ZÁPADOČESKÁ UNIVERZITA V PLZNI

FAKULTA APLIKOVANÝCH VĚD

KATEDRA MATEMATIKY

Bakalářská práce

Modelování a odhadování výsledků sportovních utkání

Plzeň, 2015 Jan Špaček

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím literatury a pramenů uvedených v seznamu.

V Plzni dne 21. května 2015 .....

Jan Špaček

Poděkování

Rád bych poděkoval vedoucímu mé bakalářské práce Ing. Patrice Markovi, Ph.D. za cenné rady a čas, který mi věnoval při konzultacích.

Abstrakt

Tato bakalářská práce se zabývá odhadováním a modelováním výsledků sportovních zápasů a následným využitím odhadů při sázení v sázkových kancelářích.

Práce popisuje základní modely používané pro odhadování výsledku fotbalového utkání, které končí buď výhrou domácího týmu, remízou, anebo výhrou hostujícího mužstva. Práce se více věnuje modelu od M. J. Dixona a S. G. Colese z roku 1997. Na základě tohoto modelu jsou odhadovány výsledky zápasů anglické, české, italské a španělské ligy v sezóně 2013/2014. Dále jsou modely ověřovány při sázení proti sázkovým kancelářím.

Klíčová slova: Poissonovo rozdělení**,** odhad sportovních výsledků, sport, sázení

Abstract

This bachelor thesis is focused on estimating and modeling results of sports matches and aftewards the application of estimates for betting in bookmakers.

The thesis describes the basic models used to estimate the result of a football game that can end with a home team win, a draw or a visiting team win. The work is based on a model by M. J. Dixon and S. G. Coles from the year 1997 and more. Based on this model results of matches English, Czech, Italian and Spanish leagues in the season 2013/2014 are estimated. Furthermore, the models are verified in betting against betting companies.

Key words: Poisson distribution, estimate of sport results, sport, betting

Obsah

1 Úvod 1

2 Pravděpodobnost a statistika 2

2.1 Poissonovo rozdělení 2

2.2 Chí-kvadrát test dobré shody 2

2.3 p hodnota 3

2.4 Bonferroniho korekce 3

3 Testování počet gólů týmu se řídí Poissonovým rozdělením 4

4 Maherovy modely 7

4.1.1 Model 0 7

4.1.2 Model 1 7

4.1.3 Model 2 8

4.1.4 Model 3 8

4.1.5 Model 4 8

4.2 Zkoumaný model 8

4.3 Sezóna 2013/14 Gambrinus liga 9

4.3.1 Parametry k^2, α a β 9

4.3.2 Ukázka užití výsledků 11

4.3.3 Chí kvadrát test 12

4.3.4 Závěr 13

5 Dixon - Colesův model 14

5.1 Popis modelu Dixon - Coles 14

5.1.1 Sdružená pravděpodobnostní funkce 14

5.1.2 Parametry λ, μ 14

5.1.3 Funkce závislosti τ 15

5.2 Způsob odhadu parametrů 16

5.2.1 Věrohodnostní funkce 16

5.2.2 Logaritmická věrohodnostní funkce 16

5.2.3 Funkce času ϕ 16

5.3 Data 17

5.4 Gambrinus liga 17
5.4.1 Odhad parametrů Gambrinus liga 18
5.4.2 Odhad výsledků zápasů 22
5.5 Další ligy 23
5.5.1 Odhady parametrů 24
6 Sázení 25
6.1 Základní pojmy 25
6.2 Systém sázení 26
6.2.1 Flat betting 26
6.2.2 X procent na kolo 27
6.3 Kurzy 27
7 Ověření modelu 28
7.1 Česká liga, ... 28
7.1.1 Strategie Flat betting 28
7.1.2 Strategie X procent na kolo 32
7.1.3 Srovnání strategií Flat betting a X procent na kolo 33
7.2 Ostatní ligy 33
7.2.1 Španělská liga 34
7.2.2 Italská liga 35
7.2.3 Anglická liga 36
7.3 Shrnutí 36
8 Závěr 37
9 Literatura a zdroje dat 38
9.1 Seznam literatury 38
9.2 Zdroj dat 38

Seznam Obrázků

Obrázek 1: Skutečný a očekávaný počet gólů vstřelený mužstvem FC Viktoria Plzeň 6
Obrázek 2: Ukázka nastavení v Microsoft Excel před první iterací 10
Obrázek 3: Funkce času 17
Obrázek 4: Nastavení řešitele Microsoft Excel 18
Obrázek 5: Odhad parametrů v Microsoft Excel 19
Obrázek 6: Vývoj parametru α u týmů FC Viktoria Plzeň a AC Sparta Praha 20
Obrázek 7: Vývoj parametru β u týmů FC Viktoria Plzeň a AC Sparta Praha 21
Obrázek 8: Vývoj parametru γ 21
Obrázek 9: Vývoj parametru ρ 22
Obrázek 10: Vývoj zisku po jednotlivých kolech pro R = 1,2 30
Obrázek 11: Vsazené a vyhrané částky pro R = 1,2 30
Obrázek 12: Závislost zisku na parametru R 31

Seznam tabulek

Tabulka 1: Četnost gólů týmu FC Viktoria Plzeň 5
Tabulka 2: Očekávané pravděpodobnosti a hodnoty počtu gólů 5
Tabulka 3: Skutečný a očekávaný počet gólů 5
Tabulka 4: Výsledky testů a p-hodnoty 6
Tabulka 5: Odhadování parametrů α a β 11
Tabulka 6: Pravděpodobnost výsledků v zápase Plzeň - Brno 12
Tabulka 7: Skutečný a očekávaný počet gólů v domácích zápasech 12
Tabulka 8: Skutečný a očekávaný počet gólů ve venkovních zápasech 13
Tabulka 9: Odhad parametrů α a β pro 30. kolo (tj. z výsledků do 29. kola včetně) 20
Tabulka 10: Odhadnuté parametry pro zápas FC Baník Ostrava - SK Slavia Praha 22
Tabulka 11: Pravděpodobnost výsledků v zápase Baník Ostrava - Slavia Praha 23
Tabulka 12: Pravděpodobnost výhry domácích, remízy, výhry hostů 23
Tabulka 13: Kurzy na zápas mezi týmy A a B 26
Tabulka 14: Shrnutí vkladů a výplat v případě ideálního rozložení sázek 26
Tabulka 15: Shrnutí vkladů a výplat v případě jiného rozložení sázek 26
Tabulka 16: Seznam vsazených zápasů pro R = 1,2 29
Tabulka 17: Porovnání parametru R 31
Tabulka 18: Zisk v závislosti na R a procentech 32
Tabulka 19: Porovnání parametru R pro strategii 5 % 33
Tabulka 20: Porovnání parametru R španělská liga 34
Tabulka 21: Porovnání parametru R italská liga 35
Tabulka 22: Porovnání parametru R anglická liga 36

1 Úvod

Sportu se věnují lidé po celém světě. Někteří lidé se sportem živí, jiní se sportu aktivně věnují ve volném čase a někteří se chodí dívat na sportovní utkání na stadiony či je sledují v televizi. Možnost, jak se ještě více vžít do zápasu, je kromě fandění také sázení. Vsadit se mohou 2 lidé či více mezi sebou, anebo je možné si vsadit v sázkové kanceláři. Pro efektivnější sázení je dobré znát pravděpodobnosti výhry jednotlivých týmů. Cílem této bakalářské práce je pomocí matematických a statistických modelů tyto pravděpodobnosti odhadnout a následně použít modely proti sázkové kanceláři.

Druhá kapitola se věnuje definici statistických pojmů a metodám, které jsou následně použity v dalších kapitolách. Jsou zde popsány: Poissonovo rozdělení, chí-kvadrát test dobré shody, *p* hodnota a Bonferroniho korekce.

Ve třetí kapitole se zkoumá, zda se počet gólů vstřelených týmy řídí Poissonovým rozděleným pravděpodobnosti.

Čtvrtá kapitola se věnuje modelům M. J. Mahera, které popisuje ve svém článku [1]. Jsou zde popsány jednotlivé modely. Dále jsou pomocí jednoho z těchto modelů předpovídány výsledky zápasů a je proveden chí kvadrát test pro kontrolu těchto výsledků.

V páté kapitole je ukázán nový model od Dixona a Colese [2], který je vylepšením předchozího Maherova modelu. V této kapitole jsou popsána data použitá k odhadu a následně celý postup odhadování a předpovídání výsledků.

Šestá kapitola se věnuje základním sázkařským pojmům a strategiím. Dále je zde popsán výběr sázkových kanceláří.

V sedmé kapitole je ověření modelu z páté kapitoly proti sázkovým kancelářím. Ověřují se zde výsledky modelu ze čtyř lig v sezóně 2013/2014. Konkrétně se jedná o českou, španělskou, italskou a anglickou ligu.

V osmé kapitole je závěrečné zhodnocení práce a shrnutí výsledků.

2 Pravděpodobnost a statistika

V této kapitole jsou popsány pojmy z pravděpodobnosti a statistiky, které jsou použity v dalších kapitolách.

2.1 Poissonovo rozdělení

Poissonovo rozdělení pravděpodobnosti náhodné veličiny je diskrétní rozdělení pravděpodobnosti s parametrem λ. Je označováno $Po(\lambda)$.

Pravděpodobnostní funkce Poissonova rozdělení je

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \ \ pro\ k = 0,1,2, \dots \tag{2.1}$$

Střední hodnota a rozptyl u Poissonova rozdělení jsou stejné a ve tvaru

$$E(x) = \lambda, \tag{2.2}$$

$$D(x) = \lambda. \tag{2.3}$$

Více o tomto rozdělení lze nalézt v knize Elementární statistická analýza [3].

2.2 Chí-kvadrát test dobré shody

V této části je čerpáno z knihy Metody matematické statistiky [4].

Je k dispozici náhodný výběr rozsahu n z náhodné veličiny X. Na hladině významnosti α se testuje hypotéza, že náhodná veličina X má nějaké pravděpodobnostní rozdělení, které je známé až na hodnotu m neznámých parametrů (může být i $m = 0$, pak jsou známy všechny parametry).

Postup testování:

Obor hodnot se rozdělí do k $tříd$ a zjistí se, kolik hodnot realizovaného náhodného výběru se nachází v jednotlivých třídách, tyto počty se označí n_i. Poté se odhadnou neznámé parametry m. Pro každou třídu se spočte očekávaný počet hodnot o_i v této třídě

$$o_i = n \cdot p_i \quad pro\ i = 1,2, \dots, k, \tag{2.4}$$

kde je

n rozsah náhodného výběru,

p_i pravděpodobnost, že X s předpokládaným rozdělením pravděpodobnosti nabude hodnoty pařící do i-té třídy.

Je-li některý očekávaný počet o_i menší než 5 (ne vždy se dodržuje, zvlášť pro málo dat, ale vždy musí platit, že o_i je větší než 1), sdruží se tato třída s některou jinou. Toto se opakuje, dokud není splněno pro každou třídu o_i větší než pět. Počet nových tříd se opět označí k. Hypotéza, že veličina X se řídí předpokládaným rozdělením, se zamítne, je-li

$$\sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} > \chi^2_{1-\alpha}(\nu), \tag{2.5}$$

kde je

$\chi^2_{1-\alpha}(\nu)$ kvantil χ^2 rozdělení,

ν počet stupňů volnosti $\nu = k - 1 - m \ (\nu > 0)$.

2.3 p hodnota

Definice p hodnoty je přebrána ze zdroje [5], kde je uvedeno, že „$P\ hodnota$ testu je u testů, kde má tato definice smysl, pravděpodobnost, s jakou testovací statistika nabývá hodnot „horších“ (více svědčících proti testované hypotéze), než je pozorovaná hodnota statistiky. $P\ hodnota$ je obvyklým výstupem počítačových programů na testování hypotéz, udává mezní hladinu významnosti, při které by hypotéza ještě byla zamítnuta. Hypotéza H_0 je zamítnuta na hladině α, právě tehdy, když $p\ hodnota$ je menší než α.“

2.4 Bonferroniho korekce

Ve statistických testech dojde k zamítnutí hypotézy H_0 v případě, že pravděpodobnost pozorovaných dat za platnosti hypotézy H_0 je malá. Problém nastává při testování složeného testu, tím jak se zvýší počet hypotéz v testu, tak dojde i ke zvýšení této pravděpodobnosti a tím dojde ke zvýšení možnosti zamítnutí H_0 za předpokladu, že H_0 platí tedy k chybě prvního druhu. Proto při testování složených hypotéz je třeba upravit hladinu významnosti α kvůli korekci chyby prvního druhu. K úpravě hladiny významnosti se používá Bonferroniho korekce

$$\alpha^* = \frac{\alpha}{m}, \tag{2.6}$$

kde je

α^* korigovaná hladina významnosti,

α původní hladina významnosti,

m počet provedených testů.

Více informací o Bonferroniho korekci lze nalézt v [6].

3 Testování počet gólů týmu se řídí Poissonovým rozdělením

Binomické rozdělení pravděpodobnosti modeluje počet příznivých výsledků z n pokusů. V jednom fotbalovém zápase dochází k velkému počtu útoků, ale jen málo z nich je úspěšných a skončí gólem. V takovém případě, kdy je vysoký počet opakování s malou pravděpodobností úspěchu jednoho pokusu, lze binomické rozdělení pravděpodobnosti aproximovat Poissonovým rozdělením, což ukázal v roce 1982 M. J. Maher ve svém článku [1] na výsledcích zápasů anglických lig.

V této kapitole se bude testovat, že Poissonovým rozdělením pravděpodobnosti se řídí i počty gólů vstřelených jednotlivými týmy v české nejvyšší fotbalové soutěži v té době Gambrinus lize (dnes Synot liga). Jako data poslouží výsledky zápasů ze sezón 2009/2010 až 2013/2014 [A]. Během tohoto období hrálo nejvyšší soutěž alespoň jednu sezónu 22 týmu. Jelikož týmy se doma před svými příznivci snaží více útočit a obvykle střílejí více branek než při venkovních zápasech, tak je zvlášť zkoumán počet branek, které mužstvo vsítilo doma a zvlášť venku.

Testování je prováděno pomocí χ^2 testu dobré shody, popsaného v kapitole 2.2. Tento test je proveden pro všechny týmy z ligy, které odehrály během sledovaného období alespoň 2 sezóny v nejvyšší soutěži, což splnilo 17 týmů a naopak nesplnilo 5 týmů: FK Bohemians Praha (Střížkov), SK Kladno, FK Ústí nad Labem, FK Viktoria Žižkov a 1.SC Znojmo FK. Tato podmínka je z důvodu, že pro mužstva, která odehrála jen jednu sezónu, je k dispozici malé množství dat, konkrétně 15 zápasů doma a 15 zápasů venku. To by mohlo vést ke zkresleným výsledkům.

