ZÁPADOČESKÁ UNIVERZITA V PLZNI
FAKULTA APLIKOVANÝCH VĚD
KATEDRA MATEMATIKY

Diplomová práce

Modelování a odhadování výsledků hokejových utkání

Plzeň, 2016 Bc. Ivana Gabrišková

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a výhradně s použitím literatury a pramenů uvedených v seznamu.

V Plzni dne 12. srpna 2016 ……

Ivana Gabrišková

Poděkování

Ráda bych poděkovala vedoucímu mé diplomové práce Ing. Patrice Markovi, Ph.D. za vedení, odborné rady a čas, který mi věnoval při zpracování této práce.

Dále bych chtěla poděkovat své mamince a rodině, která mě během mého studia neustále podporovala a bez nichž bych to nezvládla.

V neposlední řadě patří také poděkování Ing. Tomáši Pakostovi za cenné připomínky, podporu a trpělivost, kterou se mnou měl během mého studia.

Zadání!

Abstrakt

Tato diplomová práce se zabývá modelováním a odhadováním výsledků hokejových utkání a poté využitím odhadů při sázení proti sázkovým kancelářím. Práce popisuje modely používané pro odhadování výhry domácího týmu, remízy nebo výhry hostujícího týmu. Práce se věnuje Dixon-Colesovým modelům upraveným podle článku [3]. Těmito modely jsou odhadovány výsledky zápasů české, polské a národní hokejové ligy v sezóně 2015/2016. Modely jsou porovnané s triviálními schématy sázení.

Klíčová slova: Hokej, Poissonovo rozdělení, odhad hokejových výsledků, sázení

Abstract

This diploma thesis concerns with modeling and estimating results of hockey matches and use of estimation for betting against bookmakers. The thesis describes models used for estimating winning of home team, a draw or away team win. The thesis deals about Dixon-Coles models arranged according to the article [3]. Results of matches of Czech, Polish and National Hockey League in season 2015/2016 are estimated by using this models. Models are compared with trivial schemas of betting.

Key words: Hockey, Poisson distribution, estimate of hockey results, betting

OBSAH

1 Úvod

Mnoho lidí považuje sázení za hazard a mohou se při něm dostat do osobního bankrotu. Sázení naslepo bez jakýchkoliv zkušeností je velmi rizikové, a proto by bylo vhodné si vytvořit strategii sázení.

Pro tyto účely můžeme využít pravděpodobnostní modely a statistické metody. V této práci se budeme zabývat zpracováním hokejových zápasů. Data jsou popsána v první části. Za pomoci získané a zpracované historie těchto dat pro tři vybrané ligy, se budeme snažit odhadovat výsledky budoucích zápasů, ve smyslu výhry domácích, remízy a výhry hostů.

K tomuto odhadu nejprve použijeme pro českou ligu Maherův model určený pro fotbalová data, který využijeme jako základ pro odhadování výsledků sportovních utkání. Ověříme, zda tento model je stejně vhodný pro hokejová data jako pro fotbalová. Dalšími uvažovanými modely budou dvojnásobný Poissonovo model (DP) a dvourozměrný Poissonovo model (BP). Tyto modely však podhodnocují pravděpodobnosti remízy, proto využijeme ještě dalšího modelu, který je rozšířením uvedených předcházejících modelů.

Uvedené modely mezi sebou porovnáme a určíme který z nich je nejvhodnější. Na základě tohoto porovnání najdeme vhodnou strategii, resp. model, podle kterého budeme určovat, zda vsadit na výhru domácích, remízu či výhru hostujích týmů.

Na závěr zjistíme, jak výnosná strategie bude a na které zápasy sázet.

2 DATA

Pro zpracování diplomové práce budou využita data hokejových zápasů, které lze dohledat na webových stránkách Sfstats.net a Oddsportal.com. Data obsahují výsledky zápasů a sázky na ně. V této práci jsou použité údaje ze zdroje [A] pro vybrané tři hokejové ligy, a to Extraligy hrané v ČR, dále pak z polské ligy a národní hokejové ligy NHL.

Používáme jen výsledky za základní hrací dobu 60 min, což znamená, že hra může skončit za nerozhodného stavu. Samozřejmě za stejnou dobu se použijí i kurzy na výsledky ve hře. Pro naše analýzy se používají pravidelné sezónní zápasy kromě zápasů play off a play out. Hlavním důvodem tohoto omezení je, že tyto zápasy jsou hrané v rámci vyřazovacího turnaje, který se hraje na čtyři vítězná utkání, na rozdíl od zápasů v základní části. Dalším důvodem je, že zápasy v play off se mohou hrát různými taktikami oproti obvyklým zápasům odehraným v sezóně.

Data jsou rozdělena do dvou částí. První část zahrnuje všechny zápasy od sezóny 2009/2010 až do sezóny 2014/2015, druhá část obsahuje všechny zápasy ze sezóny 2015/2016 a používá se na předvídací schopnost modelu. Kurzy jsou ze stejného zdroje jako výsledky zápasů ([A]), a v tomto zdroji autoři uvádí kurzy získané z průměru pěti vybraných sázkových kanceláří, a to sportingbet, gamebookers, bwin, expekt a bet365.

Po důkladném zkoumání dat byly objevené chybějící zápasy a kurzy nebo byl tým uveden v zápase jako domácí, ale ve skutečnosti hrál jako hostující a naopak. Tyto nedostatky v datech bylo nutné dohledat ze zdroje [B]. Z tohoto zdroje byly vypsané zápasy a dostupné kurzy z vybraných sázkových kanceláří do souboru *chybějící kurzy.xlsx*, kde se z nich počítá průměr, který je poté zaznamenaný v souboru *data.xlsx* na listu dané ligy. Pro českou ligu chybělo celkem 25 kurzů, pro polskou ligu 69 a pro NHL 59.

V dalších kapitolách uvažujeme, že jeden hrací den je jedno „kolo“, z toho důvodu, že jeden hrací den může mít různý počet zápasů. Tento standard využijeme pro celé zpracování, ale například i pro váhovou funkci viz podkapitola 6.5.1.

2.1 Extraliga (CZE)

Struktura sezóny

Extraliga je nejvyšší hokejová soutěž v České republice. V této lize hraje každý rok 14 týmů. Základní část se hraje od září do února, kdy se každý tým s každým utká dvakrát doma a dvakrát venku.

Poté následuje rozšířené play off. Nejprve se hraje předkolo play off, ve kterém se utkají sedmý s desátým a osmý s devátým umístěným klubem v tabulce po základní části. Hraje se na 3 vítězné zápasy. Poté následuje klasické play off, kdy hraje první s hůře umístěným vítězem předkola, druhý s lépe umístěným vítězem předkola, třetí s šestým a čtvrtý s pátým. Dále se hraje na 4 vítězné zápasy, postoupí 4 kluby do semifinále a vítězní semifinalisté do finále. Vítěz finále se stává Mistrem extraligy, konečné druhé až desáté místo je určeno úspěšností týmů v play off.

Jedenáctý až čtrnáctý tým tabulky po základní části hrají play out skupinu o udržení v dané lize, ve které se každý s každým utká dvakrát, přičemž se započítávají body ze základní části. Týmy, které skončí ve skupině o udržení na posledních dvou místech, hrají baráž o udržení v extralize se dvěma vítězi semifinále play off 1. ligy. Tato baráž

se hraje čtyřkolově každý s každým v rámci čtyřčlenné skupiny. Týmy, které se umístí na prvním a druhém místě na konci baráže, budou postupovat v následující sezóně do Extraligy.

Data

Dostupná data obsahují 2 184 výsledků za celou hrací dobu v Extralize, a to včetně všech sezónních utkání kromě zápasů play off a play out. Soubor obsahuje všechny výsledky ze sezóny 2009/2010 až do sezóny 2014/2015. Počet hracích dní v sezóně 2015/2016 bylo 74, tj. 74 „kol".

2.2 Ekstraliga (POL)

Struktura sezóny

V této lize hraje téměř každý rok různý počet týmů. V první části ligy hraje každý s každým čtyři zápasy (tj. 36 zápasů). Následně se liga rozdělí do dvou skupin. Prvních 6 týmů hraje tzv. *silnější skupinu*, každý tým celkem 10 zápasů (každý s každým jednou doma, jednou venku) o umístění v tabulce před play off. Zbylé 4 týmy hrají *slabší skupinu*. Ve slabší skupině se hraje dvakrát doma a dvakrát venku, každý s každým (dohromady každý tým odehraje ve slabší skupině 12 utkání). Dva nejlepší celky skupiny postupují do play off a poslední tým sestupuje do 1. ligy.

Následuje play off, ve kterém se hraje první kolo na 3 vítězné zápasy a následně čtvrtfinále, semifinále a finále hrané na 4 vítězné zápasy. V polské lize se hraje i o konečné třetí místo, a to na 2 vítězné zápasy.

Data

V dostupných datech od sezóny 2009/2010 do sezóny 2014/2015 hrál Ekstraligu různý počet týmů. Například v sezóně 2009/2010 hrálo 11 týmů, v sezóně 2010/2011 a 2014/2015 hrálo 10 týmů, a v sezónách 2011/2012 a 2012/2013 hrálo dokonce jen 8 týmů. Data obsahují celkem 1 169 zápasů od sezóny 2009/2010 až do sezóny 2014/2015. Co se týče kvality dat z pohledu kurzů, tak jich zde celkově chybělo 69, které bylo potřeba doplnit z jiného propracovanějšího zdroje [B]. Ale i v tomto zdroji některé kurzy chyběly, a tak byly dohledány ze zdroje [C]. Více o chybějících kurzech včetně zápasů, kterých se to týkalo lze nalézt v souboru *chybějící kurzy.xlsx* na listu *POL*.

2.3 NHL

NHL je nejprestižnější národní hokejovou ligou světa. Účastní se jí pouze týmy z USA a Kanady. Hraje ji 30 mužstev, která jsou rozdělena do dvou konferencí, východní a západní. Obě konference jsou ještě rozděleny celkem na čtyři divize, každá po dvou. Ve východní je v každé divizi osm týmů a v západní sedm.

Struktura sezóny

Sezóna je rozdělena na dvě části. První je tzv. základní část, v níž každé družstvo odehraje 82 utkání. Po skončení základní části nastává druhé kolo, tzv. play off. Do play off postupuje osm nejlepších družstev z každé konference. Poté se hrají série na 4 vítězná utkání.

Základní část NHL začíná první středu v říjnu a končí v polovině dubna. Poté následuje play off, jenž se hraje od dubna nejpozději do poloviny června. V základní části hraje každý tým 82 zápasů, a to 41 zápasů doma a 41 venku. V současnosti hraje ve východní konferenci každý tým 30 zápasů proti soupeřům z jejich divize (4 zápasy proti pěti týmům

ze své divize a 5 zápasů proti dvěma týmům ve své divizi), 24 zápasů proti týmům ze své konference (3 zápasy proti každému týmu z druhé divize ve své konferenci) a 28 utkání se čtrnácti zbývajícími celky (2 zápasy proti každému týmu ze západní konference). Naopak v západní konferenci si každé mužstvo zahraje celkem 29 utkání proti soupeřům z jejich divize (5 zápasů proti pěti soupeřům ve své divizi a 4 zápasy proti jednomu týmu ve své divizi), 21 utkání proti celkům ze své konference (3 zápasy proti každému týmu z druhé divize ve své konferenci) a 32 zápasů s šestnácti zbývajícími soky (2 zápasy proti každému týmu z východní konference). Takovýto rozpis sezóny je platný od sezóny 2013/14 a byl vytvořený proto, aby týmy ušetřily na cestování, které fyzicky zatěžovalo hráče.

Do play off postupují nejlepší tři týmy z každé divize (celkem tedy 12 celků). Dále dvě mužstva z každé konference, která budou mít nejvyšší počet bodů bez ohledu na divize (4 týmy). Tým s nejvyšším počtem bodů po základní části z obou konferencí obdrží Presidents' Trophy. V každé konferenci hraje play off 8 týmů. Vítězové divizí jsou do play off nasazeni na prvních dvou místech (i když nějaký nevítěz divize měl více bodů než vítěz divize, tak je výše nasazen vítěz divize), a zbylé týmy jsou nasazeny jako týmy na pátém až osmém místě.

Data

Data obsahují 6 865 zápasů od sezóny 2009/2010 do 2014/2015. Chybějících kurzů bylo celkem 59 a jsou doplněné v datovém souboru *data.xlsx* (na listu *NHL*), kurzy jednotlivých sázkových kanceláří jsou opět v souboru *chybějící kurzy.xlsx* na listu *NHL*.

3 Statistické pojmy a metody

V této části se zabýváme pojmy z pravděpodobnosti a statistiky, které jsou použité v dalších kapitolách.

3.1 Poissonovo rozdělení

Poissonovo rozdělení pravděpodobnosti náhodné veličiny je označováno *Po(λ)* a jedná se o diskrétní rozdělení pravděpodobnosti s parametrem λ.

Pravděpodobnostní funkce Poissonova rozdělení má tvar

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \tag{3.1}$$

pro $k = 0,1,2, \ldots$.

Tato funkce udává pravděpodobnost, že ve velkém počtu pokusů se sledovaný jev vyskytne *k* krát, jestliže pravděpodobnost jeho výskytu v jednom pokusu je velmi malá [6].

Střední hodnota a rozptyl Poissonova rozdělení mají tvar

$$E(X) = \lambda, \tag{3.2}$$

$$D(X) = \lambda. \tag{3.3}$$

3.2 P-hodnota

Definice *p*-hodnoty podle [8] je „pravděpodobnost, s jakou testovací statistika nabývá hodnot „horších“ (více svědčících proti testované hypotéze), než je pozorovaná hodnota statistiky.“

Hypotézu H_0 zamítáme na hladině významnosti α, právě když *p-hodnota* < α.

3.3 Chí-kvadrát test dobré shody

V této části vycházíme z [6] a [7].

Máme náhodný výběr rozsahu *n* z náhodné veličiny *X*. Testujeme hypotézu na hladině významnosti α, že rozdělení veličiny *X* má nějaké rozdělení, které známe až na hodnotu *m* neznámých parametrů (jestliže známe všechny parametry, pak $m = 0$).

Při testování postupujeme následovně:

Rozdělíme obor hodnot na *k* disjunktních tříd ($k \geq 2$), a zjistíme kolik hodnot náhodného výběru se nachází v jednotlivých třídách, a tyto počty označíme n_i ($i = 1, 2, \ldots, k$). Poté odhadneme neznámé parametry *m* předpokládaného modelu. Pro každou třídu spočteme očekávaný počet hodnot o_i v této třídě podle následujícího vzorce

$$o_i = n \cdot p_i, \tag{3.4}$$

pro $i = 1, 2, \ldots, k$,

kde *n* je rozsah náhodného výběru a p_i je pravděpodobnost, že veličina *X* s předpokládaným rozdělením pravděpodobnosti nabude hodnoty patřící do *i*-té třídy. Součet očekávaných hodnot o_i je roven *n*.

Pokud je některý očekávaný počet $o_i < 5$ (někdy se nedodržuje důsledně pro všechny třídy, ale musí být vždy očekávané hodnoty $o_i > 1$), pak sdružíme danou třídu s některou z vedlejších tříd. Sdružená třída má pak očekávaný počet o_i roven součtu očekávaných počtů ze tříd, jejichž sdružením vznikla. Tento postup je nutné opakovat, dokud není splněna podmínka $o_i \geq 5$ pro každou třídu. Nový počet tříd opět označíme k.

