warszawa	61	55.14	57.21	55.95
Łódzki KS	35	32.87	35.96	31.98
Odra Wodzisław Śląski	32	37.10	34.62	36.33
Piast Gliwice	33	34.42	36.65	36.10
Polonia Bytom	35	36.59	30.90	33.99
Polonia Warszawa	54	50.60	31.87	50.60
Ruch Chorzów	34	37.71	36.41	38.26
Śląsk Wrocław	45	37.10	39.26	37.67
Wisła Kraków	64	57.70	63.18	61.22
FE value	**0**	**6.10**	**6.96**	**5.73**

Table 4: Results for season $2009/2010$

Team from S_{2009}	ANP	Results:		
		RM	WM	MM
Arka Gdynia	28	36.73	28.35	36.73

Team from S_{2009}	ANP	Results:		
		RM	WM	MM
Cracovia	34	31.12	38.15	38.40
GKS Bełchatów	48	45.10	47.04	47.34
Jagiellonia Białystok	34	40.16	31.43	37.28
Korona Kielce	37	34.32	48.07	33.08
Lech Poznań	65	50.22	51.85	45.44
Lechia Gdańsk	37	32.87	29.55	32.75
Legia Warszawa	52	52.44	62.16	56.09
Odra Wodzisław Śląski	27	34.83	35.41	33.69
Piast Gliwice	27	35.56	25.74	35.47
Polonia Bytom	37	37.83	35.93	38.45
Polonia Warszawa	33	44.40	37.08	37.37
Ruch Chorzów	53	40.50	33.29	38.47
Śląsk Wrocław	36	45.80	41.84	45.17
Wisła Kraków	62	56.21	64.35	60.53
Zagłębie Lubin	35	35.73	52.24	37.28
FE value	**0**	**6.26**	**6.86**	**6.08**

Table 5: Results for season 2010/2011

Team from S_{2010}	ANP	Results:		
		RM	WM	MM
Arka Gdynia	28	31.19	30.33	29.14
Cracovia	29	34.46	34.53	32.09
GKS Bełchatów	40	45.25	48.90	43.18
Górnik Zabrze	45	34.35	26.10	32.97
Jagiellonia Białystok	48	33.22	34.62	35.22
Korona Kielce	37	45.39	45.46	40.94
Lech Poznań	45	56.56	56.32	53.60
Lechia Gdańsk	43	40.35	32.90	40.68
Legia Warszawa	49	43.91	60.00	52.78
Polonia Bytom	27	37.81	35.96	34.41
Polonia Warszawa	44	36.19	37.11	36.07
Ruch Chorzów	38	50.80	36.15	42.26
Śląsk Wrocław	49	41.36	41.38	40.58
Widzew Łódź	43	35.73	28.70	37.54
Wisła Kraków	56	52.32	66.43	61.35

Team from S_{2010}	ANP	Results:		
		RM	WM	MM
Zagłębie Lubin	39	35.06	43.20	42.69
FE value	**0**	**7.56**	**9.01**	**5.84**

In the table below we present the weights of matches played in each of six previous seasons (expressed as a percentage) given by the exponential function with the estimated parameter a. These values determine the impact of each of the seasons on the results.

Table 6: The impact of each of seasons on results

Season's index	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$
Impact on reaults	10.54%	12.46%	14.73%	17.40%	20.56%	24.30%

The fit of rank's model (measured by FE values) for various ranges of data are shown in Table 7.

Table 7: Fit of rank's model for various ranges

Index of season:	FE for a range beginning at season:				
	1990/91	1993/94	1996/97	1999/00	2002/03
2008	6.10	6.12	6.10	6.16	6.92
2009	6.26	5.91	5.95	5.98	5.76
2010	7.56	8.05	8.32	8.93	9.53

8. Conclusions

For all considered seasons the best fit was obtained using the *mixed model*. This may means that *the selection criteria* have been constructed properly and the idea of division games based on the common history of playing teams make sense and can improve the predictions. The highest growth of fit was achieved in the season $2010/2011$ (Table 7). Using the *mixed model* for this season caused a reduction in the FE value by 35.18% and 22.75% comparing with the *weight model* and the *rank's model* respectively. The worst predictions in all three seasons were

obtained by the *weight model*. This shows that application of this approach for the designation of result's probability vectors of teams, that has played an insufficient number of common matches or which did not play together for a long time, is connected with an increase in *the fit error*.

The results presented in Table 7 show that there is a trend of decreasing in the fit error with increasing the range of data in the rank's model which is especially visible for years $2010/2011$. The exception is the season $2009/2010$ where the best fit was reached for the shortest range of data but the attention should be paid to the fact that the results of fitting for this season are also the most stable of all (the difference between the best and the worst fit is only 0.5). Based on this analysis, data from the maximum range was included in the rank's model.

FE values for considered seasons are largely increased by a small number of teams that have reached results substantially differing than expected. It is enough to ignore two teams with the worst forecasts, given by the *mixed model*, to get the mean absolute error of the rest of the 14 teams smaller than 5 for each three seasons, which seems to be satisfying.

9. References

[1] M. J. Dixont, S. G. Coles (1997), *Modelling Association Football Scores and Inefficiencies in the Football Betting Market*, Applied Statistics, Volume 46, Issue 2 , 265-280.

[2] M. J. Maher (1982), *Modelling association football scores*, Statistica Neerlandica, nr. 3.

[3] I. D. Hill (1974), *Association football and statistical inference*, Applied statistics Vol. 23, No. 2, pp. 203-208.

[4] S. R. Clarke, J. M. Norman (1995), *Home ground advantage of individual clubs in English soccer*, The Statistician Vol. 44, No. 4 , pp. 509-521.

[5] L. Fahrmeir, G. Tutz (1994), *Dynamic-stochastic models for time-dependent ordered paired comparison systems*, Journal of the American Statistical Association Vol. 89, No. 428, pp. 1438-1449.

[6] D. Brzyski (2011), *The usage of the Markov Chain theory to model the total scores of teams from the Polish Football League*, Master Thesis, Jagiellonian University, Kraków.

[7] Match results - Polish leagues, *www.90minut.pl*