Testovat se tedy budou na hladině významnosti $\alpha = 5\ \%$ dvě složené hypotézy H_0.

H_0: Počet gólů vstřelených týmy v domácích zápasech se řídí Poissonovým rozdělením pravděpodobnosti.
H_1: Počet gólů vstřelených týmy v domácích zápasech se neřídí Poissonovým rozdělením pravděpodobnosti.

H_0: Počet gólů vstřelených týmy při venkovních zápasech se řídí Poissonovým rozdělením.
H_1: Počet gólů vstřelených týmy při venkovních zápasech se neřídí Poissonovým rozdělením.

Tyto testy jsou v sešitu Poisson.xlsx v listu Poisson. Obě složené hypotézy se skládají ze 17 jednotlivých hypotéz (pro každé mužstvo jedna hypotéza).

Pro ukázku je zde uvedena část testu hypotéza H_0: počet gólů vstřelených mužstvem FC Viktoria Plzeň v jednotlivých domácích ligových utkáních se řídí Poissonovým rozdělením a alternativní hypotéza H_1: počet gólů vstřelených týmem FC Viktoria Plzeň v jednotlivých domácích ligových utkáních se neřídí Poissonovým rozdělením pravděpodobnosti.

Aby byla zachována v testu hladina významnosti α, je pro tuto hypotézu H_0 použita Bonferroniho korekce (kapitola 2.4). Vzhledem k tomu, že složená hypotéza se skládá ze 17 jednotlivých hypotéz, tak upravená hladina významnosti α^* pro tuto hypotézu je 5/17 %. Analogicky je proveden stejný postup testování hypotéz pro ostatní mužstva.

V dalším kroku jsou určeny jednotlivé třídy, což je v tomto případě počet gólů v zápase 0, 1, … a zároveň je určen počet pozorování, které padnou do jednotlivých tříd a celkový počet pozorování n. Plzeň ve sledovaném období byla v první lize ve všech pěti ročnících, každou sezónu hrála 15 zápasů doma a tedy celkově během pěti let odehrála n = 75 domácích zápasů. Osmkrát v těchto zápasech nedala ani jeden gól, devatenáctkrát vstřelila jednu branku atd. Všechna pozorování počtu gólů týmů FC Viktoria Plzeň jsou v následující tabulce.

Počet gólů x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Četnosti n_i	8	19	19	14	9	3	2	1

Tabulka 1: Četnost gólů týmu FC Viktoria Plzeň

Dále je třeba odhadnout parametr λ, což je střední hodnota. Tu lze odhadnut pomocí průměru [3]. Tj. λ se rovná aritmetickému průměru gólů za domácí zápas vstřelených týmem FC Viktoria Plzeň. Pro FC Viktoria Plzeň je $\hat{\lambda}$ = 2,25.

Nyní už je možné vypočítat očekávané hodnoty. Nejdříve se určí pravděpodobnosti p_i pomocí vzorce (2.1), že v jednom zápase (pozorování) vstřelí Viktoria 0 gólů, 1 gól atd. Podle vzorce (2.5) se pak dopočítají očekávané hodnoty.

Počet gólů x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7 a více**
Pravděpodobnost p_i	0,11	0,24	0,27	0,20	0,11	0,05	0,02	0,01
Očekávaná četnost o_i	7,88	17,75	20,00	15,02	8,46	3,81	1,43	0,63

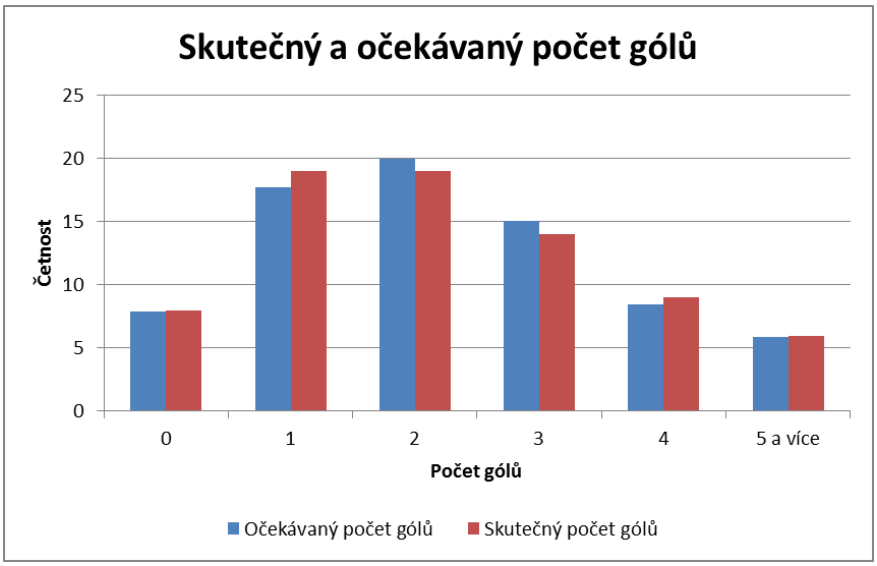
Tabulka 2: Očekávané pravděpodobnosti a hodnoty počtu gólů

V následujícím kroku je nutné spojit třídy, aby v každé byly očekáváné četnosti větší než 5. V tomto případě se spojí skupiny 5, 6 a 7 a více a vznikne jedna skupina 5 a více.

Dále se spočte testové kritérium pomocí vzorce (2.6), na základě kterého se rozhodne, zda se přijme či zamítne hypotéza H_0.

Počet gólů x	**0**	**1**	**2**	**3**	**4**	**5 a více**
Skutečná četnost n_i	8	19	19	14	9	6
Očekávaná četnost o_i	7,88	17,75	20,00	15,02	8,46	5,88

Tabulka 3: Skutečný a očekávaný počet gólů



Obrázek 1: Skutečný a očekávaný počet gólů vstřelený mužstvem FC Viktoria Plzeň

Testové kritérium vyjde v tomto příkladu 0,25. Ještě se určí kvantil $\chi^2_{1-\alpha^*}(v)$ o 4 stupních volnosti (6 – 1 – 1), který je 16,06. 0,25 je menší než 16,06, z toho vyplývá, že se hypotéza H_0 nezamítá. Na závěr je určena ještě p hodnota, která je v tomto případě 0,99.

Obdobně se otestují jednotlivé hypotézy pro všechny týmy při domácích i venkovních zápasech.

Tým	Test doma	P hodnota doma	Test venku	P hodnota venku
Bohemians	Nezamítáme H0	0,935	Nezamítáme H0	0,675
Brno	Nezamítáme H0	0,419	Nezamítáme H0	0,510
České Budějovice	Nezamítáme H0	0,401	Nezamítáme H0	0,257
Dukla	Nezamítáme H0	0,621	Nezamítáme H0	0,388
Hradec	Nezamítáme H0	0,162	Nezamítáme H0	0,038
Jablonec	Nezamítáme H0	0,843	Nezamítáme H0	0,407
Jihlava	Nezamítáme H0	0,379	Nezamítáme H0	0,991
Liberec	Nezamítáme H0	0,386	Nezamítáme H0	0,152
Mladá Boleslav	Nezamítáme H0	0,616	Nezamítáme H0	0,790
Olomouc	Nezamítáme H0	0,705	Nezamítáme H0	0,259
Ostrava	Nezamítáme H0	0,773	Nezamítáme H0	0,790
Plzeň	Nezamítáme H0	0,993	Nezamítáme H0	0,099
Příbram	Nezamítáme H0	0,159	Nezamítáme H0	0,548
Slavia	Nezamítáme H0	0,004	Nezamítáme H0	0,042
Slovácko	Nezamítáme H0	0,500	Nezamítáme H0	0,988
Sparta	Nezamítáme H0	0,624	Nezamítáme H0	0,790
Teplice	Nezamítáme H0	0,531	Nezamítáme H0	0,451

Tabulka 4: Výsledky testů a p hodnoty

Z tabulky vyplývá, že u všech týmů doma i venku není zamítnuta hypotéza H_0, a tak nejsou zamítnuty složené hypotézy H_0: počet gólů vstřelených týmy v domácích respektive venkovních zápasech se řídí Poissonovým rozdělením pravděpodobnosti.

4 Maherovy modely

M. J. Maher ve svém článku [1] popsal několik modelů pro odhad výsledků fotbalových utkání využívajících Poissonovo rozdělení. Ve všech následujících modelech se tedy očekává, že počet gólů X_{ij} vstřelených domácím týmem v zápase se řídí Poissonovým rozdělením s parametrem λ_{ij} a počet gólů Y_{ij} vstřelených hostujícím týmem v zápase se řídí Poissonovým rozdělením s parametrem μ_{ij}, což bylo ověřeno v minulé kapitole.

Parametr λ je vyjádřen následujícím vzorcem

$$\lambda_{ij} = \alpha_i \cdot \beta_j, \tag{4.1}$$

kde je

α_i síla domácího týmu v útoku,
β_j síla hostujícího týmu v obraně.

Parametr μ je vyjádřen následujícím vzorcem

$$\mu_{ij} = \gamma_i \cdot \delta_j, \tag{4.2}$$

kde je

γ_i síla domácího týmu v obraně,
δ_j síla hostujícího týmu v útoku.

Jednotlivé modely se liší výpočtem parametrů α, β, γ a δ.

4.1.1 Model 0

V tomto modelu se předpokládá, že všechny týmy jsou stejně silné. Platí tedy $\alpha_i = \alpha$, $\beta_i = \beta$, $\gamma_i = \gamma$ a $\delta_i = \delta$ pro všechna i. Výhodou tohoto modelu je, že je třeba znát pouze čtyři parametry a vzhledem k tomu, že počet branek vstřelených domácími týmy musí být stejný jako počet branek obdržených hostujícími týmy, platí $\alpha = \beta$ (analogicky $\gamma = \delta$). A z toho vyplývá, že pro tento model stačí odhadnout pouze dva nezávislé parametry. Nevýhodou tohoto modelu je, jak už bylo řečeno v předpokladu, že bere všechny týmy v lize za stejně silné, což obecně není pravda. Například mistr ligy bývá ve většině případů lepší než nováček soutěže. Výsledkem tohoto modelu je především ukázka „výhody domácího prostředí“.

4.1.2 Model 1

V dalším modelu bude stále platit, že obrana všech týmů je stejně silná pro všechny týmy, ale v útoku mají týmy už různou sílu. Platí $\beta_i = \beta$, $\gamma_i = \gamma$, $\alpha_i = \delta_i$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Z toho vyplývá, že je potřeba odhadnout $n + 1$ nezávislých parametrů, kde n je počet týmů v lize. Na rozdíl od minulého modelu je zde brána v úvahu různá síla mužstev, ale zatím jen v útoku. A stále se zde odhaduje jen relativně málo parametrů.

Analogicky lze použít, že útok všech týmů bude stejně silný a síla obrany u každého týmu se bude lišit.

4.1.3 Model 2

V tomto modelu je pro každý tým síla v obraně i v útoku různá a navíc tu je parametr k, který vyjadřuje poměr síly týmů venku a síly týmů doma a tedy platí $k \cdot \alpha_i = \delta_i$, $k \cdot \beta_i = \gamma_i$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$ pro všechna i. V tomto modelu je třeba odhadovat $2n$ nezávislých parametrů. Oproti předcházejícím modelům už je zde rozlišena síla jednotlivých mužstev v ofenzivě i v defenzivě.

4.1.4 Model 3

V modelu č. 3 je počítána síla týmu v obraně a v útoku zvlášť pro každý tým. Navíc je počítána samostatně síla obrany doma a venku. Naopak síla týmu v útoku je brána za stejnou doma i venku, a tak platí $\alpha_i = \delta_i$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Zde se odhaduje $3n - 1$ nezávislých parametrů.

Analogicky lze počítat zvlášť sílu týmu v útoku doma a venku.

4.1.5 Model 4

V posledním modelu se bere samostatně síla týmu doma, venku, v útoku i v obraně. Musí zde platit $\sum_i \alpha_i = \sum_i \beta_i$ a $\sum_i \gamma_i = \sum_i \delta_i$. V takovém případě se odhaduje $4n - 2$ nezávislých parametrů.

4.2 Zkoumaný model

V další části se bude používat model 2 (kapitola 4.1.3). Tedy stejný model, který používal i Maher ve svém článku [1]. V tomto modelu se na rozdíl od modelu 0 a modelu 1 už bere v potaz rozdílná síla jednotlivých mužstev jak v obraně, tak v útoku a oproti dalším modelům je zde třeba odhadnout o dost méně parametrů. Oproti modelu 4 stačí odhadnout téměř jen polovinu parametrů.

V modelu 2 se odhadují parametr síly týmu v útoku α_i pro všechna i, parametr síly v obraně β_j pro všechna j a parametr k respektive k^2 vyjadřující sílu na hřištích soupeřů oproti síle při domácích utkáních.

Pro odhad těchto parametrů jsou v článku [1] na str. 114 odvozeny vzorce metodou maximální věrohodnosti. Tyto vzorce mají tvar

$$\hat{k}^2 = \frac{\sum_i \sum_{j \neq i} y_{ij}}{\sum_i \sum_{j \neq i} x_{ij}}, \tag{4.3}$$

$$\hat{\alpha}_i = \frac{\sum_{j \neq i}(x_{ij} + y_{ji})}{(1 + \hat{k}^2) \cdot \sum_{i \neq j} \hat{\beta}_i}, \tag{4.4}$$

$$\hat{\beta}_j = \frac{\sum_{i \neq j}(x_{ij} + y_{ji})}{(1 + \hat{k}^2) \cdot \sum_{j \neq i} \hat{\beta}_j}, \tag{4.5}$$

kde je

x_{ij} je počet branek vstřelený týmem i v domácím zápase týmu j,

y_{ij} je počet branek vstřelený týmem j týmu i ve venkovním zápase.

A dále musí platit následující podmínky

$$\sum_i \sum_{j \neq i} \hat{\alpha}_i \cdot \hat{\beta}_j = \sum_i \sum_{j \neq i} x_{ij}, \tag{4.6}$$

$$\sum_i \hat{\alpha}_i = \sum_i \hat{\beta}_i. \tag{4.7}$$

V zápase mezi domácím týmem i a hostujícím týmem j náhodná veličina X_{ij} značí počet gólů, které vstřelí tým i a náhodná veličina Y_{ij} udává počet branek vstřelený týmem j v zápase. Je předpokládáno, že X_{ij} a Y_{ij} jsou nezávislé. Potom X_{ij} a Y_{ij} se řídí Poissonovým rozdělením

$$X_{ij} \sim Po(\alpha_i \cdot \beta_j) \tag{4.8}$$

$$Y_{ij} \sim Po\left(k^2 \cdot \alpha_j \cdot \beta_i\right). \tag{4.9}$$

4.3 Sezóna 2013/14 Gambrinus liga

V sešitě Maher.xlsm a v listu GL2013-14 je vytvořen Maherův model číslo 2. Jako data pro tento model jsou použity počty gólů vstřelené a obdržené jednotlivými týmy Gambrinus ligy v sezóně 2013/2014 a celkový počet gólů vstřelený týmy doma a počet gólů vstřelený týmy venku.