Hypotézu, že veličina se řídí předpokládaným rozdělením, zamítáme na hladině významnosti α, je-li

$$\sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} > \chi^2_{1-\alpha}(v), \tag{3.5}$$

kde $\chi^2_{1-\alpha}(v)$ je kvantil χ^2 rozdělení a v je počet stupňů volnosti $v = k - 1 - m$ $(v > 0)$.

3.4 Bonferroniho korekce

Bonferroniho korekce je jednou z nejznámnějších korekčních procedur pro násobné testování hypotéz. Při testování složených hypotéz je třeba upravit hladinu významnosti α, k čemuž se používá Bonferroniho korekce

$$\alpha^* = \frac{\alpha}{m}, \tag{3.6}$$

kde

α^* je upravená hladina významnosti,
α je původní hladina významnosti,
m je počet provedených testů.

3.5 Chí-kvadrát test v kontingenčních tabulkách

Pro testování nezávislosti používáme χ^2 test dobré shody. Pozorované četnosti zapisujeme do následující tabulky (tzv. kontingenční).

Kontingenční tabulka

	1	...	c	Σ
1	n_{11}	...	n_{1c}	$n_{1\bullet}$
2	n_{21}	...	n_{2c}	$n_{2\bullet}$
...	...	...	...	...
r	n_{r1}	...	n_{rc}	$n_{r\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet c}$	n

Obrázek 1: Ukázka kontingenční tabulky

Pro proměnné na obrázku platí

n_{ij} jsou pozorované četnosti,
n je celkový počet pozorovaných četnosti,
pro řádkové a sloupcové součty platí

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \tag{3.7}$$

kde $i = 1, \ldots, r; j = 1, \ldots, c$.

Testujeme hypotézy:

H_0: počet gólů vstřelených domácím týmem a počet gólů vstřelených hostujícím týmem jsou nezávislé náhodné veličiny

H_1: počet gólů vstřelených domácím týmem a počet gólů vstřelených hostujícím týmem nejsou nezávislé náhodné veličiny

Testovací kritérium má tvar

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - o_{ij})^2}{o_{ij}} = n \cdot \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - n. \tag{3.8}$$

Při platnosti nulové hypotézy má testové kritérium asymptoticky χ^2 rozdělení, jehož počet stupňů volnosti je roven $v = r \cdot s - (r + s - 2) = (r - 1) \cdot (s - 1)$. Pokud pro hodnotu testovacího kritéria platí, že $\chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha)$, zamítáme hypotézu o nezávislosti těchto veličin. Ke shodě s limitním rozdělením se požaduje podmínka očekávaných četností $o_{ij} > 5$. Není-li tato podmínka splněna (někdy se dodržuje striktně pro všechny třídy, ale musí být vždy alespoň 80 % očekávaných četností $o_i > 1$), pak sloučíme příslušné řádky nebo sloupce se sousedními v kontingenční tabulce, počet řádků i sloupců se však nesmí zredukovat na jeden.

4 Testování předpokladů modelů

4.1 Poissonovo rozdělení

Jak ukázal Maher ve svém článku [1], existují dobré důvody domnívat se, že počet gólů vstřelených týmem v zápase se řídí Poissonovo rozdělením: držení míče je důležitý aspekt fotbalu, a pokaždé když má tým míč, má příležitost k útoku a skórování. Pravděpodobnost *p*, že útok bude mít za následek gól je samozřejmě malá, ale kolikrát má tým míč v držení během zápasu, je velmi vysoká. Jestliže *p* je konstantní a útoky jsou nezávislé, počet gólů bude mít binomické rozdělení a za těchto okolností bude velmi dobře platit aproximace Poissonovým rozdělením, což ukázal M. J. Maher v článku [1] na výsledcích fotbalových utkání anglických lig.

V této kapitole budeme testovat pomocí χ^2 testu dobré shody (více v kapitole 3.3), zda se počty gólů vstřelených jednotlivými týmy v Extralize (resp. polské lize a NHL) řídí Poissonovo rozdělením. Dalším testovaným předpokladem bude, že počet gólů domácích a počet gólů hostů jsou dvě nezávislé náhodné veličiny, více v kapitole 4.2.

V těchto třech ligách máme k dispozici výsledky zápasů od sezóny 2009/2010 až do 2014/2015, více o datech v kapitole 2. Zkoumáme zvlášť počet gólů vstřelených doma a venku, protože týmy hrající zápas doma se zpravidla snaží více útočit a tím pádem i střílet více gólů než při hostujících zápasech.

Pomocí χ^2 testu dobré shody (kapitola 3.3) budeme testovat následující složenou nulovou hypotézu, která se skládá ze 34 jednotlivých hypotéz (pro každý tým), na hladině významnosti $\alpha = 5\ \%$.

H_0: Počet gólů vstřelených týmem v domácím/hostujícím zápase se řídí Poissonovo rozdělením.

H_1: Počet gólů vstřelených týmem v domácím/hostujícím zápase se neřídí Poissonovo rozdělením.

Testy pro jednotlivé české (polské a NHL) týmy jsou provedeny v souboru *CZE_testování předpokladů.xlsx* (*POL_testování předpokladů.xlsx* a *NHL_testování předpokladů.xlsx)*. Jako příklad nyní uvedeme jeden ze 34 testů pro českou ligu.

H_0: počet gólů vstřelených týmem Zlín v domácích zápasech se řídí Poissonovo rozdělením.

H_1: počet gólů vstřelených týmem Zlín v domácích zápasech se neřídí Poissonovo rozdělením.

Abychom zajistili, že chyba 1. druhu (α) bude korektní, je třeba použít jednu z metod pro multinomické testování, v našem případě Bonferroniho korekci (více v kapitole 3.4). Složená hypotéza se skládá ze 34 jednotlivých hypotéz, proto upravená hladina významnosti $\alpha^* = \frac{0{,}05}{34} = 0{,}147\ \%$.

V následujících tabulkách je uveden pozorovaný a očekávaný počet gólů týmu Zlín v domácích zápasech, kritická hodnota a *p-hodnota* pro tento test.

Počet gólů [x]	Pozorované četnosti [n_i]	Pravděpodobnosti [p_i]	Očekávané četnosti [o_i]	Testová statistika
0	7	0,05	8	0,21
1	21	0,16	24	0,48
2	41	0,23	36	0,77
3	37	0,22	35	0,12
4	26	0,16	26	0,01
5	14	0,10	15	0,06
6 a více	10	0,08	12	0,32
Celkem	**156**	**1,00**	**156**	**1,98**

Tabulka 1: Pozorovaný a očekávaný počet gólů pro domácí tým Zlín

testová statistika	1,98
stupeň volnosti	5
hladina významnosti [%]	0,15
kritická hodnota	8,17
p-hodnota	0,85

Tabulka 2: Výsledky chí-kvadrát testu

Přehled získaných výsledků uvedených testů pro všechny týmy z české Extraligy včetně *p*-hodnot je v tabulce 3, více lze nalézt v souboru *CZE_testování předpokladů.xlsx* na listu *Poisson_chí-kvadrát test*.

Tým	Test doma	p-hodnota doma	Test venku	p-hodnota venku
České Budějovice	nezamítáme H0	0,963	nezamítáme H0	0,860
Karlovy Vary	nezamítáme H0	0,040	nezamítáme H0	0,874
Kladno	nezamítáme H0	0,236	nezamítáme H0	0,314
Kometa Brno	nezamítáme H0	0,124	nezamítáme H0	0,132
Liberec	nezamítáme H0	0,131	nezamítáme H0	0,204
Litvínov	nezamítáme H0	0,231	nezamítáme H0	0,019
Mladá Boleslav	nezamítáme H0	0,998	nezamítáme H0	0,163
Pardubice	nezamítáme H0	0,941	nezamítáme H0	0,225
Plzeň	nezamítáme H0	0,127	nezamítáme H0	0,552
Slavia Praha	nezamítáme H0	0,132	nezamítáme H0	0,724
Sparta Praha	nezamítáme H0	0,131	nezamítáme H0	0,002
Třinec	nezamítáme H0	0,087	nezamítáme H0	0,992
Vítkovice	nezamítáme H0	0,968	nezamítáme H0	0,121
Zlín	nezamítáme H0	0,852	nezamítáme H0	0,497
Hradec Králové	nezamítáme H0	0,904	nezamítáme H0	0,138
Chomutov	nezamítáme H0	0,885	nezamítáme H0	0,327
Olomouc	nezamítáme H0	0,141	nezamítáme H0	0,040

Tabulka 3: Přehled výsledků hypotéz a jejich *p-hodnot* – Extraliga (CZE)

V dalších tabulkách jsou výsledky těchto testů pro týmy z NHL ligy a polské ligy včetně *p* - hodnot, více lze nalézt v souboru *NHL_testování předpokladů.xlsx* na listu *Poisson_chí-kvadrát test*.

Tým	Test doma	p-hodnota doma	Test venku	p-hodnota venku
Anaheim	nezamítáme H0	0,863	nezamítáme H0	0,706
Atlanta	nezamítáme H0	0,751	nezamítáme H0	0,192
Boston	nezamítáme H0	0,232	nezamítáme H0	0,386
Buffalo	nezamítáme H0	0,195	nezamítáme H0	0,004
Calgary	nezamítáme H0	0,266	nezamítáme H0	0,050
Carolina	nezamítáme H0	0,794	nezamítáme H0	0,777
Colorado	nezamítáme H0	0,540	nezamítáme H0	0,075
Columbus	nezamítáme H0	0,582	nezamítáme H0	0,299
Dallas	nezamítáme H0	0,861	nezamítáme H0	0,125
Detroit	nezamítáme H0	0,636	nezamítáme H0	0,971
Edmonton	nezamítáme H0	0,472	nezamítáme H0	0,250
Florida	nezamítáme H0	0,785	nezamítáme H0	0,721
Chicago	nezamítáme H0	0,230	nezamítáme H0	0,125
Los Angeles	nezamítáme H0	0,491	nezamítáme H0	0,370
Minnesota	nezamítáme H0	0,263	nezamítáme H0	0,009
Montreal	nezamítáme H0	0,085	nezamítáme H0	0,746
Nashville	nezamítáme H0	0,051	nezamítáme H0	0,793
New Jersey	nezamítáme H0	0,732	nezamítáme H0	0,622
NY Islanders	nezamítáme H0	0,682	nezamítáme H0	0,305
NY Rangers	nezamítáme H0	0,473	nezamítáme H0	0,937
Ottawa	nezamítáme H0	0,234	nezamítáme H0	0,051
Philadelphia	nezamítáme H0	0,001	nezamítáme H0	0,487
Phoenix	nezamítáme H0	0,963	nezamítáme H0	0,398
Pittsburgh	nezamítáme H0	0,407	nezamítáme H0	0,078
San Jose	nezamítáme H0	0,370	nezamítáme H0	0,523
St. Louis	nezamítáme H0	0,841	nezamítáme H0	0,659
Tampa Bay	nezamítáme H0	0,001	nezamítáme H0	0,495
Toronto	nezamítáme H0	0,252	nezamítáme H0	0,131
Vancouver	nezamítáme H0	0,040	nezamítáme H0	0,246
Washington	nezamítáme H0	0,007	nezamítáme H0	0,149

Tabulka 4: Přehled výsledků hypotéz a jejich *p-hodnot* – NHL

Tým	Test doma	p-hodnota doma	Test venku	p-hodnota venku
Bytom	nezamítáme H0	0,0497	nezamítáme H0	0,1306
Janów	nezamítáme H0	0,6457	nezamítáme H0	0,5272
Jastrzebie JKH GKS	nezamítáme H0	0,0001	nezamítáme H0	0,0001
Katowice	nezamítáme H0	0,2090	nezamítáme H0	0,0027
Kraków	nezamítáme H0	0,0000	nezamítáme H0	0,4973
Krynica KTH	nezamítáme H0	0,0004	nezamítáme H0	0,1877
Orlik Opole	nezamítáme H0	0,0073	nezamítáme H0	0,0327
Podhale Nowy Targ	nezamítáme H0	0,0149	nezamítáme H0	0,1074
Sanok	nezamítáme H0	0,0152	nezamítáme H0	0,0010
Stoczniowiec Gdańsk	nezamítáme H0	0,3282	nezamítáme H0	0,8745
Toruň	nezamítáme H0	0,6369	nezamítáme H0	0,6093
Tychy	nezamítáme H0	0,2704	nezamítáme H0	0,0136
Unia Oświęcim	nezamítáme H0	0,0030	nezamítáme H0	0,0009
Zagłębie Sosnowiec	nezamítáme H0	0,0101	nezamítáme H0	0,3696

Tabulka 5: Přehled výsledků hypotéz a jejich *p-hodnot* –Ekstraliga (POL)

4.2 Test nezávislosti

Testujeme hypotézy (viz kapitola 3.5):

H_0: počet gólů vstřelených domácím týmem a počet gólů vstřelených hostujícím týmem jsou nezávislé náhodné veličiny,

H_1: počet gólů vstřelených domácím týmem a počet gólů vstřelených hostujícím týmem nejsou nezávislé náhodné veličiny.

Uvedené hypotézy testujeme pro každou sezónu zvlášť pomocí χ^2 testu v kontingenčních tabulkách (kapitola 3.5).

Při testování nezávislosti jsme u české ligy došli k závěrům, že nezamítáme hypotézu H_0 v sezónách 2009/2010, 2012/2013 a 2014/2015, tj. počet gólů vstřelených domácím týmem (X_{ij}) a počet gólů vstřelených hostujícím týmem (Y_{ij}) jsou v těchto sezónách považovány za nezávislé náhodné veličiny. V ostatních sezónách 2010/2011, 2011/2012 a 2013/2014 naopak zamítáme hypotézu H_0, tj. přijímáme hypotézu H_1., že se jedná o závislé náhodné veličiny. Detaily těchto testů jsou zpracovány v souboru *CZE_testování předpokladu.xlsx* na listu *test nezávislosti*.

V NHL lize v sezóně 2009/2010, 2012/2013 a 2014/2015 jsme nezamítali hypotézu H_0 a naopak v ostatních zkoumaných sezónách H_0 zamítáme. Detail těchto testů lze nalézt v *NHL_testování předpokladu.xlsx* na listu *test nezávislosti*.

Pro poslední třetí ligu jsme nezamítali hypotézu H_0 pro sezóny 2011/2012 a 2012/2013, a ve zbývajících zkoumaných sezónách tomu bylo naopak. Výsledky všech testů jsou obsaženy v souboru *POL_testování předpokladu.xlsx* na listu *test nezávislosti*.

5 Maherovy modely

V této kapitole popíšeme, jaké modely používá M. J. Maher pro odhadování fotbalových zápasů a vybraný model použijeme pro hokejová data. V některých částech této kapitoly lze nalézt inspirace z [12].

5.1 Druhy modelů

Pravděpodobnost p, že útok bude mít za následek gól, je samozřejmě malá, ale kolikrát má tým kotouč v držení během zápasu je velmi vysoká. Jestliže p je konstantní a útoky jsou nezávislé, počet gólů bude mít binomické rozdělení a za těchto okolností bude platit Poissonova aproximace velmi dobře. Střední hodnota Poissonova rozdělení se bude lišit v závislosti na kvalitě týmu.

Na základě těchto poznatků, které byly také uvedeny ve článku [1], M. J. Maher popsal několik modelů pro odhad výsledků fotbalových zápasů za použití Poissonova rozdělení. Jestliže tým i hraje doma proti týmu j a pozorované skóre je (x_{ij},y_{ij}), Maher předpokládá, že počet gólů X_{ij} vstřelených domácím týmem se řídí Poissonovo rozdělením s parametrem λ_{ij} a počet gólů Y_{ij} vstřelených hostujícím týmem se řídí Poissonovo rozdělením s parametrem μ_{ij} (ověřeno v předchozí kapitole), a že X_{ij} a Y_{ij} jsou nezávislé.

Parametr λ je vyjádřen následujícím vzorcem

$$\lambda_{ij} = \alpha_i \cdot \beta_j, \tag{5.1}$$

kde α_i představuje sílu útoku domácího týmu i a β_j udává slabost obrany hostujícího týmu j.

Parametr μ je dán následujícím vzorcem

$$\mu_{ij} = \gamma_i \cdot \delta_j, \tag{5.2}$$

kde γ_i představuje slabost obrany domácího týmu i a δ_j sílu útoku hostujícího týmu j.

Je otázka, zda jsou všechny tyto parametry nezbytné pro zajištění odpovídajícího popisu skóre. Maher se v článku zamýšlí, jestli jsou patrnější rozdíly v útocích nebo obranách, pokud existují rozdíly mezi týmy. A zda je opravdu nutné mít odlišné parametry pro kvalitu útoku týmu doma a venku. Zvážení těchto otázek vede k možné hierarchii modelů, které se liší výpočtem jednotlivých parametrů a jsou popsány v další části podle Mahera z článku [1].

5.1.1 Model 0

Základní model je postaven na úvaze, že všechny týmy jsou stejně silné v útoku i v obraně. V takovém případě platí $\alpha_i = \alpha$, $\beta_i = \beta$, $\gamma_i = \gamma$ a $\delta_i = \delta$ pro všechna i. Aby množina odhadovaných parametrů byla jednoznačná, tak platí omezení $\alpha = \beta$ a $\gamma = \delta$, z čehož vyplývá, že pro tento model stačí odhadnout pouze dva nezávislé parametry, a to je výhodou tohoto modelu. Naopak nevýhodou je, že model bere všechny týmy za stejně silné, což pravděpodobně není pravda.

5.1.2 Model 1A, 1B

Další variantou může být úvaha, že obrana všech týmů je stejně silná, ale v útoku je síla týmů odlišná. Takovou variantu nazývá Maher jako model 1A a platí $\alpha_i = \delta_i$, $\beta_i = \beta$, $\gamma_i = \gamma$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Zde je potřeba odhadnout $n + 1$ nezávislých parametrů, kde n je počet týmů v lize.

Model 1B je opačnou variantou modelu 1A, kde útok všech týmů je stejně silný a obrana každého týmu je odlišná, tj. $\alpha_i = \alpha$, $\beta_i = \gamma_i$ a $\delta_i = \delta$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Stejně jako u modelu 1A je potřeba odhadnout $n + 1$ nezávislých parametrů.

5.1.3 Model 2

Model 2 je již variantou uvažující rozdílnou sílu týmu v útoku i obraně a navíc je zde zaveden parametr k, který vyjadřuje poměr síly týmů venku a doma, tedy platí $\delta_i = k \cdot \alpha_i$, $\gamma_i = k \cdot \beta_i$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Je potřeba odhadnout $2n$ nezávislých parametrů.

5.1.4 Model 3C, 3D

V modelu 3C se počítá síla týmu v útoku a v obraně pro každý tým zvlášť. Síla týmu v útoku je brána za stejnou doma i venku, a platí $\alpha_i = \delta_i$ pro všechna i a $\sum_i \alpha_i = \sum_i \beta_i$. Naopak síla týmu v obraně je doma a venku různá. Zde je potřeba odhadnout $3n - 1$ nezávislých parametrů.

Model 3D je opačnou variantou modelu 3C, kde platí, že síla týmu v obraně je doma a venku stejná a v útoku se počítá zvlášť doma a venku.

5.1.5 Model 4

Nejsložitější model dle Mahera uvažuje zvlášť sílu týmu doma a venku, v útoku i obraně. Platí zde $\sum_i \alpha_i = \sum_i \beta_i$ a $\sum_i \gamma_i = \sum_i \delta_i$. V tomto případě je třeba odhadnout $4n - 2$ nezávislých parametrů.

5.2 Zvolený model

Pro naše vybraná data hokejových zápasů zvolíme model 2 (kapitola 5.1.3), který Maher také použil ve svém článku [1] jako nejvhodnější pro fotbalová data. Výhodou tohoto modelu je, že se již počítá s rozdílnou silou jednotlivých týmů v útoku i v obraně, a je třeba odhadnout méně parametrů oproti složitějším modelům. V tomto modelu se odhaduje parametr síly útoku α_i pro všechna i, parametr síly obrany β_j pro všechna j a parametr k (resp. k^2), který vyjadřuje sílu při venkovních zápasech oproti síle při domácích zápasech.

Náhodná veličina X_{ij} označuje počet gólů, které vstřelí domácí tým i a náhodná veličina Y_{ij} značí počet gólů vstřelených hostujícím týmem j v zápase. Předpokládáme, že X_{ij} a Y_{ij} jsou nezávislé a řídí se Poissonovo rozdělením

$$X_{ij} \sim Po(\alpha_i \cdot \beta_j) \tag{5.3}$$

$$Y_{ij} \sim Po(k^2 \cdot \alpha_j \cdot \beta_i) \tag{5.4}$$

Pro odhad parametrů ve zvoleném modelu 2 jsou v článku [1] na str. 114 odvozené metodou maximální věrohodnosti následující vzorce

$$\hat{\alpha}_i = \frac{\sum_{j \neq i}(x_{ij} + y_{ji})}{(1 + \hat{k}^2)\sum_{j \neq i}\hat{\beta}_j} \tag{5.5}$$

$$\hat{\beta}_j = \frac{\sum_{i \neq j}(x_{ij} + y_{ji})}{\left(1 + \hat{k}^2\right)\sum_{i \neq j}\hat{\alpha}_i} \tag{5.6}$$

$$\hat{k}^2 = \frac{\sum_i \sum_{j \neq i} y_{ij}}{\sum_i \sum_{j \neq i} x_{ij}}, \tag{5.7}$$

kde x_{ij} je počet gólů vstřelených týmem *i* týmu *j* v domácím zápase, y_{ij} je počet gólů vstřelených týmem *i* týmu *j* ve venkovním zápase.

Z předchozího vyplývají následující podmínky

$$\sum_i \sum_{j \neq i} \hat{\alpha}_i \cdot \hat{\beta}_j = \sum_i \sum_{j \neq i} x_{ij}, \tag{5.8}$$

$$\sum_i \sum_{j \neq i} \hat{k}^2 \cdot \hat{\alpha}_j \cdot \hat{\beta}_i = \sum_i \sum_{j \neq i} y_{ij,} \tag{5.9}$$

což znamená, že součet středních hodnot vhodných Poissonových rozdělení se rovná pozorovanému počtu vstřelených (resp. obdržených) gólů.

Maherovy modely jsou zaměřené na data z fotbalových utkání, kde hraje každý tým proti každému pouze jednou doma a jednou venku. V Extralize hraje každý tým proti každému dvakrát doma a dvakrát venku, proto bylo potřeba upravit rovnici (5.8) a (5.9) na tvar

$$\sum_i \left(\sum_{j \neq i} \hat{\alpha}_i \cdot \hat{\beta}_j \right) \cdot 2 = \sum_i \sum_{j \neq i} x_{ij} \tag{5.10}$$

$$\sum_i \left(\sum_{j \neq i} \hat{k}^2 \cdot \hat{\alpha}_j \cdot \hat{\beta}_i \right) \cdot 2 = \sum_i \sum_{j \neq i} y_{ij}. \tag{5.11}$$

5.2.1 Zhodnocení

Na základě výsledků χ^2 testu lze říci, že pouze v některých sezónách se počet gólů řídí Poissonovo rozdělením s parametry dle zkoumaného modelu. χ^2 test byl proveden pro celé sezóny, proto jednotlivé zápasy mohou mít jiné rozdělení. Nevýhodou tohoto modelu je, že se dá modelovat až po sezóně, protože počet zápasů každého týmu musí být stejný. Další nevýhodou je, že všechny zápasy v sezóně mají stejnou váhu, a tím se neprojevuje v modelu aktuální forma z posledních zápasů.

Na závěr lze říci, že Maherův model není příliš vhodný pro hokejová data, což bylo ověřeno na české lize, proto jsme se rozhodli nezpracovávat tento model pro ostatní ligy. Nedostatky tohoto modelu budou odstraněny v dalších modelech popsaných v kapitole 6.

5.3 Model pro Extraligu

Model 2 podle Mahera popsaný v předchozí podkapitole je použit pro data české ligy od sezóny 2009/2010 do 2014/2015 v sešitu *CZE_model Maher.xlsm*. Na jednotlivých listech pojmenovaných podle sezón vycházíme z dat, která jsou uvedena na listu *data_uspořádaná*.

V následující podkapitole jsou uvedené odhady parametrů pouze pro sezónu 2014/2015. Pro předchozí sezóny je vše v sešitu *CZE_model Maher.xlsm*. Odhad sezóny 2015/2016 by byl proveden na základě parametrů odhadnutých v ukončené sezóně 2014/2015.

5.3.1 Parametry

- Nejprve potřebujeme vyjádřit poměr síly týmu venku k síle v domácím prostředí. Tento vztah vyjadřuje parametr k^2 dle rovnice (5.7). Pro sezónu 2014/2015 je $k^2 = 0{,}83$ (tj. 0,83 krát méně gólů venku oproti počtu gólů v domácím zápase).
- Poté je nutné vyjádřit pro každý tým sílu v útoku a obraně. Odhadneme tedy parametry α a β uvedené v rovnicích (5.5) a (5.6).

Postup odhadu parametrů $\boldsymbol{\alpha}$ *a* $\boldsymbol{\beta}$ *v Microsoft Excel 2013:*

1. Zvolíme výchozí hodnoty parametrů α a β pro všechny týmy viz Obrázek 2, sloupec *L* a *M*.
2. Spočteme parametry ve sloupcích *N* a *O* podle rovnice (5.5) a (5.6).
3. Spustíme řešitele s následujícím nastavením
    a. součet hodnot parametrů α_i pro všechny týmy se musí rovnat součtu hodnot parametru β_i dle podmínky definované v kapitole 5.1.3. (*N18 = O18*),
    b. nastavíme podmínky ve sloupci *P* dle rovnice (5.10).,
    c. celková hodnota v buňce *P18* musí být rovna počtu gólů domácích v celé sezóně (1077 gólů pro sezónu 2014/2015),
    d. měněné buňky jsou výchozí hodnoty parametrů α a β.
4. Na základě nových výchozích hodnot zjištěných řešitelem se dopočtou odhady parametrů α a β, které se v další iteraci použijí jako startovací hodnoty.
5. Tento postup se opakuje 5 krát, a po 5. cyklu se hodnoty ze sloupců *N* a *O* zkopírují do sloupců *Q* a *R*, čímž odhad skončí a za optimální odhad považujeme hodnoty parametrů uložené ve sloupcích pojmenovaných „alfa_řešitel“ a „beta_řešitel“.

Celý postup je proveden vytvořeným makrem ve Visual Basic viz Příloha 1, které se spouští tlačítkem „Odhad – řešitel“.

Na Obrázek 2 je ukázka počátečního nastavení před první iterací.

	K	L	M	N	O	P	Q	R
3	Tým	alfa_výchozí	beta_výchozí	alfa_odhad	beta_odhad	Nastavení podmínky	alfa_řešitel	beta_řešitel
4	Hradec Králové	1,00	1,00	5,60	5,47	866,41	1,52	1,57
5	Karlovy Vary	1,00	1,00	4,55	5,73	701,25	2,25	1,83
6	Kometa Brno	1,00	1,00	6,31	5,47	977,15	1,74	1,58
7	Liberec	1,00	1,00	5,01	5,39	776,05	1,38	1,54
8	Litvínov	1,00	1,00	6,78	4,93	1 056,23	1,87	1,43
9	Mladá Boleslav	1,00	1,00	6,57	6,52	1 002,41	1,83	1,89
10	Olomouc	1,00	1,00	5,09	6,82	774,51	1,42	1,95
11	Pardubice	1,00	1,00	5,98	6,31	914,97	1,67	1,82
12	Plzeň	1,00	1,00	5,73	5,51	885,47	1,60	1,59
13	Slavia Praha	1,00	1,00	4,34	7,54	653,09	1,23	2,14
14	Sparta Praha	1,00	1,00	7,91	5,85	1 218,70	2,22	1,72
15	Třinec	1,00	1,00	7,54	4,80	1 176,22	2,11	1,41
16	Vítkovice	1,00	1,00	5,56	6,27	851,00	1,57	1,80
17	Zlín	1,00	1,00	5,89	6,23	903,08	1,67	1,80
18	**Počet gólů celkem**	**14,00**	**14,00**	**82,85**	**82,85**	**12 756,53**	**24,08**	**24,08**
19								
20				Musí být splněna podmínka, že součet alpha = součet beta.		Tato hodnota se musí rovnat počtu gólů domácích (tj. 1 077 - buňka H20).		
21								
22								

Obrázek 2: Ukázka nastavení výchozích hodnot pro odhad parametrů α a β

V následující tabulce jsou zobrazené výsledné hodnoty parametrů α_i a β_i po 5. iteraci, když výchozí hodnoty všech parametrů byly nastavené na hodnotu 1.

Tým	alfa_řešitel	beta_řešitel
Hradec Králové	1,52	1,57
Karlovy Vary	2,25	1,83
Kometa Brno	1,74	1,58
Liberec	1,38	1,54
Litvínov	1,87	1,43
Mladá Boleslav	1,83	1,89
Olomouc	1,42	1,95
Pardubice	1,67	1,82
Plzeň	1,60	1,59
Slavia Praha	1,23	2,14
Sparta Praha	2,22	1,72
Třinec	2,11	1,41
Vítkovice	1,57	1,80
Zlín	1,67	1,80
Suma	**24,08**	**24,08**

Tabulka 6: Výsledné parametry α a β pro sezónu 2014/2015

5.3.2 Výsledky

V této podkapitole budou ukázány výsledky pouze pro dva vybrané týmy, výsledky ostatních týmů jsou v sešitu *CZE_model Maher.xlsm* na listech jednotlivých sezón.

Z odhadnutých parametrů z minulé části lze vypočítat odhady parametrů λ_{ij} dle rovnice (5.1) a μ_{ij} dle rovnice (5.2). Pokud bychom uvažovali zápas mezi Plzní a Spartou Praha, potom odhad střední hodnoty počtu gólů vstřelených byl následující:

Pro Plzeň $$\hat{\lambda} = \hat{\alpha}_{PLZ} \cdot \hat{\beta}_{SP} = 1{,}60 \cdot 1{,}72 = 2{,}75 \tag{5.12}$$

Pro Sparta Praha $$\hat{\mu} = \hat{k}^2 \cdot \hat{\alpha}_{SP} \cdot \hat{\beta}_{PLZ} = 0{,}83 \cdot 2{,}22 \cdot 1{,}59 = 2{,}93 \tag{5.13}$$

Nyní můžeme určit pravděpodobnosti kolik dá tým gólů v zápase.

Pro Plzeň, kde $\hat{\lambda}$ = 2,75 lze podle (3.1) dopočítat pravděpodobnost, že Plzeň dá Spartě 1 gól je:

$$P(X = 1) = e^{-2{,}75} \cdot \frac{2{,}75^1}{1!} = 0{,}18. \tag{5.14}$$

Nebo pravděpodobnost, že Plzeň dá Spartě více než 6 gólů je

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - F(5) = 0{,}06. \tag{5.15}$$

Pro Sparta Praha, kde $\hat{\mu}$ = 2,93 lze podle (3.1) dopočítat pravděpodobnost, že Sparta dá Plzni 1 gól je:

$$P(Y = 1) = e^{-2{,}93} \cdot \frac{2{,}93^1}{1!} = 0{,}16. \tag{5.16}$$

Můžeme také dopočítat pravděpodobnost, že zápas skončí například za stavu 1:1

$$P(X = 1, Y = 1) = P(X = 1) \cdot P(Y = 1) = 0{,}18 \cdot 0{,}16 = 0{,}03. \tag{5.17}$$

V tabulce 7 jsou uvedené pravděpodobnosti, že domácí tým dá 1 gól hostujícímu týmu v zápase mezi jakýmikoliv týmy.