4.3.1 Parametry k^2, α a β

Parametr k^2 vyjadřující poměr síly venku k síle týmů doma je odhadnut pro sezónu 2013/14 podle rovnice (4.3) jako 0,61. To znamená, pokud tým i dá průměrně doma 1 gól za zápas, potom venku dá průměrně 0,61 branky za zápas.

Následně se odhadnou parametry α, β pro každý tým, které vyjadřují sílu týmu v útoku respektive v obraně jednotlivých týmů. Odhady se provádějí iterativně podle rovnic (4.4) a (4.5). K odhadu se využije doplněk řešitel v Microsoft Excel. Ukázka počátečního nastavení Excelu před první iterací je na obrázku č. 2. Nejdříve jsou zvoleny počáteční hodnoty. Tyto počáteční hodnoty mohou být libovolné „rozumné“. Vzhledem k významu parametrů α a β, jejichž kombinace znamená průměrný počet gólů domácího týmu ve fotbalovém zápase, nemá smysl volit počáteční hodnoty záporné nebo naopak kladné vysoké (5+). Poté je možné spustit řešitel. V řešiteli je nastaveno, že buňka N18 se rovná 418, což je počet gólů vstřelený domácími týmy v sezóně 2013/2014. Tato podmínka plyne z rovnice (4.6). Dále je nastaveno I18 se rovná J18, což je rovnice (4.7). Měnícími parametry jsou startovací hodnoty tedy sloupce E a F. Po spuštění řešitele se dopočtou hodnoty do sloupců I a J, čímž je hotová první iterace. Dále se tyto výsledky nastaví jako startovací hodnoty pro druhou iteraci a opět se spustí řešitel se stejným nastavením. Toto se opakuje, dokud změna odhadu každého parametru $\hat{\alpha}_i$, $\hat{\beta}_i$ v jedné iteraci bude maximálně 0,01. Vzhledem k výsledkům pro různá nastavení počátečních podmínek lze předpokládat, že pokud jsou nastavené „rozumné“ startovací hodnoty, pak model dříve či později konverguje ke stejnému řešení. Toto bylo vyzkoušeno pro různá nastavení parametrů α a β. Výsledky jsou zaznamenány v listu jednoznačnost.

	D	E	F	G	H	I	J	K	L	M	N
1		α starovací	β startovací	suma α	suma β	α výpočet	β výpočet	α výsledné	β výsledné		Pomocné součty
2		1,00	1,00	15,00	15,00	2,65	0,87	1,95	0,68		71,54
3		1,00	1,00	15,00	15,00	1,33	1,74	1,00	1,30		34,62
4		1,00	1,00	15,00	15,00	1,78	1,66	1,35	1,26		46,67
5		1,00	1,00	15,00	15,00	1,41	2,03	1,08	1,52		36,38
6		1,00	1,00	15,00	15,00	3,23	0,79	2,37	0,63		87,46
7		1,00	1,00	15,00	15,00	1,08	1,66	0,81	1,23		28,22
8		1,00	1,00	15,00	15,00	1,37	1,78	1,04	1,33		35,64
9		1,00	1,00	15,00	15,00	1,53	1,90	1,17	1,43		39,77
10		1,00	1,00	15,00	15,00	1,78	2,19	1,38	1,67		45,71
11		1,00	1,00	15,00	15,00	2,24	1,57	1,69	1,22		58,79
12		1,00	1,00	15,00	15,00	2,11	1,45	1,59	1,12		55,79
13		1,00	1,00	15,00	15,00	1,74	2,48	1,36	1,89		44,14
14		1,00	1,00	15,00	15,00	0,99	2,11	0,76	1,56		25,59
15		1,00	1,00	15,00	15,00	1,45	1,53	1,09	1,15		38,16
16		1,00	1,00	15,00	15,00	1,86	2,07	1,43	1,58		48,07
17		1,00	1,00	15,00	15,00	1,33	2,03	1,02	1,52		34,24
18	Suma	16,00	16,00	240,00	240,00	27,87	27,87	21,08	21,08	**Řešitel**	**730,79**

Obrázek 2: Ukázka nastavení v Microsoft Excel před první iterací

V následující tabulce jsou zobrazeny výsledky po jednotlivých iteracích, pokud jsou počáteční hodnoty všech α, β parametrů nastaveny na 1.

	1. iterace		2. iterace		3. iterace		4. iterace	
Tým	α	β	α	β	α	β	**α**	**β**
1.FC Slovácko	1,35	1,24	1,35	1,26	1,35	1,26	**1,35**	**1,26**
1.FK Příbram	1,07	1,52	1,08	1,52	1,08	1,52	**1,08**	**1,52**
1.SC Znojmo FK	1,00	1,52	1,02	1,52	1,02	1,52	**1,02**	**1,52**
AC Sparta Praha	2,44	0,79	2,39	0,63	2,37	0,63	**2,37**	**0,63**
Bohemians Praha 1905	0,81	1,24	0,81	1,23	0,81	1,23	**0,81**	**1,23**
FC Baník Ostrava	1,03	1,33	1,04	1,33	1,04	1,33	**1,04**	**1,33**
FC Slovan Liberec	1,16	1,43	1,17	1,43	1,17	1,43	**1,17**	**1,43**
FC Viktoria Plzeň	2,00	0,65	1,95	0,68	1,95	0,68	**1,95**	**0,68**
FC Vysočina Jihlava	1,41	1,55	1,43	1,58	1,43	1,58	**1,43**	**1,58**
FC Zbrojovka Brno	1,00	1,30	1,00	1,30	1,00	1,30	**1,00**	**1,30**
FK Baumit Jablonec	1,35	1,64	1,37	1,67	1,38	1,67	**1,38**	**1,67**
FK Dukla Praha	1,10	1,15	1,09	1,15	1,09	1,15	**1,09**	**1,15**
FK Mladá Boleslav	1,69	1,18	1,68	1,22	1,69	1,22	**1,69**	**1,22**
FK Teplice	1,60	1,09	1,58	1,12	1,59	1,11	**1,59**	**1,12**
SK Sigma Olomouc	1,32	1,86	1,36	1,89	1,36	1,89	**1,36**	**1,89**
SK Slavia Praha	0,75	1,58	0,76	1,56	0,76	1,56	**0,76**	**1,56**

Tabulka 5: Odhadování parametrů α a β

4.3.2 Ukázka užití výsledků

Z parametrů α, β a k^2 lze vypočítat λ_{ij} (4.1) a μ_{ij} (4.2). Pro hypotetický zápas mezi domácí Plzní (ve vzorcích ozn. indexem P) a Brnem (ve vzorcích ozn. indexem B) odhad střední hodnoty počtu gólů vstřelených Plzní je

$$\lambda = \alpha_P \cdot \beta_B = 1{,}95 \cdot 1{,}30 = 2{,}53. \tag{4.10}$$

Parametr μ pro počet gólů vstřelených Brnem je

$$\mu = k^{2,} \cdot \alpha_B \cdot \beta_P = 0{,}61 \cdot 1{,}00 \cdot 0{,}68 = 0{,}42. \tag{4.11}$$

Parametry λ a μ spočtené pro zápasy mezi všemi týmy jsou v sešitu Maher v tabulce λ, respektive v tabulce μ.

Pokud jsou známy parametry λ a μ, tak je možné určit pravděpodobnosti vyjadřující kolik dá tým v zápase gólů. Pro ukázkový zápas mezi Plzní a Brnem je λ = 2,57. Podle (2.1) lze vypočítat pravděpodobnost, že Plzeň vsítí Brnu 0, 1, 2,… branek. Například pravděpodobnost, že Plzeň nedá žádný gól je

$$P(X = 0) = e^{-2{,}53} \cdot \frac{2{,}53^0}{0!} = 0{,}08. \tag{4.12}$$

Pravděpodobnost, že Plzeň dá Brnu 4 a více gólů je

$$P(X \geq 4) = 1 - F(3) = 0{,}25. \tag{4.13}$$

Pravděpodobnost, že Brno nedá Plzni gól, je

$$P(Y = 0) = e^{-0,42} \cdot \frac{0,42^0}{0!} = 0,66. \tag{4.14}$$

Nyní je možné dopočítat pravděpodobnost výsledku 0:0

$$P(X = 0, Y = 0) = P(X = 0) \cdot P(y = 0) = 0,08 \cdot 0,66 = 0,05. \tag{4.15}$$

V další tabulce jsou pravděpodobnosti všech výsledků v zápase Plzeň Brno od 0:0 do 4+:4+.

		Brno					
	Počet gólů	0	1	2	3	4+	suma
Plzeň	0	0,052	0,022	0,005	0,001	<0,001	0,08
	1	0,133	0,055	0,012	0,002	<0,001	0,20
	2	0,168	0,070	0,015	0,002	<0,001	0,26
	3	0,142	0,059	0,012	0,002	<0,001	0,22
	4+	0,164	0,069	0,014	0,002	<0,001	0,25
	suma	0,66	0,28	0,06	0,01	0,00	

Tabulka 6: Pravděpodobnost výsledků v zápase Plzeň - Brno

Pravděpodobnost, že domácí mužstvo nedá žádný gól hostujícímu týmu v zápase mezi jakýmikoliv týmy, je v tabulce $P(X = 0)$. Podobně pravděpodobnost, že mužstvo domácí dá jeden gól, je v tabulce $P(X = 1)$ atd. Obdobně pravděpodobnost, že hostující mužstvo domácímu mužstvu nedá žádnou branku je v tabulce $P(Y = 0)$ atd.

4.3.3 Chí kvadrát test

Na závěr podle Maherova článku [1] je otestováno, zda pravděpodobnosti vypočtené v předchozím modelu odpovídají skutečným výsledkům. Testování je prováděno pomocí χ^2 testu dobré shody, popsaného v kapitole 2.2. Zvlášť jsou testovány góly doma, zvlášť venku. Oba testy jsou uvedeny v sešitu Maher.xlsm v listu Chí kvadrát test.

H_0: Počty gólů vstřelených týmy doma (venku) v sezóně 2013/2014 se neliší od počtu gólů v Maherovu modelu č. 2.

H_1: Počty gólů vstřelených týmy doma (venku) v sezóně 2013/2014 se liší od počtu gólů v Maherovu modelu č. 2.

Testuje se na hladině významnosti 5 %.

Pozorované hodnoty n_i se určí z výsledků jednotlivých zápasů. Například počet utkání, kdy domácí tým nedal gól, je 52. Očekávaný počet se získá jako suma celé tabulky $P(X = 0)$, což v tomto případě vyjde 51,90 zápasů.

Počet gólů x	**0**	**1**	**2**	**3**	**4+**
Skutečný počet n_i	52	73	50	32	33
Očekávaný počet o_i	51,90	70,38	55,35	33,09	29,28

Tabulka 7: Skutečný a očekávaný počet gólů v domácích zápasech

Počet gólů x	0	1	2	3	4+
Skutečný počet n_i	84	94	39	13	10
Očekávaný počet o_i	90,19	81,55	42,85	17,16	8,24

Tabulka 8: Skutečný a očekávaný počet gólů ve venkovních zápasech

P hodnota pro test domácích týmů je 0,77 a pro test hostujících je 0,26, z toho vyplývá, že se hypotéza H_0 nezamítá ani v testu pro domácí týmy, ani pro hostující týmy.

Za povšimnutí však stojí rozdíl mezi očekávaným a skutečným počtem zápasů, ve kterých hostující týmy daly 0 nebo 1 gól. Zatímco model předpokládá větší počet zápasů s žádným gólem, tak ve skutečnosti bylo daleko více zápasů, ve kterých dal venkovní tým 1 gól.

4.3.4 Závěr

Dle chí kvadrát testu lze říct, že počty gólů se řídí Poissonovým rozdělením s parametry dle modelu č. 2. Je nutné však zmínit, že test se dělal pro celou sezónu, zatímco jednotlivé zápasy mohou mít jiné rozdělení pravděpodobnosti.

Nevýhodou takto zkonstruovaného modelu je, že se dá modelovat vždy po jednotlivých sezónách, protože počet zápasů každého týmu musí být stejný. Vzhledem k tomu, že každý rok dva nejhorší týmy z ligy sestupují, tak po více sezónách by měli některé týmy odehráno více zápasů než jiné. Další nevýhodou je, že se v modelu neprojevuje aktuální forma z posledních zápasů, ale stejnou váhu má zápas jak z prvního, tak z patnáctého i z dvacátého kola. Tyto nedostatky budou odstraněny v dalším modelu (kapitola 5).

5 Dixon - Colesův model

Maherův model vylepšili v devadesátých letech Mark J. Dixon a Stuart G. Coles. Vylepšený model popsali ve svém článku [2].

5.1 Popis modelu Dixon - Coles

Cílem tohoto modelu je opět určit, s jakou pravděpodobností dají týmy určitý počet gólů v zápase a tím odhadnout celkový výsledek utkaní. Tedy zjistit, s jakou pravděpodobností tým vyhraje, remízuje či prohraje.

Model zahrnuje různou sílu jednotlivých týmů v útoku i v obraně, „výhodu domácího prostředí“, navíc je tento model dynamický, což je důležité, protože síla týmů v čase se mění a to ať v krátkodobém období, což je dáno například měnící se formou týmů nebo příchodem nového trenéra, tak i v dlouhodobém období, na což má vliv například příchod nových hráčů. K modelování výsledků je opět použito Poissonovo rozdělení pravděpodobnosti, tentokrát dvojrozměrné. Navíc je zde přidána funkce τ kvůli závislosti mezi počtem gólů domácích a hostů.