P(X = 1)		Hosté													
		HK	KV	Brno	Liberec	Litvínov	MB	Olomouc	Pardubice	Plzeň	Slavia Praha	Sparta Praha	Třinec	Vítkovice	Zlín
Domácí	Hradec Králové	x	0,171	0,217	0,225	0,247	0,161	0,152	0,173	0,215	0,125	0,190	0,250	0,176	0,176
	Karlovy Vary	0,104	x	0,102	0,109	0,129	0,060	0,055	0,068	0,101	0,039	0,080	0,133	0,070	0,070
	Kometa Brno	0,179	0,132	x	0,185	0,207	0,123	0,115	0,134	0,175	0,090	0,150	0,211	0,136	0,137
	Liberec	0,249	0,201	0,247	x	0,274	0,192	0,183	0,204	0,245	0,154	0,220	0,278	0,206	0,207
	Litvínov	0,157	0,111	0,154	0,163	x	0,103	0,096	0,113	0,153	0,073	0,128	0,189	0,116	0,116
	Mladá Boleslav	0,163	0,117	0,161	0,170	0,192	x	0,101	0,119	0,160	0,078	0,135	0,196	0,122	0,122
	Olomouc	0,240	0,191	0,237	0,245	0,266	0,182	x	0,194	0,236	0,144	0,211	0,269	0,197	0,197
	Pardubice	0,191	0,143	0,189	0,197	0,219	0,134	0,125	x	0,187	0,100	0,161	0,223	0,148	0,148
	Plzeň	0,205	0,156	0,203	0,211	0,233	0,147	0,138	0,159	x	0,112	0,175	0,237	0,161	0,162
	Slavia Praha	0,281	0,237	0,279	0,286	0,303	0,227	0,218	0,239	0,278	x	0,254	0,306	0,241	0,242
	Sparta Praha	0,107	0,069	0,105	0,113	0,133	0,063	0,057	0,071	0,104	0,041	x	0,137	0,073	0,073
	Třinec	0,122	0,081	0,120	0,127	0,148	0,074	0,068	0,083	0,118	0,050	0,096	x	0,085	0,085
	Vítkovice	0,210	0,162	0,208	0,216	0,238	0,152	0,144	0,164	0,207	0,117	0,181	0,242	x	0,167
	Zlín	0,191	0,143	0,189	0,197	0,219	0,134	0,126	0,145	0,187	0,100	0,162	0,223	0,148	x

Tabulka 7: Pravděpodobnosti, že domácí tým dá 1 gól hostujícímu týmu

Obdobné tabulky s pravděpodobnostmi, že hostující tým dá 0, 1, … gólů domácímu týmu v zápase jsou uvedené v sešitu *CZE_model Maher.xlsm* na listu konkrétní sezóny (sloupce *AL* až *BA*).

5.3.3 χ^2 test

Lze otestovat zda vypočtené pravděpodobnosti odpovídají skutečným výsledkům. V Maherově článku [1] je použit k tomuto testování χ^2 test dobré shody viz také kapitola 3.3. Testujeme zvlášť góly doma a venku v sešitu *CZE_model Maher.xlsm* na listu *chí-kvadrát test*. Následující hypotézy testujeme na hladině významnosti $\alpha = 5\ \%$.

H_0: Počet gólů vstřelených týmy doma (resp. venku) v sezóně 2009/2010 (a také pro další sezóny) se neliší od počtu gólů určených ve zvoleném modelu.

H_1: Počet gólů vstřelených týmy doma (resp. venku) v sezóně 2009/2010 (a také pro další sezóny) se liší od počtu gólů určených ve zvoleném modelu.

Nejprve se vyjádří pozorované četnosti gólů n_i podle výsledků zápasů v každé sezóně. Dále se určí očekávané četnosti jednotlivých gólů, které se získají například jako součet tabulky $P(X = 0)$ vynásobené dvěma (kvůli dvěma zápasům hraným doma a dvěma venku, jak již bylo uvedeno výše) pro počet zápasů kdy domácí tým nedal gól.

DOMA	Sezóna 2014/2015	
Počet gólů	Pozorované četnosti	Očekávané četnosti
0	24	22,52
1	52	59,16
2	87	80,79
3	78	76,52
4	52	56,59
5	36	34,84
6 a více	35	33,57
Počet gólů celkem	**364**	**364,00**

Tabulka 8: Pozorované a očekávané četnosti pro domácí zápasy

VENKU	Sezóna 2014/2015	
Počet gólů	Pozorované četnosti	Očekávané četnosti
0	21	35,60
1	96	78,81
2	107	90,78
3	57	72,61
4	36	45,37
5	26	23,61
6 a více	21	17,22
Počet gólů celkem	**364**	**364,00**

Tabulka 9: Pozorované a očekávané četnosti pro venkovní zápasy

Ostatní sezóny jsou zpracované na listu *chí-kvadrát test*. *P-hodnota* u testu domácích zápasů pro sezónu 2014/2015 vychází 0,86, což značí, že bychom hypotézu H_0 neměli zamítat. Naopak u testu venkovních zápasů vychází *p-hodnota* 0,002, proto bychom měli hypotézu H_0 zamítat, tj. přijímat hypotézu H_1. V ostatních sezónách bychom některé hypotézy také zamítali, a to je dáno velkým rozdílem mezi pozorovanými a očekávanými četnostmi gólů.

6 Dixon-Colesovy modely

Mark J. Dixon a Stuart G. Coles vylepšili Maherův model, a tento model popisují ve svém článku [2]. Jejich model byl vytvořený pouze pro fotbalová data, proto používáme upravené Dixon-Colesovy modely z článku [3], které se řadí do skupiny dynamických modelů a jsou určené pro hokejová data.

U statistického modelu jsou požadovány různé vlastnosti:

- měl by brát v úvahu rozdílné schopnosti obou týmů v utkání,
- měl by uvažovat skutečnost, že týmy hrající doma mají všeobecně nějakou výhodu tzv. „domácí výhodu“,
- nejrozumnější měřítko schopnosti týmu by mělo být založeno na tom, jakou výkonnost měl tým v minulosti, kdy největší váhu by měly mít nejnovější výsledky,
- měl by brát zvlášť schopnost týmu útočit (dát gól) i jeho schopnost bránit se (nedostat gól).

Všechny tyto předpoklady splňují upravené Dixon-Colesovy modely.

6.1 Pravděpodobnost výhry, remízy a prohry

Chceme určit, s jakou pravděpodobností dají týmy určitý počet gólů v utkání, a tím odhadnout zda tým vyhraje, remízuje či prohraje. Náhodné proměnné X_{ij} a Y_{ij} označují výsledky zápasu mezi týmy i *a* j, kde náhodná proměnná X_{ij} představuje počet gólů vstřelených domácím mužstvem a náhodná proměnná Y_{ij} je počet gólů od hostujícího týmu, jak již bylo uvedeno v Maherově modelu (kapitola 5.1). Za předpokladu, že je známá sdružená pravděpodobnostní funkce (X_{ij}, Y_{ij}), můžeme spočítat pravděpodobnosti výhry domácího týmu $\left(p_{ij}^{H}\right)$[1], remízy $\left(p_{ij}^{D}\right)$[2] a prohry, čili vítězství hostujícího týmu $\left(p_{ij}^{A}\right)$[3], dle následujících vzorců

$$p_{ij}^{H} = \sum_{x>y} P\left(X_{ij} = x, Y_{ij} = y\right), \tag{6.1}$$

$$p_{ij}^{D} = \sum_{x=y} P\left(X_{ij} = x, Y_{ij} = y\right), \tag{6.2}$$

$$p_{ij}^{A} = \sum_{x<y} P\left(X_{ij} = x, Y_{ij} = y\right). \tag{6.3}$$

V následujících podkapitolách jsou popsány čtyři modely, které mohou být použity pro odhad spojité pravděpodobnostní funkce (X_{ij}, Y_{ij}). Inspirace pro modely se bere z fotbalu, kde můžeme najít některé modely, které uvažují počet gólů domácího a hostujícího týmu jako nezávislé náhodné proměnné. Některé modely však považují skóre domácího a hostujícího týmu jako závislé náhodné proměnné. Nelze přesně říci, která volba je celkově vhodnější pro lední hokej, a proto se zkoumají oba modely.

[1] H = home, používá se u parametrů týkajících se domácího týmu,
[2] D = draw, používá se u parametrů týkajících se remízy,
[3] A = away, používá se u parametrů týkajících se hostujícího týmu.

Dále nás přesvědčují výsledky od Karlise a Ntzoufrase [4], kde se podhodnocují pravděpodobnosti remíz, proto je navržena modifikace obou modelů, a je zde použita.

6.2 Dvojnásobný Poissonovo (DP) model

V tomto modelu, původně užitém Maherem [1], jsou obě náhodné proměnné X_{ij} a Y_{ij} nezávislé, pak platí[4]

$$X_{ij} \sim Po\left(\lambda_H = \mu \cdot \alpha_i \cdot \beta_j \cdot \gamma\right), \tag{6.4}$$

$$Y_{ij} \sim Po\left(\lambda_A = \mu \cdot \alpha_j \cdot \beta_i\right), \tag{6.5}$$

kde

α_i je síla útoku týmu (čím vyšší, tím lepší), $\alpha_i > 0$ pro všechna *i*,

β_i je naopak síla obrany týmu (čím menší, tím lepší), $\beta_i > 0$ pro všechna *i*,

μ je parametr, který vyjadřuje průměrný počet gólů hostujícího týmu, $\mu > 0$

γ je míra výhody domácího utkání (více v článku [1] nebo [2]), $\gamma > 0$

6.2.1 Sdružená pravděpodobnostní funkce

Sdružená pravděpodobnostní funkce pro výsledek zápasu mezi domácím týmem *i* a hostujícím týmem *j* má tvar

$$P(x, y) = \frac{\lambda_H^x \cdot e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y \cdot e^{-\lambda_A}}{y!} \tag{6.6}$$

pro x, y = 0, 1, 2, …

Abychom dosáhli identifikovatelnosti modelu, je nezbytné použít omezující podmínky $\sum_i \alpha_i = N$ a $\sum_i \beta_i = N$, kde *N* je počet týmů, které vstupují do modelu. Tyto podmínky se týkají nejen tohoto modelu, ale i těch následujících, a zajišťují, že průměrná hodnota všech útočných i obranných parametrů je 1.

6.3 Dvourozměrný Poissonovo (BP) model

V tomto modelu z článku [3] autoři navrhují užít dvourozměrné Poissonovo rozdělení. Toto rozdělení však nelze použít vzhledem k záporné korelaci mezi počtem gólů domácích a hostů. Proto použijeme dvourozměrné Poissonovo rozdělení, které je definované takto[5] (detail tohoto rozdělení viz [5])

$$P(x, y) = \lambda_H^x \cdot \lambda_A^y \cdot e^{-\lambda_H - \lambda_A} \cdot \frac{1 + \lambda \cdot \left(e^{-x} - e^{-d \cdot \lambda_H}\right) \cdot \left(e^{-y} - e^{-d \cdot \lambda_A}\right)}{x! \cdot y!} \tag{6.7}$$

pro x, y = 0, 1, 2, …,

[4] Model použitý Maherem, ale upraveným pro hokejová data, kde provádíme také reparametrizaci parametru μ.

[5] Tímto postupem nahrazujeme složitější a problematické řešení, které využívá propojení dvou Poissonovo rozdělení pomocí kopule [3].

kde $d = 1 - e^{-1}$, λ_H a λ_A jsou získány stejným způsobem jako v rovnicích (6.4) a (6.5). Korelační koeficient (záporný, kladný nebo nulový) mezi X_{ij} a Y_{ij} se může spočítat užitím parametru λ takto

$$\rho = \lambda \cdot \sqrt{\lambda_H \cdot \lambda_A} \cdot d^2 \cdot e^{-d\cdot(\lambda_H+\lambda_A)}. \tag{6.8}$$

Z toho plyne, že musí být odhadnutý parametr λ. Stejně jako parametry γ a μ, i parametr λ je globální pro všechny týmy a určuje závislost mezi góly vstřelenými domácími a hostujícími týmy, $\lambda \in R$.

6.4 Diagonálně rozšířený model

Tento model je rozšířením předchozích dvou modelů. Výsledky v článku [2] a [4] nás přesvědčují, že modely užité pro fotbalová data podhodnocují pravděpodobnost výsledku remízy. Karlis a Ntzoufras (2003) navrhli diagonálně rozšířený model v této formě

$$P_{DI}(x,y) = \begin{cases} (1-p)P(x,y), & x \neq y \\ (1-p)P(x,y) + pD(x), & x = y \end{cases} \tag{6.9}$$

kde $D(x)$ je diskrétní rozdělení, $P(x,y)$ je sdružená pravděpodobnostní funkce, která je definovaná předešlými dvěma podkapitolami, a $p \in [0,1]$.

Budeme uvažovat obecné rozdělení $D(x)$, pro které platí $P(X = k) = \theta_k$ pro k = 0, 1, …, kde $\theta_k \geq 0$ pro všechna k a $\sum_{k=0}^{K} \theta_k = 1$. Karlis a Ntzoufras (2003) používali $K = 3$ jako postačující hodnotu pro fotbalová data. Naše data obsahují 463 remízových her (pro českou ligu), ale jen 4 z nich mají 6 nebo více gólů na každé straně. A tak bude užit jako dostatečný koeficient $K = 5$. Stejná hodnota tohoto parametru se užije i pro ostatní ligy.

Za použití rovnice (6.9) ve dvojnásobném Poissonovo modelu a dvourozměrném Poissonovo modelu nám budou poskytnuty dva nové modely (DP-DI a BP-DI). Tyto diagonálně rozšířené verze mají několik nových parametrů, tj. p a θ_k pro k = 0, 1, …, 5.

6.5 Odhad parametrů

V této části je čerpáno z článku [3].

Nejprve bylo potřeba ve všech modelech odhadnout všechny parametry pro poslední ukončenou sezónu, pro kterou máme dostupné výsledky ze všech utkání, tj. sezónu 2014/2015. Až na základě této sezóny je možné maximalizací $S(\xi)$ (rovnice (6.18)) určit optimální hodnotu parametru ξ potřebnou pro novou sezónu 2015/2016.

Parametry z modelů definovaných v předchozích kapitolách 6.2 až 6.4 se odhadují pomocí metody maximální věrohodnosti.

6.5.1 Věrohodnostní funkce

V kapitolách 6.2 až 6.4 jsme uvažovali dynamické modely, které do věrohodnostní funkce zavádějí i váhovou funkci $\tau(t_m)$.

Věrohodnostní funkce pro odhad parametru zápasů označovaných jako $m = 1, 2, \ldots, M$ ($m = 1$ je nejstarší dostupné utkání a $m = M$ je naopak nejnovější dostupný zápas)

a zaznamenané skóre mezi týmy *i* a *j* v *m*-tém zápasu (x_m, y_m) se definuje pro každý model z podkapitol 6.2 až 6.4.