5.1.1 Sdružená pravděpodobnostní funkce

Pro výsledek zápasu mezi domácím týmem i a hostujícím týmem j je sdružená pravděpodobnostní funkce ve tvaru

$$P\left(X_{i,j} = x, Y_{i,j} = y\right) = \tau_{\lambda,\mu}(x,y) \cdot \frac{\lambda^x \cdot e^{-\lambda}}{x!} \cdot \frac{\mu^y \cdot e^{-\mu}}{y!}, \tag{5.1}$$

kde je

X_{ij} náhodná veličina vyjadřující počet gólů vstřelených domácím týmem i,
Y_{ij} náhodná veličina vyjadřující počet gólů vstřelených hostujícím týmem j,
λ parametr určující počet gólů domácích,
μ parametr určující počet gólů hostů,
τ funkce vyjadřující závislost mezi X_{ij} a $Y_{ij.}$

5.1.2 Parametry λ, μ

Parametr λ je vyjádřen následujícím vzorcem

$$\lambda_{ij} = \alpha_i \cdot \beta_j \cdot \gamma, \tag{5.2}$$

kde je

α_i síla domácího týmu v útoku,
β_j síla hostujícího týmu v obraně,
γ parametr vyjadřující výhodu domácího prostředí

Parametr μ je vyjádřen následujícím vzorcem

$$\mu_{ij} = \alpha_j \cdot \beta_i, \tag{5.3}$$

kde je

α_j síla hostujícího týmu v útoku,

β_i síla domácího týmu v obraně.

Jako ochrana před přeparametrizováním modelu je dána podmínka pro α

$$\sum_{i=1}^{n} \alpha_i = n, \tag{5.4}$$

kde je

n počet týmu, pro které se odhadují parametry α a β.

5.1.3 Funkce závislosti τ

Počty gólů domácích a počty gólů hostů nejsou nezávislé veličiny. Jinak hraje tým, který vede, a jinak hraje tým, který prohrává. To má vliv na počet gólů domácích i hostů a různou četnost jednotlivých výsledků. Toho si všimli Dixon a Coles, a proto do modelu použili funkci τ, která upravuje nejčastější výsledky fotbalových zápasů 0:0, 1:1, 1:0 a 0:1. $\rho = 0$ určuje nezávislost mezi X, Y. Funkce τ má tvar

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1 - \lambda\mu\rho, & pro\ x = 0\ y = 0 \\ 1 + \lambda\rho, & pro\ x = 0\ y = 1 \\ 1 + \mu\rho, & pro\ x = 1\ y = 0 \\ 1 - \rho, & pro\ x = 1\ y = 1 \\ 1, & jinak, \end{cases} \tag{5.5}$$

kde je

λ parametr určující počet gólů domácích,

μ parametr určující počet gólů hostů,

x počet gólů domácích,

y počet gólů hostů,

ρ parametr závislosti.

Pro ρ platí

$$\max\left(-\frac{1}{\lambda}, -\frac{1}{\mu}\right) \leq \rho \leq \min\left(\frac{1}{\lambda\mu}, 1\right). \tag{5.6}$$

5.2 Způsob odhadu parametrů

Parametry v tomto modelu jsou odhadovány pomocí metody maximální věrohodnosti.

5.2.1 Věrohodnostní funkce

Jak bylo napsáno výše, v této části se pracuje s dynamickým modelem, tak je do věrohodnostní funkce zanesena i funkce času $\phi(t)$. Základní tvar věrohodnostní funkce je

$$V(\alpha_i, \beta_i, \varrho, \gamma; i = 1, \dots, n) = \prod_{k=1}^{n} \left(\tau_{\lambda_k, \mu_k}(x_k, y_k) \cdot \frac{\lambda_k^{x_k} \cdot e^{-\lambda_k}}{x_k!} \cdot \frac{\mu_k^{y_k} \cdot e^{-\mu_k}}{y_k!} \right)^{\phi(t-t_k)}, \tag{5.7}$$

kde je

λ_k	parametr určující počet gólů domácích,
μ_k	parametr určující počet gólů hostů,
τ	funkce vyjadřující závislost mezi X_{ij} a Y_{ij},
x_k	počet gólů domácího týmu i v zápase k,
y_k	počet gólů hostujícího týmu j v zápase k,
$\phi(t - t_k)$	funkce času (kapitola 5.2.3).

5.2.2 Logaritmická věrohodnostní funkce

Protože pro odhad parametrů není důležité absolutní číslo, ale jen polohy bodů maxima, tak je možné věrohodnostní funkci zlogaritmovat a tím se odhady parametrů nezmění. Ze stejného důvodu je možné vynechat členy $\ln x_k!$ respektive $\ln y_k!$. Zlogaritmována funkce má následující tvar

$$\begin{aligned} &L(\alpha_i, \beta_i, \rho, \gamma; i = 1, \dots, n) = \\ &= \textstyle\sum_{k=1}^{n} (\phi(t - t_k) \cdot (\ln \tau_{\lambda_k, \mu_k}(x_k, y_k) + x_k \cdot \ln \lambda_k - \lambda_k + y_k \cdot \ln \mu_k - \mu_k)). \end{aligned} \tag{5.8}$$

5.2.3 Funkce času ϕ

Funkce $\phi(t)$ je funkce času. Pomocí ní je možné v odhadu preferovat zápasy odehrané v nedávné době oproti výsledkům, které se zrodily před delším časem.

Funkci $\phi(t)$ je možné definovat různými způsoby. V této práci je použita podobná funkce, kterou použili Dixon a Coles ve svém modelu [2]. Rozdíl je v tom, že zde je čas t počítán ve dnech, zatímco v Dixon - Colesovo modelu byl počítán v „polotýdnech“

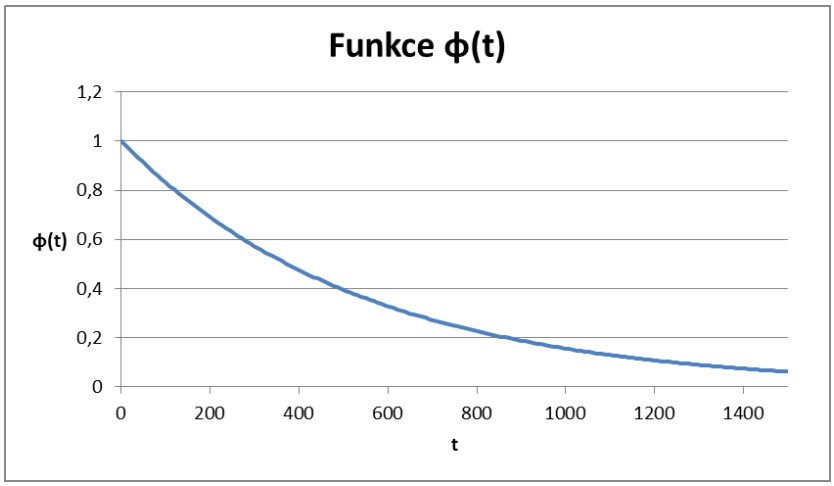
$$\phi(t) = e^{-\xi \cdot t}, \tag{5.9}$$

kde je

ξ	váha,
t	počet dní, které uplynuly od doby zápasu ke dni odhadu parametrů.

Nyní je třeba ještě určit váhu ξ. Toto určení je problematické, protože váha ξ nezávisí na pravděpodobnostech a nedá se odhadovat z věrohodnostní funkce, ale je nutné ji určit předem. V tomto modelu je zvolená váha ξ = 0,0018671, což je váha zvolená Dixonem a Colesem přepočtená z „polotýdnů“ na dny vydělením jejich původní váhy 3,5 dny.

Například pokud se bude odhadovat kolo hypoteticky hrané $1.1.2014$, potom čas hypotetického zápasu t hraného $1.1.2013$ je 365 a $\phi(t)$ je 0,508.



Obrázek 3: Funkce času

5.3 Data

K odhadu parametrů metodou maximální věrohodnosti je potřeba znát výsledky z minulých zápasů (sezón), na jejichž základě budou odhadnuty parametry α_i, β_i, γ a ρ a z nich budou následně odhadovány výsledky budoucích utkání.

V této práci se budou odhadovat výsledky zápasů české nejvyšší soutěže Gambrinus ligy (od sezóny 2014/2015 Synot ligy), dále anglické Premier League, španělské La Liga a italské Seria A.

5.4 Gambrinus liga

V české nejvyšší soutěži hraje 16 mužstev. Každé dva týmy během jedné sezóny spolu sehrají 2 zápasy jeden doma a jeden venku. Jeden ročník má 30 kol a je v něm odehráno 240 utkání. Poslední dva týmy po posledním kole sestupují do nižší soutěže a 2 nejlepší týmy z druhé ligy postoupí do první.

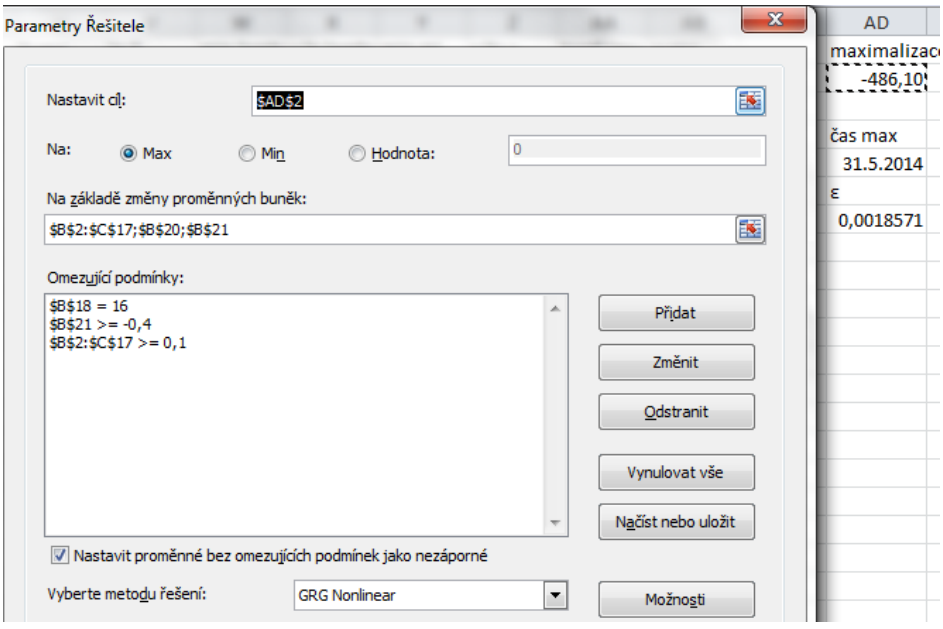
České kluby hrají mezi sebou také pohár FAČR. V něm v první fázi začínají hrát týmy z nižších soutěží a mužstva z první ligy jsou nasazena až do druhého či třetího kola. Pro většinu českých klubů však pohár není až tak zajímavá soutěž a do zápasu v poháru často staví náhradníky a dochází zde často k hodně nečekaným výsledkům. Proto nebyly zápasy poháru zaneseny do tohoto modelu na rozdíl od Dixona a Colese, kteří do svého modelování zařadili i výsledky z anglických pohárů. Výhodou zanesení zápasů v poháru do modelu je možnost porovnání lig mezi sebou tj. první s druhou atd. Zde nastává tedy odlišnost od modelu Dixon - Colese, kteří odhadovali parametry pro týmy z více lig v jedné zemi zároveň, a zde se bude odhadovat pouze pro jednu nejvyšší soutěž. Dalším důvodem, proč se odhadují parametry týmů jen v nejvyšší soutěži, je zavedení tzv. juniorské ligy od sezóny 2012/2013 [7]. To mělo za následek zrušení „B-týmu“ většiny prvoligových mužstev, které obvykle hrály druhou, třetí či čtvrtou ligu. Tím pádem došlo k velkým obměnám týmů v nižších českých soutěžích. Naopak v anglických soutěžích vždy

postupují a sestupují jen 3, respektive 4 týmy. Problémem modelu, kde se odhaduje pouze nejvyšší soutěž, je, že pro nováčka, který nehrál za sledované období nejvyšší soutěž, nejsou na začátku sezóny k dispozici žádná data.

Odhadovat se budou výsledky zápasů v sezóně 2013/2014 od 6. kola a to právě kvůli nováčkovi v nejvyšší české lize týmu 1. SC Znojmo FK, pro který nebyla k dispozici data z minulých let, protože tento tým hrál jen nižší soutěž. Výsledky pro odhad zápasů jsou sesbírány od sezóny 2010/2011. Vzhledem k časové funkci $\phi(t)$ a jejímu parametru ξ nemá cenu pracovat v modelu se staršími zápasy, protože jejich váha by byla velmi nízká. Od začátku sezóny 2010/2011 do konce sezóny 2012/2013 bylo sehráno 720 utkání. Do modelu bylo zaneseno pouze 488 z nich. Konkrétně byly vynechány zápasy týmů, které v sezóně 2013/2014 nehrají první ligu. Jedná se o týmy FK Ústí nad Labem, FC Viktoria Žižkov, FC Hradec Králové a SK Dynamo České Budějovice. Vzhledem k dostatečnému množství výsledků ostatních zápasů, vynechání těchto utkání výrazně neovlivní odhady parametrů ostatních mužstev a zároveň to zabrání nestabilitě parametrů pro tato mužstva, kdyby se tyto parametry musely odhadovat.

5.4.1 Odhad parametrů Gambrinus liga

Odhadování výsledků české ligy je prováděno v sešitu CZEDixon.xlsx v listu Odhad. Odhad probíhá maximalizací věrohodnostní funkce rovnice (5.8), která je v tomto případě v buňce $AD2$. K maximalizaci je použit řešitel, což je doplněk programu Microsoft Excel. V něm je vybrána metoda řešení GRG Nonlinear [8] a nastavena zastavovací podmínka 0,0001, což znamená, pokud se žádný z parametrů nezmění o víc než 0,0001, tak výpočet skončí.



Obrázek 4: Nastavení řešitele Microsoft Excel

V průběhu výpočtu se mění parametry síly v útoku α_i, síly v obraně β_i pro všechny týmy i a dále parametr domácího prostředí γ a parametr závislosti ρ. Všechny tyto parametry jsou ve sloupcích B a C.

	A	B	C
1	Tým	α	β
2	FC Viktoria Plzeň	1,45	0,64
3	FC Zbrojovka Brno	0,80	1,17
4	1.FC Slovácko	0,95	0,98
5	1.FK Příbram	0,80	1,21
6	AC Sparta Praha	1,60	0,56
7	Bohemians Praha 1905	0,63	1,07
8	FC Baník Ostrava	0,79	1,13
9	FC Slovan Liberec	1,07	1,08
10	FK Baumit Jablonec	1,14	1,28
11	FK Mladá Boleslav	1,16	1,02
12	FK Teplice	1,05	1,00
13	SK Sigma Olomouc	1,01	1,29
14	SK Slavia Praha	0,72	1,09
15	FK Dukla Praha	0,91	1,04
16	FC Vysočina Jihlava	1,06	1,20
17	1.SC Znojmo FK	0,84	1,28
18		16,00	17,05
19			
20	γ	1,55	
21	ρ	-0,11	

Obrázek 5: Odhad parametrů v Microsoft Excel

Definičním oborem parametrů α_i, β_i, γ, τ_k, λ_k a μ_k pro všechny týmy i a zápasy k jsou nezáporná reálná čísla, což vyplývá z logaritmické věrohodnostní funkce (5.8) a také z významu parametrů α a β, které vyjadřují sílu v útoku a obraně. Navíc pro parametry α_i a ρ jsou nastaveny podmínky z rovnic (5.4) a (5.6).