Pro dvojnásobný Poissonovo (DP) model má věrohodnostní funkce tvar

$$L(\alpha_i, \beta_i, \gamma, \mu; i = 1, \dots, N; k = 0, \dots, K) = \prod_{m=1}^{M} P(x_m, y_m), \tag{6.10}$$

kde $P(x_m, y_m)$ je pravděpodobnostní funkce dvojnásobného Poissonovo rozdělení.

Pro dvourozměrný Poissonovo (BP) model, je věrohodnostní funkce následující

$$L(\alpha_i, \beta_i, \gamma, \mu, \lambda; i = 1, \dots, N; k = 0, \dots, K) = \prod_{m=1}^{M} P(x_m, y_m), \tag{6.11}$$

kde $P(x_m, y_m)$ je pravděpodobnostní funkce dvourozměrného Poissonovo rozdělení definovaná v rovnici (6.7).

Pro diagonálně rozšířený dvojnásobný Poissonovo (DP-DI) model je věrohodnostní funkce tato

$$L(\alpha_i, \beta_i, \gamma, \mu, p, \theta_k; i = 1, \dots, N; k = 0, \dots, K) = \prod_{m=1}^{M} P(x_m, y_m), \tag{6.12}$$

kde $P(x_m, y_m)$ je pravděpodobnostní funkce definovaná v rovnici (6.9), do které se dosadí pravděpodobnostní funkce dvojnásobného Poissonovo rozdělení definovaná v rovnici (6.6).

Pro diagonálně rozšířený dvourozměrný Poissonovo (BP–DI) model platí věrohodnostní funkce

$$L(\alpha_i, \beta_i, \gamma, \mu, \lambda, p, \theta_k; i = 1, \dots, N; k = 0, \dots, K) = \prod_{m=1}^{M} P(x_m, y_m), \tag{6.13}$$

kde $P(x_m, y_m)$ je pravděpodobnostní funkce definovaná v rovnici (6.9), do které se dosadí pravděpodobnostní funkce dvourozměrného Poissonovo rozdělení definovaná v rovnici (6.7).

Výkonnosti týmů se mohou v průběhu sezóny měnit, a tak užijeme podobné přiblížení jako Dixon a Coles (1997), což znamená, že odhadujeme parametr modelu pro každý čas, když se hraje nějaký zápas, spíše než konstantní parametry po celou sezónu. V této práci, stejně jako ve článku [2], zvolíme exponenciální funkci jako váhu pravděpodobnosti, a tak informace ze všech předchozích výsledků exponenciálně klesá. Dixon a Coles měřili čas v polo týdnech. Podle článku [3] však volíme časomíru v přesných údajích, protože se bere do úvahy také informace o přestávkách mezi zápasy, například čas mezi sezónami nebo některými krátkými přestávkami. Váhová funkce v čase odhadu *T* je

$$\tau(t_m) = \begin{cases} e^{-\xi \cdot (T - t_m)/365{,}25}, & t_m < T, \\ 0, & t_m \geq T, \end{cases} \tag{6.14}$$

kde t_m je čas zápasu, $\xi \geq 0$; $\xi = 0$ znamená, že všechny zápasy mají stejnou váhu.

Z rovnic (6.10) až (6.13) s použitím váhové funkce sestavujeme věrohodnostní funkci, kde je pravděpodobnostní funkce umocněna na $\tau(t_m)$

$$L(i = 1, \dots, N; k = 0, \dots, K) = \prod_{m=1}^{M} \{P(x_m, y_m)\}^{\tau(t_m)}, \tag{6.15}$$

m = 1, 2, ..., M (*m = 1* je nejstarší zápas a *m = M* je nejnovější zápas),

(x_m, y_m) je skóre mezi týmem *i* a *j* v *m*-tém zápase.

6.5.2 Logaritmická věrohodnostní funkce

Pro odhad parametrů není důležité absolutní číslo, ale pouze polohy bodů maxima, proto je možné věrohodnostní funkci zlogaritmovat, čímž se nezmění odhady parametrů.

Logaritmická věrohodnostní funkce pro DP a BP model má následující tvar

$$ln\, L = ln\left(\prod_{m=1}^{M} \{P(x_m, y_m)\}^{\tau(t_m)}\right) = \sum_{m=1}^{M} \tau(t_m) \cdot \ln(P(x_m, y_m)). \tag{6.16}$$

Pro diagonálně rozšířené modely (DP-DI a BP-DI) má logaritmická věrohodnostní funkce tvar

$$ln\, L = ln\left(\prod_{m=1}^{M} \{P_{DI}(x_m, y_m)\}^{\tau(t_m)}\right) = \sum_{m=1}^{M} \tau(t_m) \cdot ln(P_{DI}(x_m, y_m)), \tag{6.17}$$

kde $P_{DI}(x_m, y_m)$ je definovaná v rovnici (6.9).

6.5.3 Odhad parametru ξ

Parametr ξ nemůže být optimalizován rovnicí (6.15), protože tento postup by vedl k $\xi \to +\infty$. K získání optimální volby parametru ξ přijmeme postup používaný Dixonem a Colesem (1997). Tito autoři zvolili ξ tak, aby optimalizovali předvídání výsledků spíše než skóre zápasů. K nalezení nejlepší hodnoty parametru ξ definovali

$$S(\xi) = \sum_{m=1}^{M} (\delta_m^H \cdot \ln p_m^H + \delta_m^D \cdot \ln p_m^D + \delta_m^A \cdot \ln p_m^A), \tag{6.18}$$

kde δ_m je indikační funkce, pro kterou platí například:

$\delta_m^H = 1$, jestliže zápas *m* končí vítězstvím domácího týmu,

$\delta_m^H = 0$ v jiných případech,

p_m^H, p_m^D a p_m^A jsou pravděpodobnosti výhry domácího týmu spočítané podle rovnic (6.1), (6.2) a (6.3) za použití maximálně věrohodných odhadů z rovnice (6.15) s váhovým parametrem nastaveným na ξ.

Funkci $S(\xi)$ lze považovat jako měřítko předvídací chyby výsledků, protože pro výsledky s odhadnutou malou pravděpodobností dává velké záporné číslo, zatímco pro výsledky s odhadnutou vysokou pravděpodobností udává záporné číslo blížící se nule, tudíž lze říci, že maximalizace $S(\xi)$ vede k nejlepšímu odhadu parametru ξ. Jak bude ukázáno později, výsledek je robustní v celém okruhu hodnot parametru ξ.

Každý model byl sestaven z dat ze sezóny 2014/2015, a zkoumali jsme předvídací schopnost každého modelu. Nejvhodnější model se poté porovná se sázkovými kancelářemi na datech ze sezóny 2015/2016.

Pro každý model se užije následující ověřovací postup převzatý z [3]:

1) Soubor dat historických výsledků je rozdělen do dvou souborů. První soubor obsahuje všechny zápasy, které byly před sezónou 2014/2015 a druhý soubor obsahuje zápasy, které se odehrály právě v této sezóně. Předpokládejme, že výsledky všech zápasů v druhém souboru jsou neznámé.
2) Nejprve vybereme nejbližší datum, kdy se hrály nějaké zápasy s neznámým výsledkem v sezóně 2014/2015 a stanovíme *T*, které je použito v rovnici (6.14), rovné tomuto datu. (V této práci se odhaduje pro českou ligu až po uplynutí 5 hracích dní.)
3) Stanovíme *M* použité v rovnicích (6.10) až (6.13), jako počet zápasů odehraných před tímto datem *T*.
4) Použijeme model definovaný v rovnici (6.15) nebo jeho logaritmickou verzi a odhadneme příslušné parametry pro daný model užitím všech informací ze souboru historických výsledků.
5) Užijeme odhadované parametry k výpočtu pravděpodobností definovaných v rovnicích (6.1), (6.2) a (6.3) pro každý zápas odehraný v čase *T*.
6) Přesuneme všechny zápasy a jejich reálné výsledky odehrané v čase *T* ze souboru s neznámými výsledky do souboru s historickými výsledky.
7) Budou-li v sezóně stále zápasy s neznámým výsledkem, vrátíme se na 2. bod tohoto postupu a použijeme aktualizované soubory z předchozího bodu.

Tento postup se opakuje pro různé hodnoty parametru ξ a odpovídající $S(\xi)$ je počítán po uplynutí celé sezóny. Jak jsme zmiňovali již dříve, $S(\xi)$ může mít pouze záporné hodnoty, a čím více se hodnota $S(\xi)$ blíží k nule, tím lepší predikci získáme. Z toho důvodu hledáme maximální hodnotu parametru $S(\xi)$, což vede k nejlepšímu odhadu parametru ξ. Jak bude ukázáno dále, výsledek je robustní v celé řadě hodnot parametru ξ.

6.6 Model pro Extraligu (CZE)

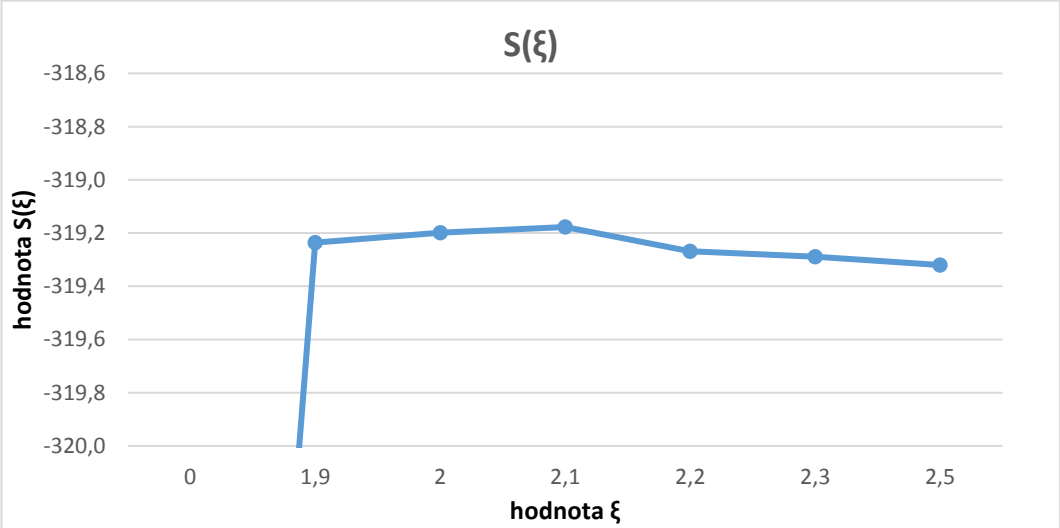
V této kapitole se budou odhadovat výsledky zápasů v sezóně 2015/2016 od 7. kola, tak aby každý tým v této sezóně hrál jako domácí tým alespoň dvakrát, tj. 34 zápasů je vynecháno. Je to z toho důvodu, že odhady mohou být kolísavé v první části sezóny, protože výkonnost jednoho mužstva před letní přestávkou a po ní se významně mění. Odhady parametrů od 7. kola se tedy uvažují již jako stabilní, což lze vidět na Obrázek 5 a Obrázek 6.

Výsledky pro odhad zápasů jsou získány od sezóny 2009/2010. V modelu nemá cenu pracovat se staršími zápasy vzhledem k váhové funkci $\tau(t_m)$ a parametru ξ, protože jejich váha by byla velmi malá. Abychom mohli odhadovat výsledky utkání, nejprve je třeba určit váhu jednotlivých historických zápasů, k čemuž potřebujeme parametr ξ získaný maximalizací funkce $S(\xi)$ definované v rovnici (6.18). Poté se tento parametr použije pro odhadovanou sezónu 2015/2016.

6.6.1 Parametr ξ pro sezónu 2014/2015

V kapitole 6.2 až 6.4 byly definovány čtyři modely, a to dvojnásobný Poissonovo model (DP), dvourozměrný Poissonovo model (BP), diagonálně rozšířený dvojnásobný Poissonovo model (DP–DI) a diagonálně rozšířený dvourozměrný Poissonovo model (BP–DI). Podle postupu z předchozí podkapitoly 6.5.3 se každý model užil na odhadování pravděpodobností během sezóny 2014/2015 a jejich síla je porovnána v této části. Výpočty pro všechny čtyři modely jsou v souboru *CZE_volba ksi.xlsm*.

Při určování hodnoty parametru ξ (čerpáno z článku [3]) odhadujeme parametry pro sezónu 2014/2015 od 6. kola. Nejdříve musíme určit hodnotu $S(\xi)$, tj. míry chyby predikce výsledků v celé sezóně. Jak již bylo zmíněno výše, parametr ξ váhové funkce je definován v rovnici (6.14). Funkce $S(\xi)$ nám umožňuje porovnat modely a vybrat z nich ten nejlepší. Na Obrázek 3 jsou uvedené hodnoty $S(\xi)$ proti ξ pro celou sezónu. Co se týče všech modelů, tak BP model poskytoval maximální hodnotu funkce $S(\xi)$. Optimální volba parametru ξ pro českou ligu je 2,1. U všech modelů závisí $S(\xi)$ na hodnotě parametru ξ a ve všech případech je optimální hodnota ξ 2,1 a výsledky nejsou citlivé na malé změny parametru ξ v okolí optimální hodnoty. Ostatní modely zaznamenaly stejnou závislost $S(\xi)$ na hodnotě parametru ξ, a ve všech případech je optimální hodnota ξ rovna 2,1. To znamená, že z pohledu funkce $S(\xi)$, všechny modely vytvářejí nejlepší odhady, když je parametr ξ nastaven na hodnotu 2,1. Po dosazení této hodnoty do rovnice (6.14) můžeme říci, že váha výsledků zápasů starých jeden rok je 12,26 %, zatímco stejné výsledky před dvěma lety mají tuto váhu 1,50 %.



Obrázek 3: Hodnoty $S(\xi)$ proti ξ s maximem v bodě 2,1 pro BP model

V následující tabulce je přehled hodnot $S(\xi)$ jednotlivých modelů pro různé hodnoty parametru ξ. V tabulce je vidět, že mezi jednotlivými modely jsou nepatrné rozdíly v hodnotách statistiky $S(\xi)$ v optimálním bodě ξ, a jsou označené žlutě. Na základě maximalizace $S(\xi)$ byl vybrán model BP, který vyšel s hodnotou -319,177, tj. nepatrně větší hodnotou oproti ostatním modelům, a v tabulce je označen zeleně.

ksí	0,0	1,9	2,0	2,1	2,2	2,3	2,5
S(ksí) - DP	-325,240	-319,321	-319,311	-319,306	-319,309	-319,321	-319,368
S(ksí) - BP	-325,235	-319,236	-319,199	-319,177	-319,268	-319,289	-319,320
S(ksí) - DP-DI	-327,507	-319,980	-319,923	-319,476	-320,240	-320,252	-320,274
S(ksí) - BP-DI	-327,259	-319,865	-319,564	-319,500	-319,529	-319,541	-319,583

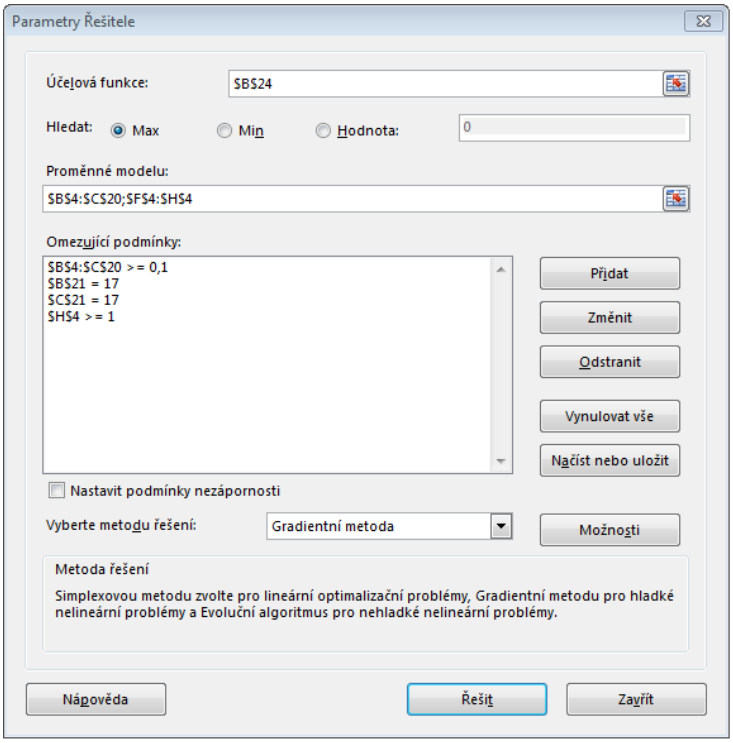
Tabulka 10: Hodnoty $S(\xi)$ pro jednotlivé modely

6.6.2 Odhad parametrů BP model

Všechny prezentované výsledky jsou založeny na nejlepším získaném modelu, což je podle předchozích výsledků BP model s parametrem ξ nastaveným na 2,1.