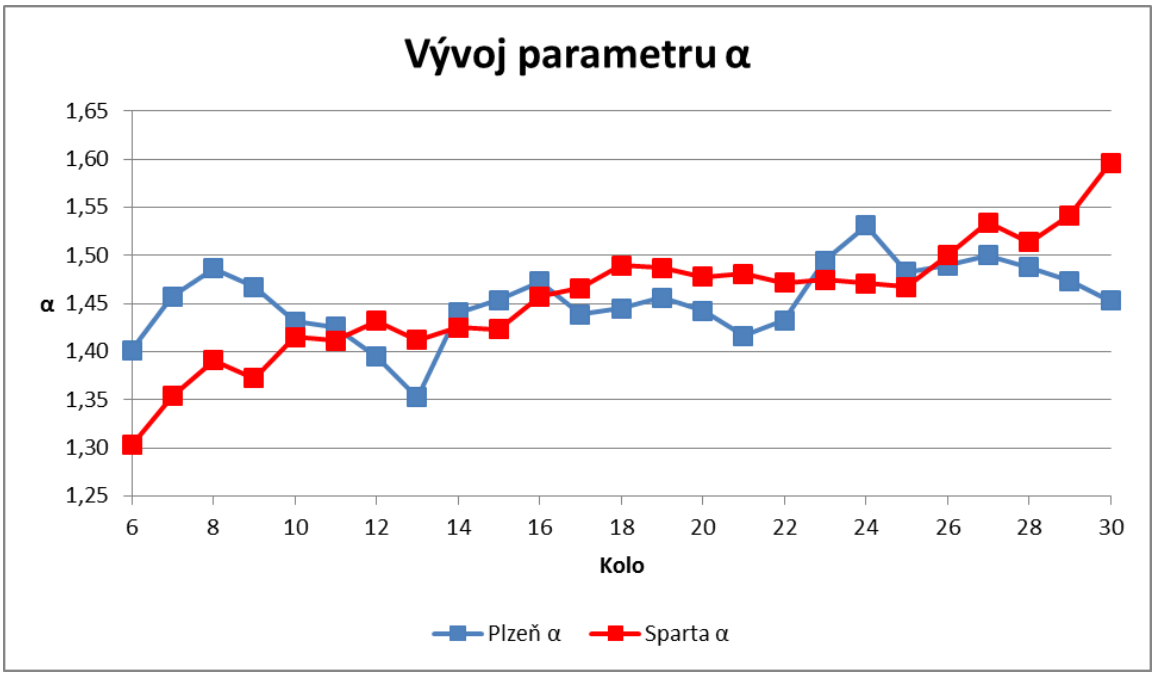
Aby bylo možné spustit řešitel, je třeba nastavit počáteční hodnoty parametrů. Zde byly nastaveny pro šesté kolo všechny parametry α a β na 1, parametr γ na 1,5 a parametr ρ na 0. Pro další kola se vždy bere za počáteční hodnoty kolo předcházející a to především z důvodu rychlejší konvergence. Ta je zapříčiněna tím, že se parametry během jednoho kola nemohou o tolik změnit. Vzhledem k výsledkům pro několik různých nastavení počátečních podmínek lze však předpokládat, že pokud bude model konvergovat, dříve či později dojde ke stejnému řešení. To je ukázáno v listu Jednoznačnost, kde jsou pro různá nastavení počátečních podmínek pro odhad 30. kola dopočteny odhady jednotlivých parametrů. V tomto listu je vidět, že pro všechny počáteční nastavení se dospělo ke stejným hodnotám s výjimkou dvou nastavení, kdy řešitel během výpočtu nahlásil chybu. Ta je způsobena tím, že pro některé zápasy k se během výpočtu dostane τ_k do záporných čísel tedy mimo svůj definiční obor, výpočet nemůže pokračovat, a proto řešitel nahlásí chybu. V tomto případě nelze nastavit podmínku nezápornosti τ_k, protože řešitel umožnuje nastavit pouze 200 buněk s podmínkou, zatímco zápasů je více.

Odhadnuté parametry pro všechna kola jsou v listu Parametry. V následující tabulce jsou zobrazeny parametry α a β odhadnuté pro poslední 30. kolo.

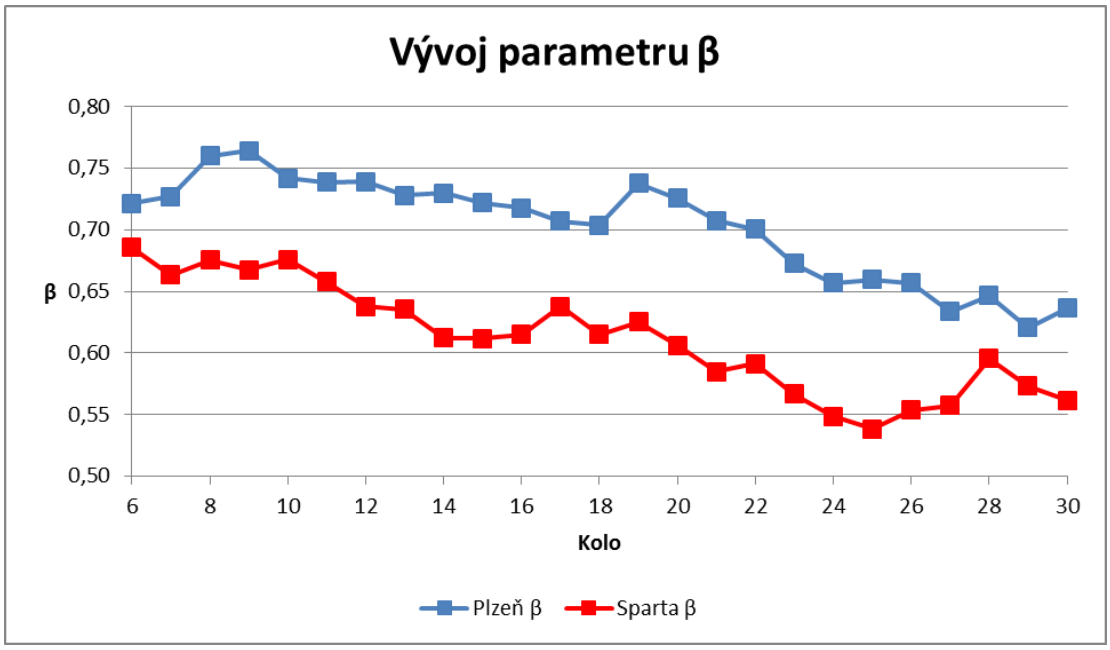
Tým	α	β
1.FC Slovácko	0,95	0,98
1.FK Příbram	0,80	1,21
1.SC Znojmo FK	0,84	1,28
AC Sparta Praha	1,60	0,56
Bohemians Praha 1905	0,63	1,07
FC Baník Ostrava	0,79	1,13
FC Slovan Liberec	1,07	1,08
FC Viktoria Plzeň	1,45	0,64
FC Vysočina Jihlava	1,06	1,20
FC Zbrojovka Brno	0,80	1,17
FK Baumit Jablonec	1,14	1,28
FK Dukla Praha	0,91	1,04
FK Mladá Boleslav	1,16	1,02
FK Teplice	1,05	1,00
SK Sigma Olomouc	1,01	1,29
SK Slavia Praha	0,72	1,09

Tabulka 9: Odhad parametrů α a β pro 30. kolo (tj. z výsledků do 29. kola včetně)

Pro představu, jak se mění parametry α, β během sezóny, je zde uveden vývoj těchto parametrů u dvou nejúspěšnějších týmů v české lize za poslední roky. Jedná se o týmy FC Viktoria Plzeň a AC Sparta Praha.



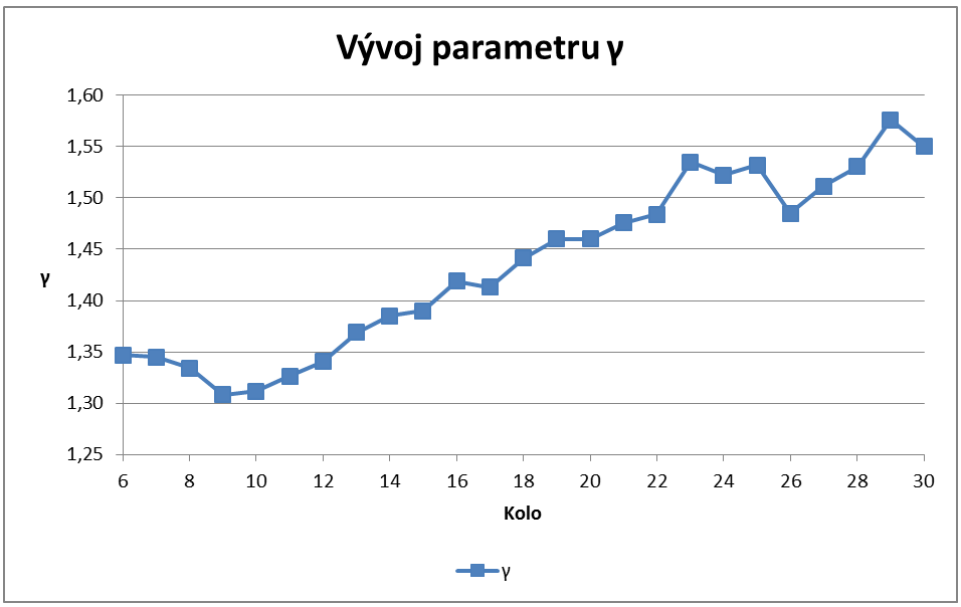
Obrázek 6: Vývoj parametru α u týmů FC Viktoria Plzeň a AC Sparta Praha



Obrázek 7: Vývoj parametru β u týmů FC Viktoria Plzeň a AC Sparta Praha

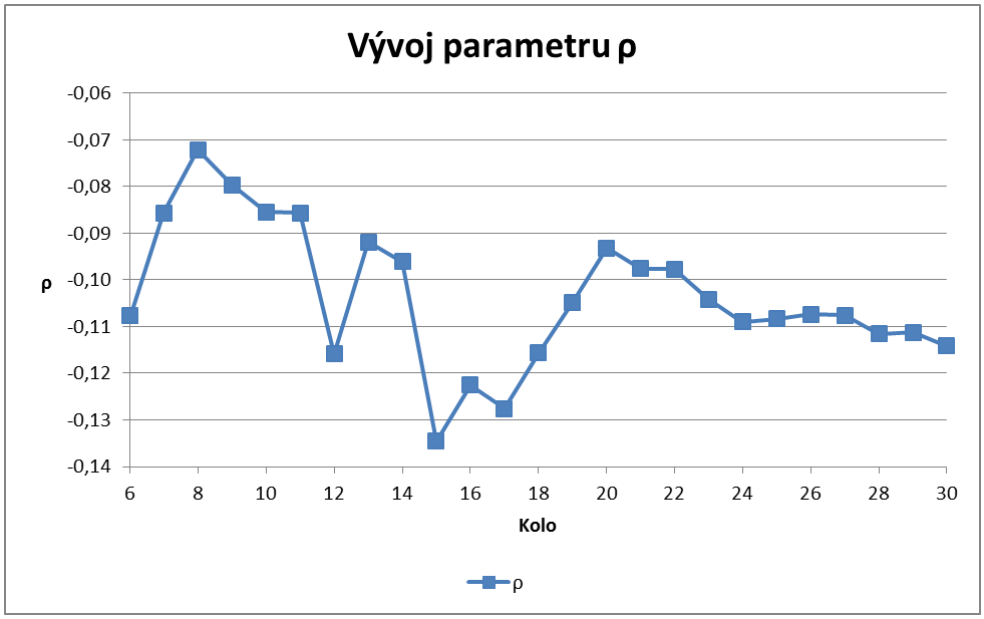
Z grafu jsou patrné rozdíly v parametrech mezi jednotlivými koly. To značí, jak byl tým silný v jednotlivých částech sezóny, tedy jeho aktuální formu. Za povšimnutí stojí větší rozdíly v parametru α u týmu FC Viktoria Plzeň mezi 13. a 14. kolem či 22. a 23. kolem. To je způsobeno tím, že ve 13. kole i v 22. kole vstřelila FC Viktoria Plzeň svým soupeřům 6 gólů. U AC Sparta Praha roste rychle parametr α mezi 6. až 8. kolem, protože v 6. i 7. kole vstřelila Sparta 4 góly.

Na následujícím obrázku je zobrazen parametr γ. Parametr γ během sezóny převážně rostl. To znamená, že se zvětšovala výhoda domácího prostředí.



Obrázek 8: Vývoj parametru γ

Na dalším obrázku je zobrazen vývoj parametru ρ, který určuje závislost mezi počtem gólů domácích a hostů. Tento parametr se během sezóny pohyboval v záporných číslech, což znamená, že parametr ρ zvětšoval pravděpodobnost výsledků 0:0 a 1:1, které by byly v případě použití „nezávislého“ modelu podhodnoceny, a naopak snižoval pravděpodobnost u výsledků 1:0 a 0:1.



Obrázek 9: Vývoj parametru ρ

5.4.2 Odhad výsledků zápasů

Pokud jsou známy všechny parametry, je možné odhadnout výsledky zápasů pomocí sdružené pravděpodobnostní funkce viz. rovnice (5.1). Odhady zápasů v jednotlivých kolech jsou prováděny v listu Kolo a závěry jsou zaznamenávány v listu 2013-14.

Pro ukázku zde bude uveden odhad výsledku zápasu 30. kola mezi týmy FC Baník Ostrava (ve vzorcích ozn. indexem O) a SK Slavia Praha (ve vzorcích ozn. indexem S). Odhadnuté parametry pro tento zápas jsou v následující tabulce.