Jak již bylo uvedeno výše, odhady parametrů mohou být nestabilní v první části sezóny. Z tohoto důvodu jsme vynechali odhad parametrů pro prvních 6 kol. Parametry tedy odhadujeme od 7. kola. Parametry odhadnuté v tomto kole se následně použijí k odhadování pravděpodobností 8. kola (výhra domácích, remíza a výhra hostů).

Odhadování výsledků české ligy je vytvořeno v sešitu *CZE_BP 15-16.xlsm* na listu *BP model*. Odhad je proveden maximalizací věrohodnostní funkce (rovnice (6.21)), která je v buňce *B24*. K maximalizaci se používá řešitel[6], ve kterém je nastavena přesnost omezující podmínky na 0,00001, tj. pokud se žádný z parametrů nezmění o více než 0,00001, výpočet skončí.



Obrázek 4: Nastavení řešitele pro BP model v Microsoft Excel 2013

Při výpočtu se mění parametry síly útoku α_i (buňky *B4:B20*) a obrany β_i (buňky *C4:C20*) pro všechny týmy i, dále se mění parametr výhody domácího prostředí γ (buňka *F4*),

[6] Byla použita Gradientní metoda pro řešení nelineárních problémů (v Microsoft Excel 2010 se tato metoda nazývá GRG Nonlinear), více o této metodě lze nalézt v [11].

parametr určující závislost mezi góly domácích a hostujících týmů λ (buňka *G4*) a parametr μ (buňka *H4*).

Parametry α_i, β_i, γ a μ jsou pro všechny týmy i nezáporná reálná čísla, což vyplývá z logaritmické věrohodnostní funkce (6.21) a z významu parametrů α a β, které vyjadřují sílu útoku a obrany. Parametr λ je reálné číslo, které může být oproti ostatním uvedeným parametrům i záporné. Pro parametry α_i a β_i jsou nastavené podmínky z podkapitoly 6.2.1.

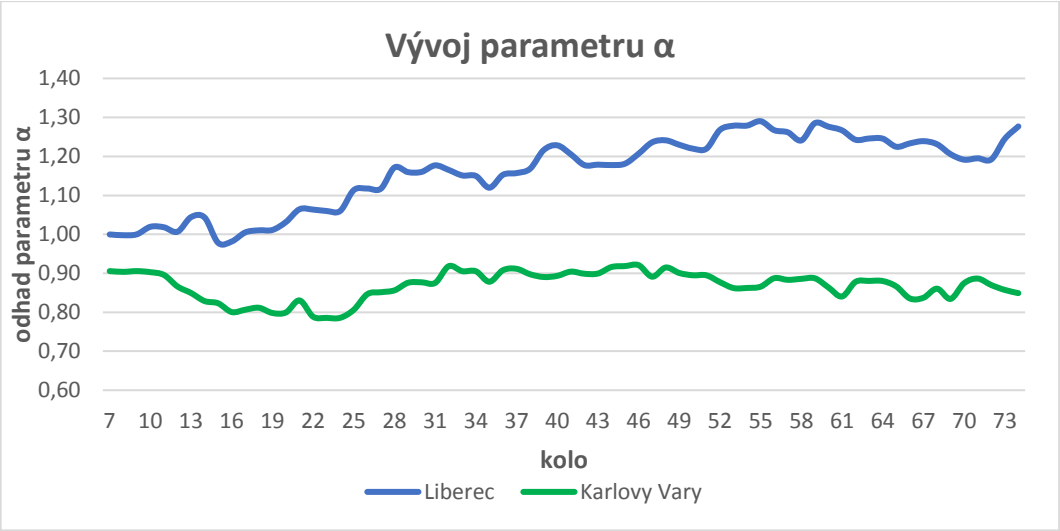
Před spuštěním řešitele, je třeba nastavit počáteční hodnoty parametrů. V tomto modelu byly nastaveny pro 7. kolo všechny parametry α, β, γ, λ a μ na 1. Pro další kola se vždy jako počáteční hodnoty používají hodnoty z předcházejícího kola.

Odhadnuté parametry 7. až 74. kola jsou uvedené v souboru *CZE_BP model 15-16.xlsm* na listu *BP model_parametry*. V následující tabulce jsou uvedené parametry α a β 73. kola použité k odhadu 74. kola.

Tým	alfa	beta
České Budějovice	0,93	1,04
Hradec Králové	1,03	0,96
Chomutov	1,03	1,02
Karlovy Vary	0,86	1,24
Kladno	0,68	1,11
Kometa Brno	0,99	1,05
Liberec	1,24	0,79
Litvínov	0,84	0,93
Mladá Boleslav	1,15	1,00
Olomouc	0,84	0,83
Pardubice	0,90	1,07
Plzeň	1,18	0,98
Slavia Praha	0,82	1,28
Sparta Praha	1,50	0,92
Třinec	0,95	0,85
Vítkovice	1,08	1,07
Zlín	0,98	0,87
Celkem	**17,00**	**17,00**

Tabulka 11: Odhad parametrů pro 74. kolo (tj. z výsledků do 73. kola včetně)

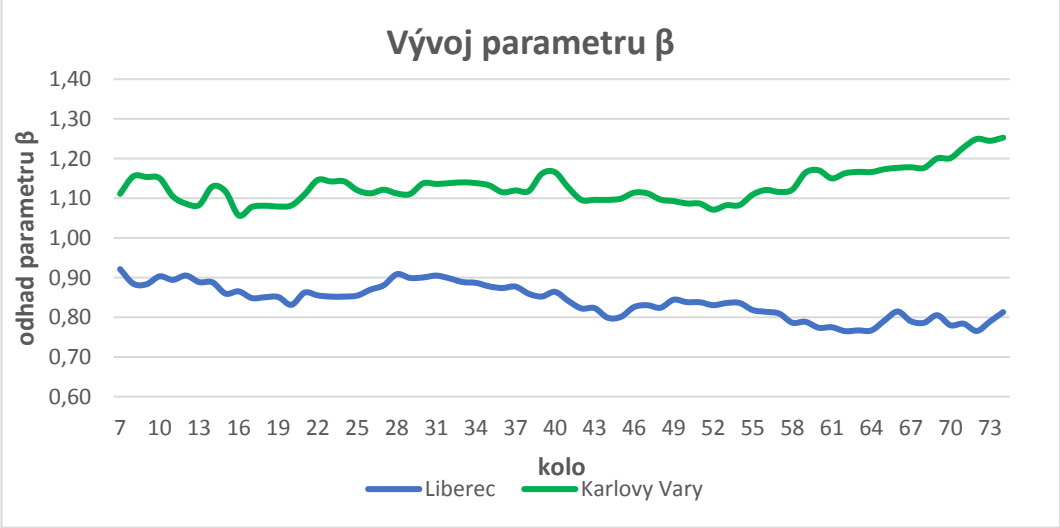
Na Obrázek 5 je vidět, jak se mění parametr útoku α během celé sezóny 2015/2016 pro nejlepší tým sezóny Liberec, a nejhorší mužstvo Karlovy Vary. Jak již bylo uvedeno výše, parametry se odhadovaly pokaždé, když se hrál zápas, a proto se získala pro každý parametr časová posloupnost 68 záznamů (pro 7. až 74. kolo).



Obrázek 5: parametru útoku α pro nejlepší a nejhorší tým

Jak bylo znázorněno na obrázku 4, odhady jsou v souladu se skutečností, že Liberec měl lepší útočný výkon než Karlovy Vary, a to pro celou sezónu. Parametr síly útoku α pro tým Liberec měl rostoucí trend po celou sezónu. Od 7. kola po 74. kolo jsou odhady parametrů téměř stabilní. To je způsobeno definicí váhové funkce dané v rovnici (6.14), která bere v úvahu čas mezi koncem předchozí sezóny a začátkem současné sezóny.

Dobrá hodnota parametru pro útok ještě není zárukou toho, že mužstvo bude úspěšné, protože musí být také podpořeno parametrem obrany β. Vývoj odhadu parametru obrany pro Liberec, nejlepší tým sezóny, a Karlovy Vary, nejhorší tým, je znázorněn na Obrázek 6.



Obrázek 6: Maximálně věrohodné odhady parametru obrany β pro nejlepší a nejhorší tým

Jak již bylo uvedeno, čím nižší hodnota tohoto parametru, tím je to lepší pro mužstvo. Ve druhé části sezóny je zřejmé, že Liberec zlepšil svoji obranu a snížil tak hodnotu parametru pro obranu ještě více pod průměr, tj. pod hodnotu 1. Naopak Karlovy Vary zhoršily svoji obranu a zvýšily hodnotu β až na 1,2.

Další tři parametry jsou globální pro všechny týmy. Odhady parametru výhody domácího prostředí γ a konstantní parametr μ se téměř nemění v průběhu času, protože popisují celou ligu, kterou tvoří stejná mužstva v jedné sezóně. Větší změny těchto parametrů

se mohou očekávat mezi sezónami. Průměrná hodnota odhadu parametru γ je 1,232 a znamená to, že domácí tým skóruje přibližně 1,232 krát více gólů než hostující tým. Průměrná hodnota odhadu parametru μ je 2,289, a to lze chápat jako průměrný počet gólů hostujícího mužstva. Jestliže se tento parametr násobí odhadem parametru γ, může to být uvažováno jako průměrný počet gólů domácího týmu, tj. 2,820. Průměrná hodnota odhadu parametru λ je -0,489, což znamená, že korelační koeficient ρ je záporný (dle rovnice (6.8)), tj. X_{ij} a Y_{ij} jsou závislé náhodné veličiny.

6.6.3 Odhad výsledků zápasů

Výsledky zápasů lze odhadnout pomocí sdružené pravděpodobnostní funkce pro BP model definované v rovnici (6.7). Pravděpodobnosti konkrétních výsledků zápasů se vypočítávají na listu *BP sezóna 2015-2016* (sloupce *AB* až *JY*) s použitím odhadnutých parametrů z listu *BP model_parametry*. Poté se vyjádří celkové pravděpodobnosti výher domácích, remíz a výher hostů, které jsou zaznamenány na listu *BP sezóna 2015-2016* ve sloupcích *K* až *M*.

Odhad parametrů pro zápas posledního hracího dne sezóny (tj. 74. „kola“) mezi týmem Plzeň a Chomutov jsou zobrazeny v následující tabulce.

Parametry	Domácí - Plzeň	Hosté - Chomutov
α	1,185	1,031
β	0,977	1,024
γ	1,284	
λ	-0,867	
μ	2,241	

Tabulka 12: Odhadnuté parametry pro zápas Plzeň - Chomutov

Pro tento zápas je odhad průměrného počtu gólů domácích (λ_{PLZ}) a průměrného počtu gólů hostů (λ_{CHO}) uveden v následujících rovnicích

$$\hat{\lambda}_{PLZ} = \hat{\mu} \cdot \hat{\alpha}_{PLZ} \cdot \hat{\beta}_{CHO} \cdot \hat{\gamma} = 2{,}241 \cdot 1{,}185 \cdot 1{,}024 \cdot 1{,}284 = 3{,}491, \tag{6.19}$$

$$\hat{\lambda}_{CHO} = \hat{\mu} \cdot \hat{\alpha}_{CHO} \cdot \hat{\beta}_{PLZ} = 2{,}241 \cdot 1{,}031 \cdot 0{,}977 = 2{,}257. \tag{6.20}$$

Známe-li tyto parametry, můžeme je dosadit do pravděpodobnostní funkce pro BP model (rovnice (6.7)) a určit tak pravděpodobnost konkrétního výsledku.

Tento zápas skončil 3:1 výhrou pro domácí Plzeň, proto v následující rovnici uvádíme pravděpodobnost tohoto výsledku určeného před zápasem dle zvoleného BP modelu

$$P(X = 3, Y = 1) = (3{,}491^3 \cdot 2{,}257^1 \cdot e^{-3{,}491-2{,}257}) \cdot$$
$$\cdot \frac{1 - 0{,}867 \cdot \left(e^{-3} - e^{-\left(1-e^{-1}\right)\cdot 3{,}491}\right) \cdot \left(e^{-1} - e^{-\left(1-e^{-1}\right)\cdot 2{,}257}\right)}{3! \cdot 1!} = 0{,}051 \tag{6.21}$$

Stejným způsobem se vypočítává pravděpodobnost pro všechny možné výsledky až do výsledku 15:15 (teoreticky však až do výsledku ∞:∞).

V následující tabulce jsou pravděpodobnosti pro jednotlivé výsledky tohoto zápasu.

Počet gólů domácí / hosté		Chomutov							
		0	1	2	3	4	5	6	7 a více
Plzeň	0	0,001	0,006	0,009	0,007	0,004	0,002	0,001	0,000
	1	0,009	0,024	0,029	0,022	0,013	0,006	0,002	0,001
	2	0,019	0,044	0,050	0,037	0,021	0,010	0,004	0,002
	3	0,024	0,051	0,057	0,043	0,024	0,011	0,004	0,002
	4	0,021	0,045	0,050	0,037	0,021	0,009	0,004	0,002
	5	0,015	0,031	0,035	0,026	0,015	0,007	0,002	0,001
	6	0,009	0,018	0,020	0,015	0,008	0,004	0,001	0,001
	7 a více	0,007	0,015	0,017	0,013	0,007	0,003	0,001	0,000

Tabulka 13: Pravděpodobnosti výsledků pro zápas Plzeň - Chomutov

Výsledek		
Výhra domácí [1]	Remíza [0]	Výhra hosté [2]
0,617	0,148	0,235

Tabulka 14: Pravděpodobnost výhry domácích, remízy a výhry hostů

V předchozí tabulce jsou uvedené pravděpodobnosti výsledku tohoto zápasu, kde největší pravděpodobnost je, že vyhraje domácí tým Plzeň. Konkrétní výsledek zápasu nepotřebujeme znát a ani se ho nesnažíme odhadnout, chceme jen vědět, pravděpodobnosti konečného stavu zápasu.

6.7 Model pro Ekstraligu (POL)

Stejně jako v předchozí kapitole, i zde se budou odhadovat výsledky zápasů v sezóně 2015/2016, ale nyní pro polskou ligu. Odhadování výsledků pro polskou ligu od 8. kola je provedeno v sešitu *POL_BP 15-16.xlsm* na listu *BP model*. V této lize odhadujeme veškeré parametry až od 8. kola ze stejného důvodu jako v případě české ligy. Vynecháváme 36 zápasů, tak aby v této sezóně hrál tým, který nehrál v předchozí sezóně, alespoň třikrát doma. Jedná se o týmy Toruň a Zagłębie Sosnowiec.