Parametr	Hodnota
α_O	0,79
β_O	1,13
α_S	0,72
β_S	1,09
γ	1,55
ρ	-0,11

Tabulka 10: Odhadnuté parametry pro zápas FC Baník Ostrava - SK Slavia Praha

Výpočet parametru λ pro tento zápas

$$\lambda_{O,S} = \alpha_O \cdot \beta_S \cdot \gamma = 0{,}79 \cdot 1{,}09 \cdot 1{,}55 = 1{,}33. \tag{5.10}$$

Výpočet parametru μ pro tento zápas

$$\mu_{O,S} = \alpha_S \cdot \beta_O = 0{,}72 \cdot 1{,}13 = 0{,}81. \tag{5.11}$$

Nyní už je možné dosadit do pravděpodobnostní funkce rovnice (5.1). Pro výsledek 0:0 je pravděpodobnost

$$P(X = 0, Y = 0) = [1 - 1.33 \cdot 0{,}81 \cdot (-0{,}11)] \cdot \frac{1{,}33^0 \cdot e^{-1{,}33}}{0!} \cdot \frac{0{,}81^0 \cdot e^{-0{,}81}}{0!} = 0{,}130. \tag{5.12}$$

Ve skutečnosti tento zápas skončil vítězstvím Baníku Ostrava 2:0. Pravděpodobnost tohoto výsledku počítána modelem před zápasem byla

$$P(X = 2, Y = 0) = \frac{1{,}33^2 \cdot e^{-1{,}33}}{2!} \cdot \frac{0{,}81^0 \cdot e^{-0{,}81}}{0!} = 0{,}103. \tag{5.13}$$

Podobně se dopočítává pravděpodobnost pro všechny možné výsledky teoreticky až do výsledku ∞: ∞.

V následující tabulce je vypočtena pravděpodobnost pro jednotlivé výsledky.

		Baník Ostrava					
	Počet gólů	0	1	2	3	4	5+
Slavia Praha	0	0,13	0,14	0,10	0,05	0,02	<0,01
	1	0,08	0,14	0,08	0,04	0,01	<0,01
	2	0,04	0,05	0,03	0,02	0,01	<0,01
	3	0,01	0,01	0,01	<0,01	<0,01	<0,01
	4	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
	5+	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01

Tabulka 11: Pravděpodobnost výsledků v zápase Baník Ostrava - Slavia Praha

Hlavním cílem není zjistit, jaká je pravděpodobnost jednotlivých výsledků, ale důležité je zjistit pravděpodobnost výhry domácích, hostů a remízy. Pokud se sečtou v tabulce všechny výsledky, při kterých vyhraje Baník, tak výsledek je 0,473. Součet výsledků výher Slavia je 0,216 a remízy je 0,311.

Výsledek	**Pravděpodobnost**
Výhra Baníku Ostrava	0,473
Remíza	0,311
Výhra Slavia Praha	0,216

Tabulka 12: Pravděpodobnost výhry domácích, remízy, výhry hostů

5.5 Další ligy

Španělská Primera División, italská Seria A a anglická Premier League jsou další 3 soutěže, které se budou odhadovat pomocí Dixon - Colesova modelu.

V každé z těchto 3 lig hraje 20 týmů. Stejně jako v české lize každé dva týmy během jedné sezóny spolu sehrají 2 zápasy jeden doma a jeden venku. Během jednoho ročníku je tedy odehráno 38 kol a 380 utkání. Poslední tři týmy na konci soutěže sestupují do nižší ligy a tři nejlepší týmy z druhé ligy postoupí do první.

Stejně jako v případě české ligy se budou i ve španělské, italské a anglické lize odhadovat výsledky v sezóně 2013/2014 na základě předchozích ligových výsledků od sezóny 2010/2011. Podobně jako v české lize jsou i zde vynechány zápasy týmů, které nehrají nejvyšší soutěž v sezóně 2013/2014. Konkrétně ve Španělsku se jedná o mužstva Hércules CF, Sporting Gijón, Racing Santander, Deportivo La Coruña, RCD Mallorca a Real Zaragoza. V Itálii jde o mužstva AC Siena, Delfino Pescara, Palermo, US Lecce, Novara Calcio, AC Cesena, Brescia Calcio a AS Baria a v Anglii se jedná o týmy Wigan Athletic, Reading FC, Queens Park Rangers, Bolton Wanderers, Blackburn Rovers, Wolverhampton Wanderers, Birmingham City a Blackpool FC.

5.5.1 Odhady parametrů

Odhadování výsledků jednotlivých lig je vždy prováděno v listu odhad v sešitu SPADixon.xlsx pro španělskou ligu, v sešitu ITADixon.xlsx pro italskou ligu a v sešitu ENGDixon.xlsx pro anglickou ligu. Odhad v případě španělské a italské ligy je prováděn úplně stejným způsobem jako v případě české ligy. Jediný problém ve španělské lize nastává v zápase 34. kola mezi týmy Real Valladolid - Real Madrid. Tento zápas byl odložen a odehrán až po 36. kole. Proto jsou odhadnuty parametry zvlášť pro tento zápas. V italské lize nastává podobný problém pro zápas 22. kola mezi týmy AS Řím - Parma FC. Tento zápas byl odehrán až po 31. kole a pro tento zápas jsou odhadnuty parametry opět zvlášť. V anglické lize je takto dohrávaných a předehrávaných zápasů více. Proto anglická liga není odhadována po jednotlivých kolech, ale po skupině zápasů, tak aby v žádné skupině nehrál nějaký tým více než jedno utkání. Kromě této změny jsou parametry odhadovány stejným způsobem jako u české ligy. Všechny odhady parametrů jsou v listu Parametry. Odhady výsledků zápasů v jednotlivých kolech jsou prováděny v listu kolo a závěry jsou zaznamenávány v listu 2013-14.

6 Sázení

Sázení je oblíbená činnost spousty lidí po celém světě. Předmětem sázky může být jakýkoliv náhodný pokus, který s nenulovou pravděpodobností nabývá alespoň dvou různých výsledků. Vsadit si mohou například dva lidé či více mezi sebou, anebo si jednotlivec může vsadit v sázkové kanceláři na pobočce či na internetu. Výhodou internetového sázení bývá, že je sázka obvykle bez manipulačního poplatku. Zatímco na pobočce se k sázce musí zaplatit ještě tento poplatek.

Cílem každého sázejícího je vyhrát. Nejčastěji (u sázkových kanceláří vždy) bývá výhra vyplacena v penězích. Stejný cíl, tedy zisk, má i sázková kancelář. Ta však nemusí být v zisku v souboji s každým sázejícím, ale chce být v zisku v souboji se všemi sázejícími dohromady.

V kapitole 5 byl počítán Dixon - Colesův model pro výpočet pravděpodobností výsledků fotbalových utkání. V následující kapitole bude zkoumáno, jak si povede tento model ve srovnání se sázkovou kanceláří.

6.1 Základní pojmy

V této kapitole bude čerpáno z přednášek předmětu KIV/ZTI [9].

U fotbalového zápasu je možné vsadit na výhru domácích ozn. 1, remízu ozn. 0 nebo výhru hostů ozn. 2. Pro každý tento výsledek zápasu existuje pravděpodobnost p_i pro i = 1, 0, 2, s kterou tento výsledek nastane. Tato pravděpodobnost je však neznámá, jak pro sázejícího, tak i pro sázkovou kancelář.

Sázející se snaží odhadnout tyto pravděpodobnosti p_i pravděpodobnostmi r_i, a to buď na základě znalostí síly jednotlivých sportovních týmů získaných sledováním sportovních utkání, anebo například pomocí matematických modelů, jako je Dixon - Colesův model.

Sázková kancelář nejdříve odhadne pravděpodobnosti výsledku zápasu q_i. Z těchto pravděpodobností sázková kancelář vychází při tvorbě kurzů. Pro sázkovou kancelář je ideální stav, pokud sázkaři rozloží svoje sázky tak, aby při libovolném výsledku zápasu vyplácela v součtu stejnou částku. V takovém případě, vzhledem k marži sázkové kanceláře ζ, bude sázková kancelář vždy v zisku.

Kurz o_i, což je hodnota výplaty sázejícímu při úspěšné sázce s vkladem 1 jednotka, je počítán dle následujícího vzorce

$$o_i = \frac{1 - \zeta}{q_i}, \tag{6.1}$$

kde je

ζ marže sázkové kanceláře,

q_i pravděpodobnost výsledku odhadnutá sázkovou kanceláří.

Pro ukázku je uveden zápas mezi domácím týmem A a hostujícím týmem B. Pravděpodobnost q_i výhry domácích odhadla sázková kancelář na 50 %, pravděpodobnost výhry hostů na 20 % a pravděpodobnost remízy na 30 %. Zároveň chce sázková kancelář pro sebe marži 10 %. Kurzy o_i, které sázková kancelář na takový zápas vypíše, jsou v následující tabulce.

Výsledek	1	0	2
Kurz	1,8	3,0	4,5

Tabulka 13: Kurzy na zápas mezi týmy A a B

Pokud na tento zápas bude vsazeno 1 000 jednotek, tak bude pro sázkovou kancelář ideální, pokud 500 jednotek bude vsazeno na výhru domácích, 300 jednotek na remízu a 200 jednotek na výhru hostů. V takovém případě, ať zápas dopadne jakýmkoli výsledkem, sázková kancelář bude v zisku 100 jednotek.

Výsledek	1	0	2
Kurz	1,8	3	4,5
Vsazeno	500	300	200
Výplata	900	900	900

Tabulka 14: Shrnutí vkladů a výplat v případě ideálního rozložení sázek

Sázkaři však můžou vsadit i v jiném poměru, např. 800 na výhru domácích, 100 na remízu a 100 na výhru hostů.

Výsledek	1	0	2
Kurz	1,8	3	4,5
Vsazeno	800	100	100
Výplata	1440	300	450

Tabulka 15: Shrnutí vkladů a výplat v případě jiného rozložení sázek

V tomto případě, pokud vyhrají domácí, bude sázková kancelář ve ztrátě 440 jednotek. Pokud by se však zápas opakoval mnohokrát, pak by střední hodnota výplaty $E(X)$, kde X je výplata sázkové kanceláře, opět byla 900.

$$E(X) = 0{,}5 \cdot 1440 + 0{,}3 \cdot 300 + 0{,}2 \cdot 200 = 900 \tag{6.2}$$

Ve skutečnosti se sice neopakuje jeden zápas několikrát, ale každý den se hrají stovky zápasů a z toho plyne, že ve střední hodnotě $E(X)$ jsou sázkové kanceláře v plusu.

6.2 Systém sázení

Existuje mnoho různých strategií jakým způsobem sázet. Zde budou použity dvě strategie: Flat betting a X procent na kolo.

6.2.1 Flat betting

V této strategii je vložen na každou sázku stejný vklad, např. 1 jednotka. Výhodou tohoto modelu pro sázkaře je, že nemusí řešit při každé sázce, kolik má vsadit a pořád sází stejně. Více o tomto systému v [10].

6.2.2 X procent na kolo

Druhou strategií sázení, která je použita v této práci, je strategie, v níž se v každém kole vsadí stejné procento z banku peněz, který je k dispozici, a následně se v kole rozdělí rovnoměrně mezi všechny zápasy, na které se bude sázet. Pro ukázku, když na každé kolo se bude brát 10 % z banku a na začátku bude k dispozici 1 000 jednotek, tak v prvním kole bude na sázky 100 jednotek. Pokud v tomto kole se bude sázet na dva zápasy, pak na každý zápas bude vsazeno 50 jednotek.

6.3 Kurzy

Sázkové kanceláře vypisují na jednotlivá fotbalová utkání kurzy o_i dle svých pravděpodobností q_i a svých marží ζ. Kurzy se tak u různých sázkových kanceláří většinou mírně liší. Nemůžou se však lišit příliš, protože v takovém případě by sázkař mohl vsadit u jedné společnosti na výhru jednoho týmu, u druhé na výhru jeho soupeře a u třetí na remízu a potom by sázkař vydělal, aniž by záleželo na výsledku utkání. V realitě se občas i taková možnost sázení naskytne.

V této práci jsou použity kurzy z internetové stránky www.oddsportal.com [B]. Na této stránce jsou k dispozici kurzy na různé sporty, jako jsou fotbal, hokej, tenis atd. Kurzy jsou k dispozici několik let dozadu. Pro tuto práci jsou důležité kurzy z fotbalových lig ze sezóny 2013/14. Tyto kurzy nejsou od jedné sázkové kanceláře, ale jedná se o průměrné kurzy z vypsaných kurzů 18 „prémiových“ sázkových kanceláří (10Bet, 1xbet, bet-at-home, bet365, Betadonis, Betrally, Betsafe, BetVictor, Betway, bwin, MarathonBet, Matchbook, Pinnacle Sports, Tempobet, TonyBet, Unibet, Winlinebet a Winner). Dále je možno k tomu přidat kurzy i z jiných sázkových kanceláří, to v tomto případě nebylo zvoleno.

7 Ověření modelu

Pokud jsou známy kurzy sázkových kanceláří o_i na jednotlivá utkání a také jsou z modelu vypočtené pravděpodobnosti r_i, tak dalším krokem je výběr zápasů, na které se bude sázet. Sázet se bude na zápasy, kde alespoň pro jedno i (i = 1, 0, 2) je splněna následující nerovnice

$$r_i \cdot o_i > R, \tag{7.1}$$

kde je

r_i pravděpodobnost výsledku předpokládaná modelem (sázejícím),
o_i kurz,
R předem zvolený parametr pro sázení na zápas.

Parametr R lze volit např. 1,0; 1,1; 1,2; atd. Parametr R menší než 1 nemá smysl volit, protože tento parametr R udává minimální střední hodnotu výhry při sázce 1, pokud by pravděpodobnosti r_i odhadnuté modelem byly naprosto stejné jako skutečné neznámé pravděpodobnosti p_i. Teoreticky je nejlepší volit co největší R (1,7 a vyšší). Problémem je, že takových zápasů za sezónu je velmi málo.

7.1 Česká liga

7.1.1 Strategie Flat betting

V této kapitole bude ukázáno ověření Dixon - Colesova modelu použitého pro českou první ligu (kapitola 5.4) při sázení strategií Flat betting (kapitola 6.2.1) tj. vklad na všechny sázky je stejný. V tomto případě je vklad 1 jednotka. Sází se na vybrané zápasy 6. - 30. kola Gambrinus ligy v sezóně 2013/2014 podle rovnice (7.1).

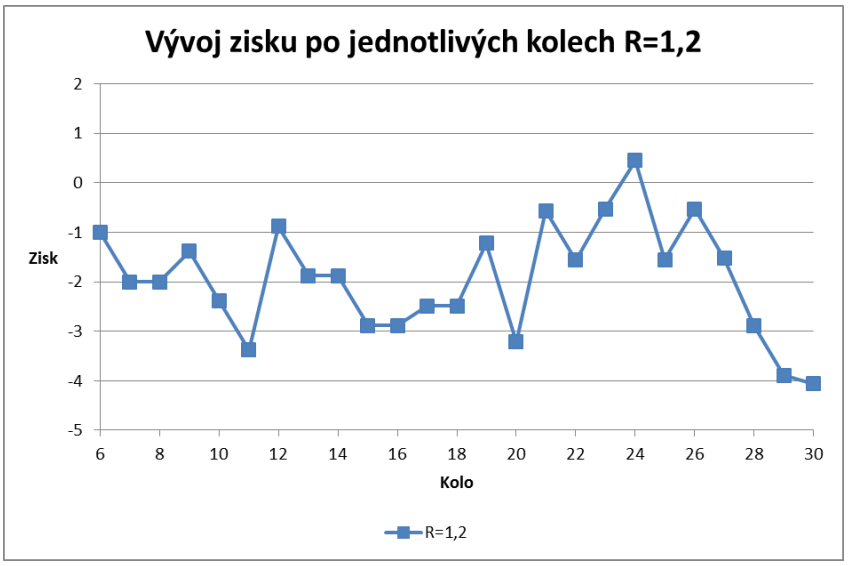
V následující tabulce jsou vypsány všechny zápasy, na které se bude sázet při daném parametru R = 1,2. Ve sloupcích Kurzy jsou kurzy sázkových kanceláří postupně na výhru domácích, remízu a výhru hostů. Ve sloupcích Vsazeno je opět řazení výhra domácích, remíza a výhra hostů. V jednotlivých políčkách je 1, pokud je na tu možnost vsazeno, a 0, pokud není vsazeno. V sloupci Výhra je výhra ze sázek na jednotlivé zápasy. Pokud je 0, tak sázka nevyšla. Pokud je tam číslo větší než 0, tak sázka byla úspěšná a bylo vyhráno právě tolik jednotek.

Kolo	Domácí	Hosté	Výsledek			Kurzy		Vsazeno			Výhra
6	FK Teplice	FC Vysočina Jihlava	4	2	1,55	3,92	5,9	0	0	1	**0**
7	Bohemians 1905	1.SC Znojmo FK	2	0	1,87	3,31	4,18	0	0	1	**0**
9	Bohemians 1905	SK Sigma Olomouc	0	2	2,61	3,23	2,62	0	0	1	**2,62**
9	FK Mladá Boleslav	FC Vysočina Jihlava	3	1	1,55	3,94	5,77	0	0	1	**0**
10	FC Vysočina Jihlava	FC Viktoria Plzeň	1	2	6,74	4,13	1,47	1	0	0	**0**
11	Bohemians 1905	FC Vysočina Jihlava	0	0	2,15	3,28	3,29	0	0	1	**0**
12	FC Vysočina Jihlava	FK Baumit Jablonec	3	2	3,5	3,34	2,04	1	0	0	**3,5**
13	FC Zbrojovka Brno	FC Vysočina Jihlava	1	0	2,02	3,33	3,56	0	0	1	**0**
15	1.FK Příbram	FC Vysočina Jihlava	5	0	2,07	3,26	3,55	0	0	1	**0**
17	SK Sigma Olomouc	SK Slavia Praha	5	1	2,39	3,13	2,98	1	0	0	**2,39**
17	FC Slovan Liberec	FK Mladá Boleslav	2	2	2,59	3,18	2,67	1	0	0	**0**
19	1.SC Znojmo FK	FK Baumit Jablonec	4	0	3,27	3,25	2,18	1	0	0	**3,27**
19	SK Sigma Olomouc	FK Mladá Boleslav	1	1	2,73	3,25	2,49	1	0	0	**0**
20	SK Slavia Praha	1.SC Znojmo FK	2	1	1,6	3,7	5,55	0	0	1	**0**
20	Bohemians 1905	FC Slovan Liberec	1	0	2,63	3,11	2,71	0	0	1	**0**
21	1.SC Znojmo FK	Bohemians 1905	0	0	2,14	3,14	3,51	1	0	0	**0**
21	SK Slavia Praha	FC Vysočina Jihlava	1	2	1,62	3,64	5,66	0	0	1	**5,66**
21	FK Teplice	FK Mladá Boleslav	0	1	2,4	3,18	2,94	1	0	0	**0**
22	1.FK Příbram	1.FC Slovácko	3	2	1,95	3,35	3,84	0	0	1	**0**
23	FC Slovan Liberec	FC Baník Ostrava	3	2	2,02	3,22	3,78	1	0	0	**2,02**
23	1.SC Znojmo FK	FC Zbrojovka Brno	1	1	2,38	3,13	2,96	1	0	0	**0**
23	FC Vysočina Jihlava	FK Mladá Boleslav	2	1	3,01	3,17	2,34	1	0	0	**3,01**
23	SK Slavia Praha	FK Baumit Jablonec	0	0	1,96	3,44	3,67	0	0	1	**0**
24	1.FK Příbram	FC Slovan Liberec	3	1	2,19	3,26	3,23	0	0	1	**0**
24	FC Baník Ostrava	1.FC Slovácko	0	1	1,91	3,35	3,98	0	0	1	**3,98**
24	FK Baumit Jablonec	FK Dukla Praha	1	4	2,52	3,24	2,72	1	0	0	**0**
25	1.SC Znojmo FK	FC Baník Ostrava	0	4	2,82	3,13	2,51	1	0	0	**0**
25	SK Sigma Olomouc	1.FK Příbram	0	0	2,38	3,15	2,99	1	0	0	**0**
26	1.FC Slovácko	SK Sigma Olomouc	3	1	3,02	3,23	2,3	1	0	0	**3,02**
26	1.FK Příbram	1.SC Znojmo FK	1	1	1,55	3,79	6,14	0	0	1	**0**
27	SK Sigma Olomouc	FC Baník Ostrava	2	3	2,38	3,15	2,99	1	0	0	**0**
28	Bohemians 1905	FK Dukla Praha	3	2	1,99	3,37	3,67	0	0	1	**0**
28	1.FK Příbram	FK Baumit Jablonec	3	0	1,68	3,8	4,62	0	0	1	**0**
28	FC Baník Ostrava	FC Vysočina Jihlava	3	1	1,71	3,6	4,77	0	0	1	**0**
28	FC Slovan Liberec	FK Teplice	2	1	2,64	3,24	2,59	1	0	0	**2,64**
29	SK Slavia Praha	1.FC Slovácko	1	1	1,73	3,61	4,51	0	0	1	**0**
30	Bohemians 1905	FC Viktoria Plzeň	0	0	2,55	3,4	2,55	0	0	1	**0**
30	FC Baník Ostrava	SK Slavia Praha	2	0	2,83	3,34	2,36	1	0	0	**2,83**
30	FC Slovan Liberec	SK Sigma Olomouc	1	1	2,45	3,46	2,64	1	0	0	**0**

Tabulka 16: Seznam vsazených zápasů pro R = 1,2

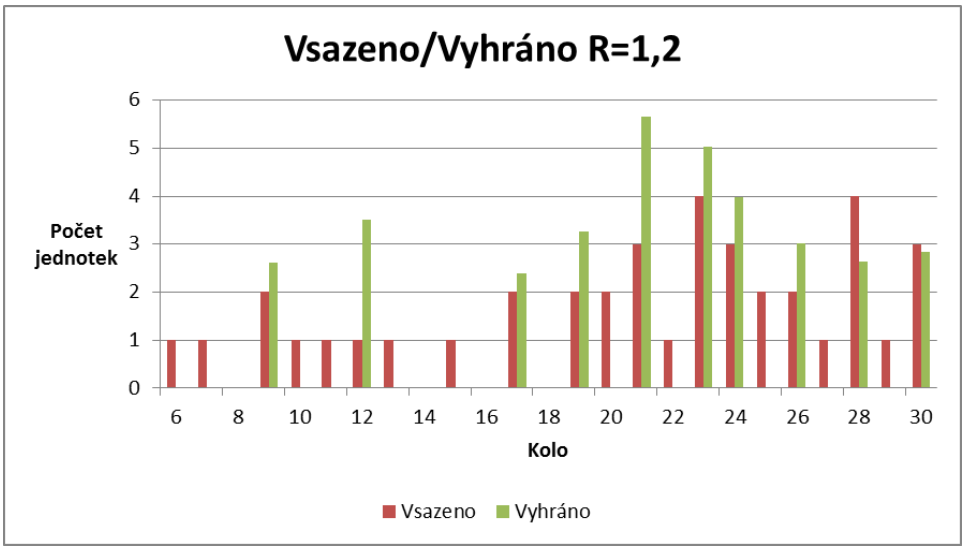
Ztráta v tomto případě je 4,06.

V následujícím grafu je pro hodnotu parametru R = 1,2 zobrazen vývoj celkového zisku po jednotlivých kolech.



Obrázek 10: Vývoj zisku po jednotlivých kolech pro R = 1,2

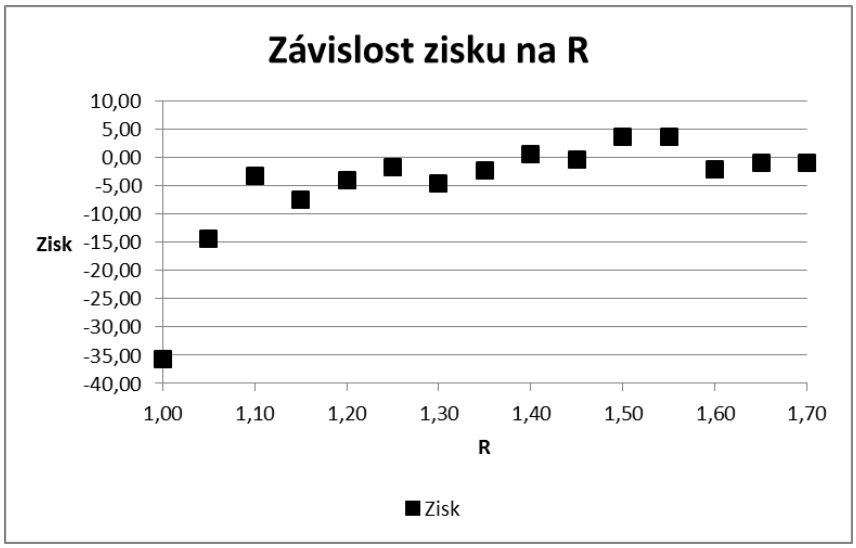
V dalším grafu jsou zobrazeny vsazené částky a výhry v jednotlivých kolech pro hodnotu parametru R = 1,2.



Obrázek 11: Vsazené a vyhrané částky pro R = 1,2

Pro hodnotu parametru R = 1,2 se během celé sezóny pouze po 24. kole dostane sázkař, sázející dle modelu, do kladných čísel a to díky úspěšným kolům 21, 23 a 24. Od 24. kola jsou všechna kola ztrátová s výjimkou 26. kola, a proto sázkař, který by sázel dle modelu při hodnotě parametr R = 1,2, by byl na konci sezóny ve ztrátě 4,06 jednotky.

V následujícím grafu je shrnuta závislost výše zisku na parametru R.



Obrázek 12: Závislost zisku na parametru R

V následující tabulce je shrnuto porovnání počtu sázek, počtu výher a zisku pro různé hodnoty parametru R.

R	Počet sázek	Vsazená částka	Vítězné sázky	Vyhraná částka	Zisk
1,00	163	163	55	127,33	-35,67
1,05	107	107	40	92,62	-14,38
1,10	71	71	26	67,70	-3,30
1,15	55	55	16	47,59	-7,41
1,20	39	39	11	34,94	-4,06
1,25	28	28	8	26,41	-1,59
1,30	22	22	5	17,51	-4,49
1,35	17	17	4	14,68	-2,32
1,40	12	12	3	12,66	0,66
1,45	9	9	2	8,68	-0,32
1,50	5	5	2	8,68	3,68
1,55	5	5	2	8,68	3,68
1,60	2	2	0	0,00	-2,00

Tabulka 17: Porovnání parametru R

Z grafu i tabulky je patrné, že největšího zisku je dosaženo, pokud R je nastaveno kolem 1,5, a to zásluhou dvou vítězných sázek z pěti vsazených zápasů při této hodnotě parametru R. Jedná se o zápas 21. kola, kdy bylo vsazeno na výhru Jihlavy s kurzem 5,66 proti domácí Slavii Praha, a druhá vítězná sázka byla v zápase 26. kola mezi Slovákem a Olomoucí, kde bylo vsazeno na výhru Slovácka s kurzem 3,02.

7.1.2 Strategie X procent na kolo

Druhou strategií sázení, kterou se bude ověřovat Dixon - Colesův model pro českou Gambrinus ligu (kapitola 5.4), je strategie, ve které se bude sázet určité procento z peněz, které jsou k dispozici před začátkem kola (kapitola 6.2.2).

Vzhledem k tomu, že výběr zápasů nezávisí na strategii sázení, ale jen na modelu, kurzech sázkových kanceláří a parametru R, tak zápasy, na které se bude sázet, jsou při daném parametru R stejné jako u předchozí strategie (kapitola 7.1.1).

V této strategii je kromě parametru R potřeba určit i kolik procent z banku se bude sázet na každé kolo. Počáteční bank je vždy 1 000 jednotek. V následující tabulce je zisk pro různá R a procenta.

R / Procenta	1 %	5 %	10 %	20 %	50 %	100 %
1,00	-53	-244	-438	-703	-973	-1000
1,05	-20	-110	-234	-483	-931	-1000
1,10	-18	-101	-224	-488	-952	-1000
1,15	-76	-341	-587	-861	-999	-1000
1,20	-57	-293	-490	-785	-996	-1000
1,25	-22	-128	-280	-583	-983	-1000
1,30	-28	-149	-309	-601	-977	-1000
1,35	-13	-83	-205	-485	-962	-1000
1,40	25	89	94	-92	-870	-1000
1,45	15	50	41	-111	-791	-1000
1,50	36	164	285	389	-163	-1000
1,55	36	164	285	389	-163	-1000
1,60	-20	-98	-190	-360	-750	-1000

Tabulka 18: Zisk v závislosti na R a procentech

Z tabulky je patrné, že nemá cenu volit 50 % či více. V takových případech dojde k velkým ztrátám. Při volbě 20 % lze dosáhnout největšího zisku, avšak oproti nižším procentům je tu velká citlivost na hodnotu parametru R. Proto bude lepší volit 5 - 10 %.

V další tabulce je shrnuto porovnání počtu sázek, počtu výher a zisku pro volbu 5 % a různé hodnoty parametru R.