Ale ještě před samotným odhadováním sezóny 2015/2016 je potřeba, stejně jako u české ligy, vhodně zvolit hodnotu parametru ξ na sezóně 2014/2015 a poté tento parametr použít při odhadování sezóny 2015/2016.

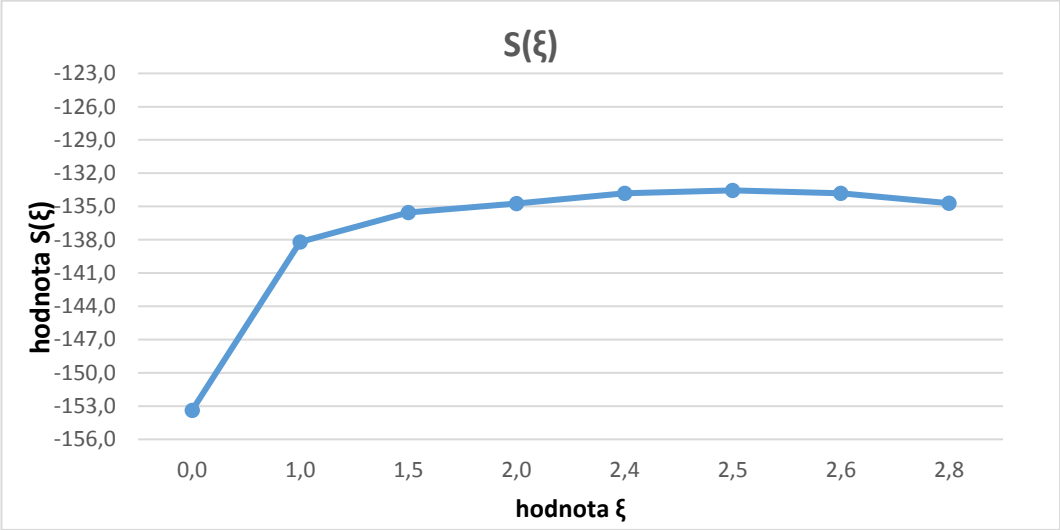
6.7.1 Parametr ξ pro sezónu 2014/2015

Při určování hodnoty parametru ξ odhadujeme parametry pro sezónu 2014/2015 od 7. kola, tj. vynechali jsme 20 zápasů. Výpočty pro všechny čtyři modely jsou v souboru *POL_volba ksi.xlsm*.

Zde uvádíme výsledky pro funkci $S(\xi)$, která je, jak již bylo uvedeno výše, definována v rovnici (6.18). Na obrázku 8 jsou uvedené hodnoty $S(\xi)$ proti ξ pro celou sezónu. Co se týče všech modelů, tak BP-DI model sice poskytoval maximální hodnotu funkce $S(\xi)$, ale oproti BP modelu byla tato hodnota pouze o 0,008 větší a odhady parametru p pro jednotlivá kola vycházely většinou nulové. To znamená, že parametr p nemá téměř žádný vliv na zvýšení pravděpodobností remíz. Z tohoto důvodu byl zvolen jako nejvhodnější BP model, a tento závěr potvrzuje i skutečnost, že počet remíz v polské

lize je velmi nízký, například v sezóně 2014/2015 jich bylo pouze 22.

U všech modelů závisí $S(\xi)$ na hodnotě parametru ξ. V modelu DP a DP-DI je optimální hodnota ξ 2,6 a pro model BP a BP-DI vychází hodnota tohoto parametru 2,5. Po dosazení této hodnoty do rovnice (6.14) můžeme říci, že váha výsledků zápasů starých jeden rok je 8,22 %, zatímco stejné výsledky před dvěma lety mají tuto váhu 0,68 %.



Obrázek 7: Hodnoty $S(\xi)$ proti ξ pro BP model – Ekstraliga (POL)

V následující tabulce je uveden přehled hodnot $S(\xi)$ jednotlivých modelů pro různé hodnoty parametru ξ. Z tabulky je zřejmé, že mezi jednotlivými modely jsou velmi malé rozdíly v hodnotách $S(\xi)$ v optimálních bodech ξ, které jsou vyznačené žlutě. Pro zvolený BP model je optimální hodnota ξ označena v tabulce zelenou barvou.

ksí	0,0	1,0	2,0	2,4	2,5	2,6	2,7	3,0
S(ksí) - DP	-153,365	-138,206	-133,982	-133,584	-133,555	-133,550	-133,566	-133,724
S(ksí) - BP	-153,372	-138,204	-134,731	-133,818	-133,549	-133,802	-134,051	-134,709
S(ksí) - DP-DI	-153,484	-138,456	-133,987	-133,585	-133,557	-133,555	-133,575	-133,712
S(ksí) - BP-DI	-153,340	-139,035	-134,354	-133,573	-133,541	-133,546	-133,616	-133,700

Tabulka 15: Hodnoty $S(\xi)$ pro jednotlivé modely – Ekstraliga (POL)

6.7.2 Odhad parametrů BP model

V této části jsou všechny výsledky založeny na nejvhodnějším modelu, což je podle předchozí kapitoly BP model s parametrem ξ nastaveným na 2,5.

Odhad je proveden maximalizací věrohodnostní funkce (6.17), která je v buňce *B22*. K maximalizaci se používá opět řešitel, ve kterém je nastavena stejná přesnost omezující podmínky jako u české ligy. Před spuštěním řešitele, je třeba zvolit počáteční hodnoty parametrů, které byly v tomto modelu pro 8. kolo nastavené stejně jako u české ligy.

Při výpočtu se opět mění parametry α_i (buňky *B4:B19*), β_i (buňky *C4:C19*) pro všechny týmy i, dále se mění parametr γ (buňka *F4*), λ (buňka *G4*) a parametr μ (buňka *H4*).

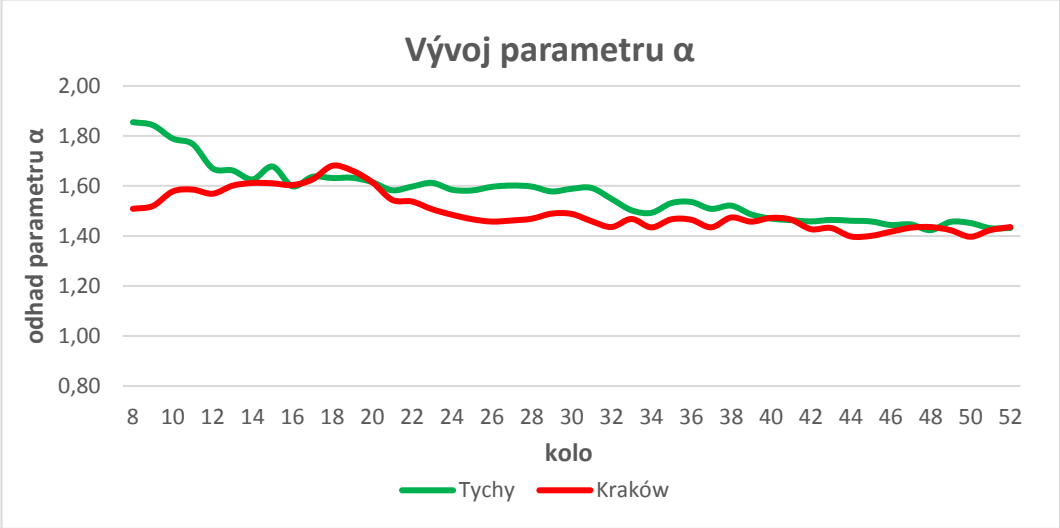
Tým	alfa	beta
Bytom	1,03	0,92
Janów	0,76	1,12
Jastrzebie JKH GKS	1,02	0,53
Katowice	0,70	1,99
Kraków	1,42	0,49
Krynica KTH	0,74	1,29
Orlik Opole	1,04	0,86
Podhale Nowy Targ	1,31	0,60
Sanok	1,07	0,54
Sosnoviec SMS	0,47	1,90
Stoczniowiec Gdańsk	1,30	0,94
Toruň	0,86	1,38
Tychy	1,43	0,47
Unia Oświęcim	1,07	0,78
Zagłębie Sosnowiec	0,78	1,21
Celkem	15,00	15,00

γ	λ	μ
1,13	0,00	3,70

Obrázek 8: Odhad parametrů pro 52. kolo (tj. z výsledků do 51. kola včetně)

Parametry α_i, β_i, γ, μ, jsou pro všechny týmy i nezáporná reálná čísla. Parametr λ je reálné číslo, které může být záporné. Pro parametry α_i a β_i jsou nastavené podmínky z podkapitoly 6.2.1.

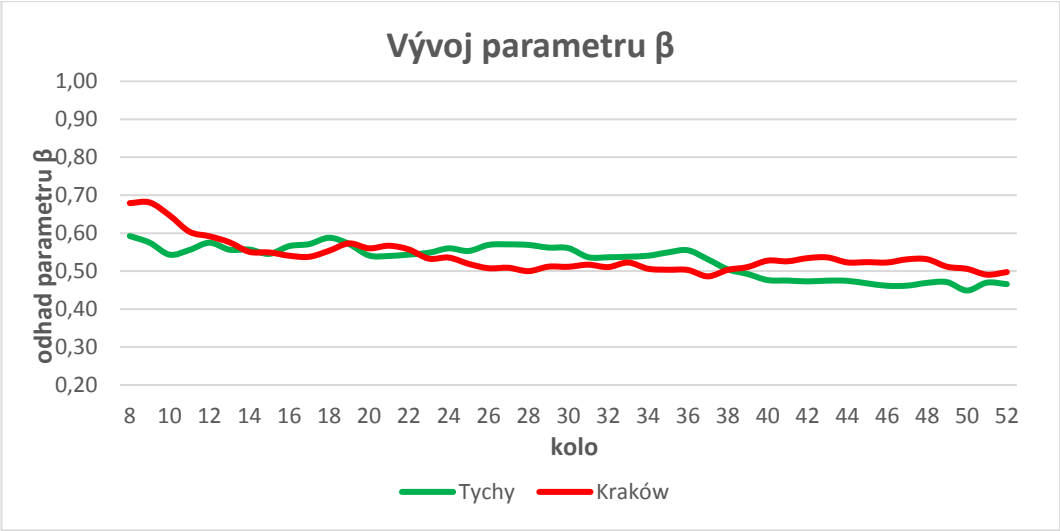
Na obrázku 9 je vývoj parametru útoku α během celé sezóny 2015/2016 pro dva nejlepší týmy sezóny – Kraków a Tychy. Parametry se odhadovaly pokaždé, když se hrál zápas, a získala se tak časová posloupnost 45 záznamů (pro 8. až 52. kolo).



Obrázek 9: Maximálně věrohodné odhady parametru útoku α pro dva nejlepší týmy

Jak bylo znázorněno na Obrázek 9, odhady odpovídají skutečnosti, že oba týmy měly od 13. kola až do konce sezóny velmi vyrovnanou sílu v útoku. Od 8. kola po 52. kolo jsou odhady parametrů stabilní.

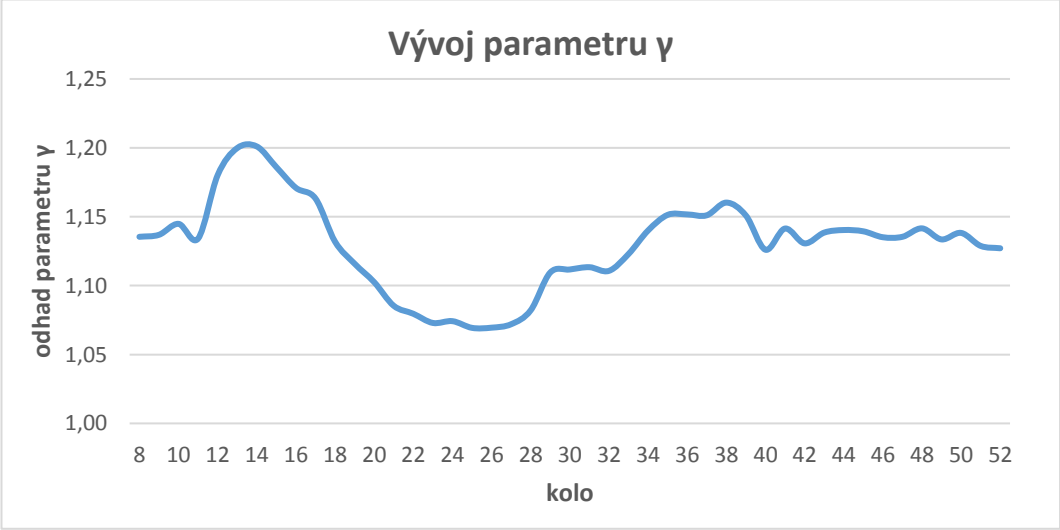
Vývoj odhadu parametru obrany pro Kraków, nejlepší tým sezóny, a Tychy, 2. nejlepší tým, je zobrazen na Obrázek 10.



Obrázek 10: Maximálně věrohodné odhady parametru obrany β pro dva nejlepší týmy

Od 39. kola Tychy zlepšil svoji obranu a snížil tak hodnotu parametru síly v obraně až na hodnoty pod 0,5. Naopak Kraków zhoršil svoji obranu, a proto se zvýšila hodnota β.

Na následujícím obrázku je zobrazen vývoj parametru γ. Od 15. do 27. kola parametr převážně klesal, což znamená, že se snižovala výhoda domácího prostředí.



Obrázek 11: Vývoj parametru γ

Průměrná hodnota odhadu parametru γ je 1,130 tj. domácí tým skóruje přibližně 1,130 krát více gólů než hostující tým. Průměrná hodnota odhadu parametru μ je 3,553. Jestliže vynásobíme tento parametr odhadem parametru γ, může to být uvažováno jako průměrný počet gólů domácího týmu, tj. 4,014. Průměrná hodnota odhadu parametru λ je -0,148, to znamená, že korelační koeficient ρ je záporný (dle rovnice (6.8)), tj. X_{ij} a Y_{ij} jsou dvě závislé náhodné veličiny. Nyní když jsou známé všechny parametry, můžeme odhadnout výsledky zápasů pomocí sdružené pravděpodobnostní funkce pro BP model. Výsledky zápasů se odhadují na listu *BP sezóna 2015-2016* (sloupce *AB* až *JY*) a pravděpodobnosti výher domácích, remíz a výher hostů jsou zaznamenány na tomto listu ve sloupcích *K* až *M*.

6.8 Model pro NHL

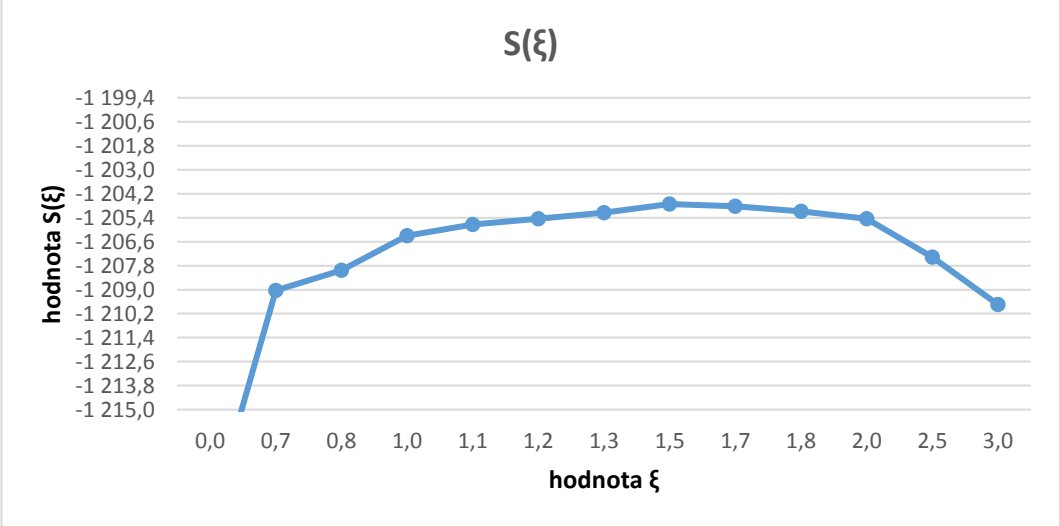
V této kapitole se budou odhadovat výsledky zápasů pro stejnou sezónu jako u ostatní lig, tentokrát však od 16. kola, tak aby v sezóně 2015/2016 každý tým opět hrál jako domácí alespoň dvakrát, tj. 90 zápasů je vynecháno. Ale hlavní důvod proč parametry pro zápasy do 15. kola neodhadujeme je vysvětlen již u české ligy v kapitole 6.6.