	Počet sázek	Vsazená částka	Vítězných sázek	Vyhraná částka	Zisk
1,00	142	1075,19	55	830,73	-244,46
1,05	101	1181,76	40	1071,42	-110,34
1,10	68	1165,03	26	1063,87	-101,16
1,15	52	924,88	16	583,70	-341,19
1,20	39	898,84	11	630,44	-268,41
1,25	28	823,38	8	695,81	-127,57
1,30	22	605,55	5	456,79	-148,76
1,35	17	592,71	4	509,22	-83,49
1,40	12	550,75	3	640,02	89,27
1,45	9	364,60	2	415,04	50,43
1,50	5	272,97	2	436,88	163,91
1,55	5	272,97	2	436,88	163,91
1,60	2	97,50	0	0,00	-97,50

Tabulka 19: Porovnání parametru R pro strategii 5 %

7.1.3 Srovnání strategií Flat betting a X procent na kolo

Zatímco při strategii Flat betting není nutné se zabývat tím, jakou částku by se mělo na jednotlivé zápasy vsázet, protože se na každý zápas vsadí vždy stejná částka, tak při druhé strategii je nutné sázku na zápasy v každém kole přepočítávat. To má pak vliv i na zisk. Například pokud by se zvolilo sázení 5 % na kolo, tak by se sázkař dostal do zisku v sezóně 2013/2014 i pro hodnoty parametru R = 1,4 či 1,45, což by se mu při strategii Flat betting nepovedlo. Na druhou stranu pokud by zvolil sázkař strategii 50 % na kolo, tak by byl ve ztrátě pro jakékoliv hodnoty parametru R.

Z toho vyplývá, že správná volba strategie je při sázení důležitá a může vylepšit celkový zisk ze sázení nezávisle na kvalitě odhadu výsledků.

7.2 Ostatní ligy

V této kapitole bude ukázáno ověření Dixon - Colesova modelu pro zbylé ligy, které byly odhadovány v kapitole 5. Konkrétně se jedná o španělskou, italskou a anglickou ligu. Z důvodu, že ve všech těchto ligách se některé zápasy jednotlivých kol dohrávaly po několika dalších kolech, tak je k ověření modelu použita pro tyto ligy jen strategie Flat betting (kapitola 6.2.1) tj. vklad na všechny sázky je stejný, v tomto případě bude vždy 1 jednotka.

7.2.1 Španělská liga

V následující tabulce je shrnuto porovnání počtu sázek, počtu výher a zisku pro různé hodnoty parametru R ve španělské lize.

R	Počet sázek	Vsazená částka	Vítězných sázek	Vyhraná částka	Zisk
1,00	346	346	101	386,07	40,07
1,05	260	260	65	282,86	22,86
1,10	195	195	44	216,91	21,91
1,15	151	151	33	179,27	28,27
1,20	117	117	22	121,56	4,56
1,25	92	92	13	81,96	-10,04
1,30	72	72	9	63,53	-8,47
1,35	54	54	7	57,77	3,77
1,40	45	45	6	52,91	7,91
1,45	31	31	5	47,81	16,81
1,50	31	31	5	47,81	16,81
1,55	25	25	4	33,97	8,97
1,60	17	17	2	22,6	5,6
1,65	13	13	2	22,6	9,6
1,70	12	12	2	22,6	10,6
1,75	10	10	1	13,4	3,4
1,80	7	7	1	13,4	6,4
1,85	6	6	1	13,4	7,4

Tabulka 20: Porovnání parametru R španělská liga

Španělská liga dopadla pro sázejícího dle modelu velmi příznivě a nejlépe ze všech 4 zkoumaných lig. Španělská liga je zvláštní v tom, že v ní hrají dva dominantní týmy FC Barcelona a Real Madrid. V sezóně 2013/2014 se k nim přidal i tým Atlético Madrid. Tyto týmy vyhrávají většinu svých zápasů a často i větším rozdílem. Proto jsou na jejich výhru vypisované velmi nízké kurzy a naopak na jejich soupeře vysoké. Proto když se podaří odhadnout zápas, ve kterém jeden z těchto týmu nevyhraje, tak je z této sázky vysoký zisk.

Toto se stalo například v 36. kole, kdy ztratily všechny tři týmy. Atlético Madrid prohrálo na hřišti Levante, když na Levante byl vypsán kurz 13,4. Především díky tomuto zápasu byl sázkař, sázející dle modelu, v zisku pro většinu hodnot parametru R, protože na tento zápas bylo vsazeno i při hodnotě parametru R = 1,85. Při nižších hodnotách parametru R bylo v tomto kole vsazeno i na remízy Barcelony s Getafe a Realu Madrid s Valencií s kurzy 16,01, respektive 9,72. Hlavně kvůli těmto 3 zápasům bylo 36. kolo pro sázejícího dle modelu velmi úspěšné. Například pro hodnotu parametru R = 1 byl zisk v tomto kole 28,87 jednotek, když bylo vsazeno 14 jednotek a vyhráno 42,87 jednotek. V případě vynechání zmíněných 3 zápasů, potom by toto kolo bylo naopak ztrátové.

7.2.2 Italská liga

V následující tabulce je shrnuto porovnání počtu sázek, počtu výher a zisku pro různé hodnoty parametru R v italské lize.

R	Počet sázek	Vsazená částka	Vítězných sázek	Vyhraná částka	Zisk
1,00	331	331	90	272,99	-58,01
1,05	234	234	57	188,71	-45,29
1,10	175	175	40	134,8	-40,2
1,15	131	131	26	92,46	-38,54
1,20	102	102	18	64,19	-37,81
1,25	73	73	16	59,35	-13,65
1,30	60	60	13	44,12	-15,88
1,35	49	49	11	39,82	-9,18
1,40	39	39	8	30,67	-8,33
1,45	33	33	6	22,94	-10,06
1,50	25	25	4	15,11	-9,89
1,55	19	19	2	9,4	-9,6
1,60	14	14	2	9,4	-4,6
1,65	10	10	1	3,22	-6,78
1,70	9	9	1	3,22	-5,78
1,75	8	8	1	3,22	-4,78
1,80	6	6	0	0	-6

Tabulka 21: Porovnání parametru R italská liga

Sázkař sázející dle modelu na italskou ligu v sezóně 2013/2014 byl ve ztrátě pro všechny hodnoty parametru R. Toto se stalo pouze v této lize. Nejlépe sázkař dopadl, pokud volil parametr R = 1,6. V takovém případě sázkař prohrál za sezónu 4,6 jednotky. Celkově vsadil v sezóně 14 sázek, ale pouze 2 byly úspěšné. Konkrétně se jedná o sázku na zápas 22. kola mezi domácím US Sassuolo a hostujícím mužstvem Hellas Verona, kdy bylo vsazeno na Hellas s kurzem 3,22 a Hellas vyhrál 2:1. Druhý úspěšný zápas byl v 37. kole, kdy domácí mužstvo Atlanta Bergamo porazilo AC Milán 2:1 a bylo na něj vsazeno s kurzem 6,18.

7.2.3 Anglická liga

V následující tabulce je shrnuto porovnání počtu sázek, počtu výher a zisku pro různé hodnoty parametru R v anglické lize.

R	Počet sázek	Vsazená částka	Vítězných sázek	Vyhraná částka	Zisk
1,00	362	362	93	345,58	-16,42
1,05	273	273	68	268,91	-4,09
1,10	191	191	37	147,34	-43,66
1,15	122	122	19	86,55	-35,45
1,20	83	83	11	57,74	-25,26
1,25	60	60	7	34,13	-25,87
1,30	43	43	5	28,16	-14,84
1,35	32	32	3	16,41	-15,59
1,40	22	22	2	12,56	-9,44
1,45	10	10	1	5,07	-4,93
1,50	10	10	1	5,07	-4,93
1,55	10	10	1	5,07	-4,93
1,60	5	5	1	5,07	0,07
1,65	5	5	1	5,07	0,07
1,70	2	2	0	0	-2

Tabulka 22: Porovnání parametru R anglická liga

V anglické lize v sezóně 2013/2014 se sázkař sázející dle modelu dostal do zisku pouze v případě, že volil hodnotu parametru R kolem 1,6. V tomto případě byl celkový zisk 0,07 jednotky. Sázkař vsadil během sezóny na 5 zápasů a jeden zápas byl výherní. Konkrétně byla správně odhadnuta výhra Chelsea v Liverpoolu s kurzem 5,07.

7.3 Shrnutí

Pro sezónu 2013/2014 vyšly nejlépe předpovědi výsledků a následný souboj se sázkovou kanceláří pro španělskou ligu. Naopak v italské lize se model pro žádnou hodnotu parametru R nedostal do zisku. Z odhadů těchto čtyř lig lze předpokládat, že nejlepší hodnota parametru R je někde mezi 1,4 a 1,6. Pro jistotu tohoto tvrzení a i zpřesnění by však bylo nutné odhadnout mnohem více lig i ročníků, protože při takto nastaveném parametru R bylo málo zápasů, na které bylo vsazeno, a tudíž proběhlo málo pozorování během jedné sezóny.

Obecně lze říct, že sázkové kanceláře velmi dobře odhadují pravděpodobnosti výsledků jednotlivých zápasů a je velmi těžké sázkové kanceláře porazit.

8 Závěr

Cílem práce bylo najít matematické modely pro odhadování sportovních výsledků a následně je ověřit v souboji se sázkovými kancelářemi.

Pro odhadování či předpovídání výsledků byl použit Maherův model z roku 1982 a následně i vylepšení tohoto modelu od dvojice Dixon a Coles z roku 1997. Tento vylepšený model byl následně podroben souboji se sázkovými kancelářemi. Ukázalo se, že sázkové kanceláře odhadují pravděpodobnosti výsledků zápasů poměrně přesně a je těžké je porazit pomocí matematického modelu. Avšak v několika málo případech vyšla možnost vítězství nad sázkovou kanceláří pro některé hodnoty parametru R, což je předem určená hodnota, která musí být menší než součin kurzu a odhadnuté pravděpodobnosti výsledku utkání modelem, aby se vsadilo na zápas.

Nejlépe si model vedl ve španělské lize, kde se sázkař sázející podle modelu mohl dostat do zisku při různých hodnotách parametru R. Na druhou stranu v italské lize se sázkař nedostal do zisku při žádné hodnotě parametru R. Ze čtyř zkoumaných lig lze usuzovat, že parametr R je nejlepší volit mezi 1,4 až 1,6. V takovém případě většinou sázkař dosahoval největšího zisku či nejmenší ztráty. Důležité je však připomenout, že byl odhadován jen jeden ročník ve čtyřech soutěžích. Pro optimalizování parametru R a možnost tvrzení, ve které soutěži si model vede nejlépe, by bylo nutné odhadnout více sezón i soutěží a nejlépe za stejného odhadu výsledků sázkových kanceláří. V tomto však nastává problém, protože i sázkové kanceláře vylepšují své modely a zpřesňují odhady výsledků.

Fotbal je nejrozšířenější sport na světě a patří mezi sporty, na které se nejvíce sází. Proto se sázkové kanceláře na fotbal zaměřují a mají velmi dobré odhady výsledků fotbalových utkání. Dalším rozšířením práce by tedy mohlo být zahrnutí jiných sportů, např. hokej, florbal a futsal, ve kterých by si model proti sázkové kanceláři mohl vést lépe.

9 Literatura a zdroje dat

9.1 Seznam literatury

1 - MAHER, M. J. *Modelling association fotball scores*. Statistica Neerlandica. 1983, č. 3, s. 109-118.

2 - DIXON, Mark a Stuart COLES. *Modelling Association Fotball Scores and Inefficiencies in the Fotball Betting Market*. Journal of the Royal Statistical Society. 1997, č. 2, s. 265-280.

3 - CYHELSKÝ, Lubomír. *Elementární statistická analýza*. 2. vyd. Praha: Management Press, 2001. ISBN 80-7261-003-1.

4 - REIF, Jiří. *Metody matematické statistiky*. Plzeň: Západočeská univerzita, 2004, s. 61-63. ISBN 80-7043-302-7.

5 - Pravděpodobnost a statistika HYPERTEXTOVĚ. *P-hodnota.* [online]. 2014 [cit. 2015-04-23]. Dostupné z: http://home.zcu.cz/~friesl/hpsb/phodn.html

6 - ABDI, Herve. The University of Texas at Dallas. *The Bonferonni and Šidák Corrections for Multiple Comparisons.* [online]. 2007 [cit. 2015-04-23]. Dostupné z: http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf

7 – Česká televize. *Juniorská fotbalová liga už má konkrétní obrysy.* [online]. 2012 [cit. 2015-04-23]. Dostupné z: http://www.ceskatelevize.cz/sport/fotbal/171354-juniorska-fotbalova-liga-uz-ma-konkretni-obrysy/

8 - Microsoft. *GRG Algorithm.* [online]. © 2015 [cit. 2015-04-23]. Dostupné z: http://support.microsoft.com/en-us/kb/82890

9 - MAREK, Patrice. *Přednášky ZTI.* [online]. 2014 [cit. 2015-04-23]. Dostupné z: http://www.kma-old.zcu.cz/main.php?KMAfile=./CLENOVE/main.php&DRC=./STRUCTURE/06_IT/02_www/&DRL=CZ&DROF=0&nick=PaMar&kam=vyuka.php

10 - Kurzové sázení. *Flat betting*. [online]. 2015 [cit. 2015-04-23]. Dostupné z: http://www.kurzovesazeni.com/stabilni-vyse-sazek-flat-betting/

9.2 Zdroj dat

A - EuroFotbal. [online]. © 2007 [cit. 2015-04-23]. Dostupné z: http://www.eurofotbal.cz/

B - Odds Portal. [online]. © 2008-2015 [cit. 2015-04-23]. Dostupné z: http://www.oddsportal.com/

Příloha

Seznam přiložených souborů

BP_Spacek.pdf – elektronická verze bakalářské práce

Poisson.xlsx – testování, zda se počty gólů ve fotbalových utkáních řídí Poissonovým rozdělením pravděpodobnosti

Maher.xlsm – odhadování pravděpodobnosti výsledků zápasů pomocí Maherova modelu

CZEDixon.xlsx – odhadování pravděpodobnosti výsledků zápasů v české lize v sezóně 2013/2014 pomocí Dixon - Colesova modelu a shrnutí sázení modelu proti sázkové kanceláři

ENGDixon.xlsx – odhadování pravděpodobnosti výsledků zápasů v anglické lize v sezóně 2013/2014 pomocí Dixon - Colesova modelu a shrnutí sázení modelu proti sázkové kanceláři

ITADixon.xlsx – odhadování pravděpodobnosti výsledků zápasů v italské lize v sezóně 2013/2014 pomocí Dixon - Colesova modelu a shrnutí sázení modelu proti sázkové kanceláři.

SPADixon.xlsx – odhadování pravděpodobnosti výsledků zápasů ve španělské lize v sezóně 2013/2014 pomocí Dixon - Colesova modelu a shrnutí sázení modelu proti sázkové kanceláři