Odhadování výsledků NHL ligy pro nejnovější sezónu je provedeno v sešitu *NHL_BP-DI 15-16.xlsm* na listu *BP-DI model*.

Zde je stejný způsob odhadování jako u předchozích lig, tj. nejprve musíme zvolit hodnotu parametru ξ na předchozí sezóně.

6.8.1 Parametr ξ pro sezónu 2014/2015

Při určování hodnoty parametru ξ odhadujeme parametry pro sezónu 2014/2015 od 15 kola, tzn., že neodhadujeme parametry pro 88 zápasů. Výpočty jsou rozdělené do čtyř souborů nazvaných podle konkrétního modelu, například *NHL_DP_odhad ksi.xlsm*, atd. Nejdříve opět určíme hodnotu $S(\xi)$. Na obrázku 13 jsou hodnoty $S(\xi)$ pro různé ξ pro celou sezónu. Porovnáme-li všechny čtyři modely na základě hodnoty $S(\xi)$, tak model BP-DI poskytoval maximální hodnotu $S(\xi)$. Optimální volba parametru ξ pro BP-DI a DP-DI model je 1,5. V modelu BP je optimální hodnota ξ rovna 1,9 a u modelu DP tato hodnota vychází 0,2. Optimální hodnota 1,5 pro BP-DI model v našem případě znamená, že váha výsledků zápasů starých jeden rok je 22,34 %, kdežto stejných výsledků před dvěma lety pouze 4,99 %.



Obrázek 12: Hodnoty $S(\xi)$ proti ξ pro BP-DI model

V následující tabulce je přehled hodnot $S(\xi)$ jednotlivých modelů pro různé hodnoty parametru ξ. Mezi modely jsou výrazné rozdíly v hodnotách $S(\xi)$ v optimálním bodě ξ, a jsou zvýrazněné žlutě. Na základě maximalizace $S(\xi)$ byl vybrán model BP-DI, který vyšel s hodnotou -1 204,719, tj. o téměř 600 vyšší hodnota oproti BP modelu, a v tabulce je vyznačen zeleně.

ksí	0,0	1,0	1,2	1,3	1,4	1,5	1,6	1,8	2,0	2,5	3,0
S(ksí) - DP-DI	-1 220,468	-1 206,284	-1 205,308	-1 204,972	-1 204,808	-1 204,721	-1 204,759	-1 205,050	-1 205,529	-1 207,470	-1 209,881
S(ksí) - BP-DI	-1 220,511	-1 206,302	-1 205,447	-1 205,156	-1 204,815	-1 204,719	-1 204,737	-1 205,087	-1 205,451	-1 207,379	-1 209,744

Tabulka 16: Hodnoty $S(\xi)$ pro diagonálně rozšířené modely – NHL

ksí	0,0	0,1	0,2	0,3	0,4	0,5	1,0	2,0	2,1	2,2	2,6
S(ksí) - DP	-1 809,837	-1 809,213	-1 808,961	-1 809,082	-1 809,493	-1 810,149	-1 814,499	-1 821,466	-1 822,094	-1 822,719	-1 825,198

Tabulka 17: Hodnoty $S(\xi)$ pro DP model – NHL

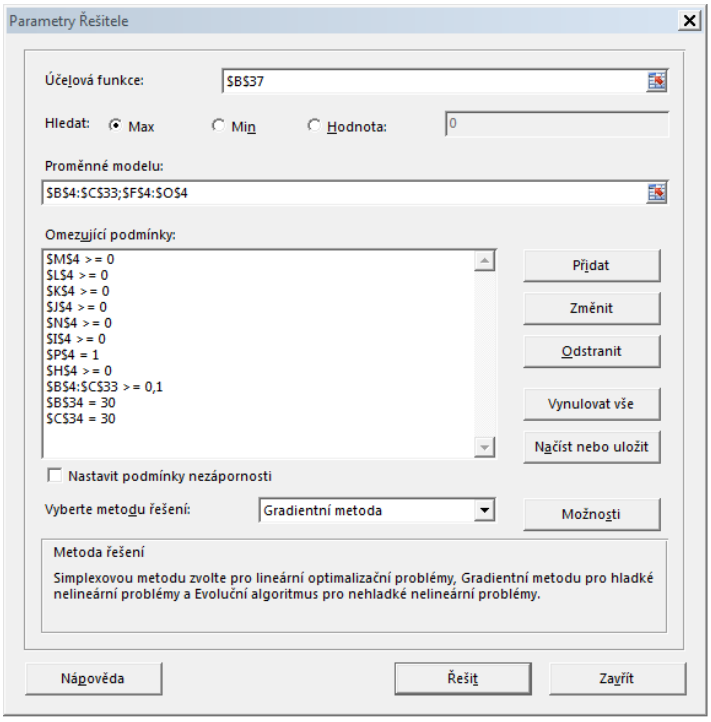
ksí	0,0	0,4	0,5	1,0	1,5	1,7	1,8	1,9	2,0	2,1	2,5	2,8
S(ksí) - BP	-1 809,331	-1 807,524	-1 807,151	-1 801,833	-1 798,949	-1 798,957	-1 798,168	-1 798,106	-1 798,310	-1 798,980	-1 799,419	-1 800,246

Tabulka 18: Hodnoty $S(\xi)$ pro BP model – NHL

6.8.2 Odhad parametrů BP-DI model

Všechny uvedené výsledky pro NHL ligu jsou založeny na nejlepším získaném modelu, což je podle předchozí podkapitoly BP-DI model s parametrem ξ nastaveným na 1,5.

Odhad je proveden maximalizací věrohodnostní funkce (6.17), která je v buňce *B37*. K maximalizaci se používá opět řešitel, ve kterém je nastavena stejná přesnost omezující podmínky jako u předchozích lig.



Obrázek 13: Nastavení řešitele pro BP-DI model v Microsoft Excel 2013

Při výpočtu se mění parametry α_i (buňky *B4:B32*) a β_i (buňky *C4:C32*) pro všechny týmy i, dále se mění parametr γ (buňka *F4*), parametr λ (buňka *G4*), parametr μ (buňka *H4*), parametr p (buňka *I4*) a parametry $\theta_{0,\dots,5}$ (buňky *J4: O4*).

Tým	α	β
Anaheim	1,02	0,89
Atlanta	0,98	1,04
Boston	1,06	1,03
Buffalo	0,87	1,02
Calgary	1,06	1,16
Carolina	0,91	1,01
Colorado	0,97	1,09
Columbus	1,01	1,12
Dallas	1,20	1,06
Detroit	0,97	1,02
Edmonton	0,91	1,16
Florida	1,04	0,95
Chicago	1,05	0,92
Los Angeles	0,98	0,88
Minnesota	0,99	0,90
Montreal	0,96	1,06
Nashville	1,04	0,93
New Jersey	0,79	0,94
NY Islanders	1,05	0,98
NY Rangers	1,09	0,95
Ottawa	1,05	1,08
Philadelphia	0,97	0,96
Phoenix	0,89	1,11
Pittsburgh	1,12	0,91
San Jose	1,07	0,97
St. Louis	1,03	0,88
Tampa Bay	1,05	0,92
Toronto	0,89	1,11
Vancouver	0,87	1,08
Washington	1,12	0,90
Celkem	30,00	30,00

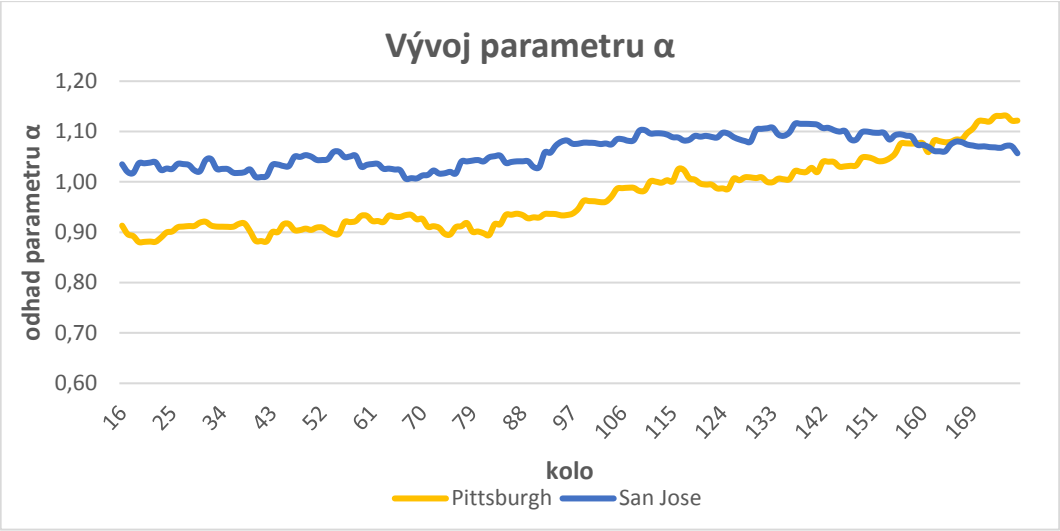
γ	λ	μ	p	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	suma θ_k
1,09	-1,59	2,57	0,08	0,10	0,39	0,27	0,16	0,07	0,00	1,00

Obrázek 14: Odhad parametrů pro 177. kolo (tj. z výsledků do 176. kola včetně)

Parametry α_i, β_i, γ, μ, jsou pro všechny týmy i nezáporná reálná čísla. Parametr $p \in [0;1]$ a θ_k pro k = 0,…,5 platí $\theta_k \geq 0$ pro všechna k, $\sum_{k=0}^{5} \theta_k = 1$. Parametr λ je reálné číslo, které může být záporné.

Před spuštěním řešitele, je třeba nastavit počáteční hodnoty parametrů. V tomto modelu byly nastaveny pro 16. kolo parametry α, β, γ, λ a μ na 1, p na 0,5, $\theta_0 = 0{,}35$, $\theta_1 = 0{,}30$, $\theta_2 = 0{,}20$, $\theta_3 = 0{,}10$, $\theta_4 = 0{,}05$ a $\theta_5 = 0$. Pro další kola se vždy jako počáteční hodnoty používají hodnoty z předcházejícího kola.

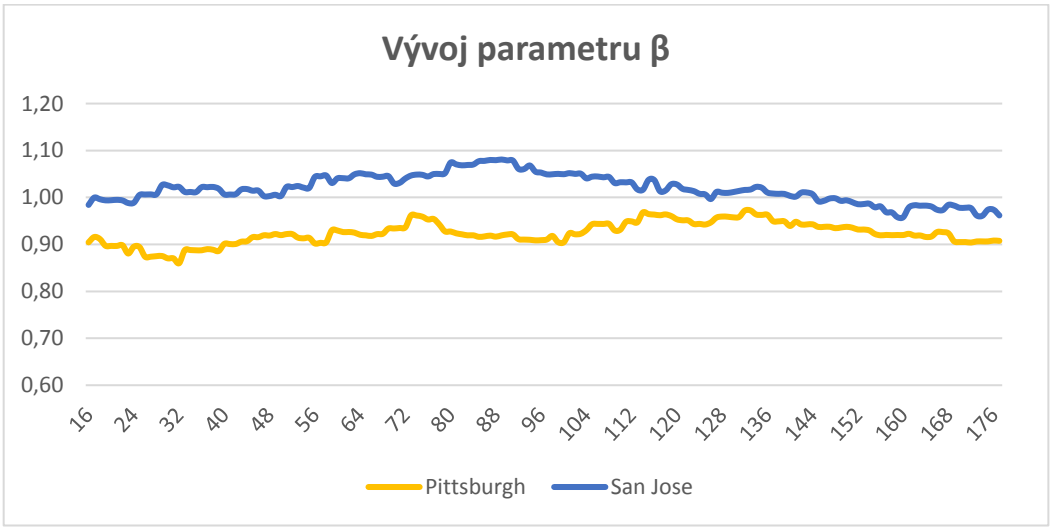
Na Obrázek 15 je vidět, jak se mění parametr útoku α během celé sezóny 2015/2016 pro dva nejlepší týmy sezóny, Pittsburgh a San Jose. Jak již bylo uvedeno výše, parametry se odhadovaly pokaždé, když se hrál zápas, a tak se získala pro každý parametr časová posloupnost 162 záznamů (pro 16. až 177. kolo).



Obrázek 15: Maximálně věrohodné odhady parametru útoku α pro dva nejlepší týmy

Parametr síly útoku α pro tým Pittsburgh měl rostoucí trend po celou sezónu. Tým San Jose byl téměř celou sezónu lepší v útoku než Pittsburgh.

Vývoj odhadu parametru obrany pro Pittsburgh, nejlepší tým sezóny, a San Jose, 2. nejlepší tým, je znázorněn na obrázku 16.

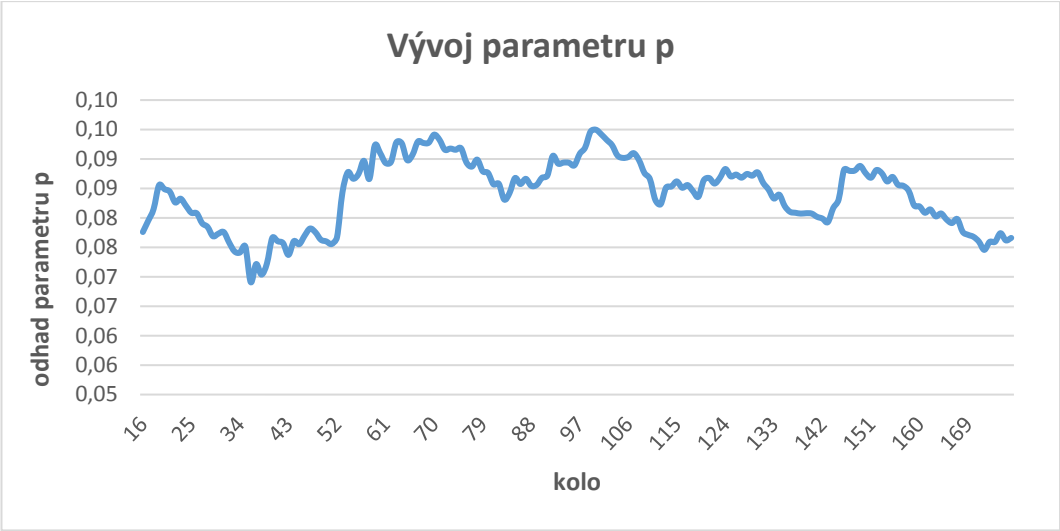


Obrázek 16: Maximálně věrohodné odhady parametru obrany β pro dva nejlepší týmy

Z předchozího obrázku je zřejmé, že Pittsburgh byl silnější v obraně oproti San Jose.

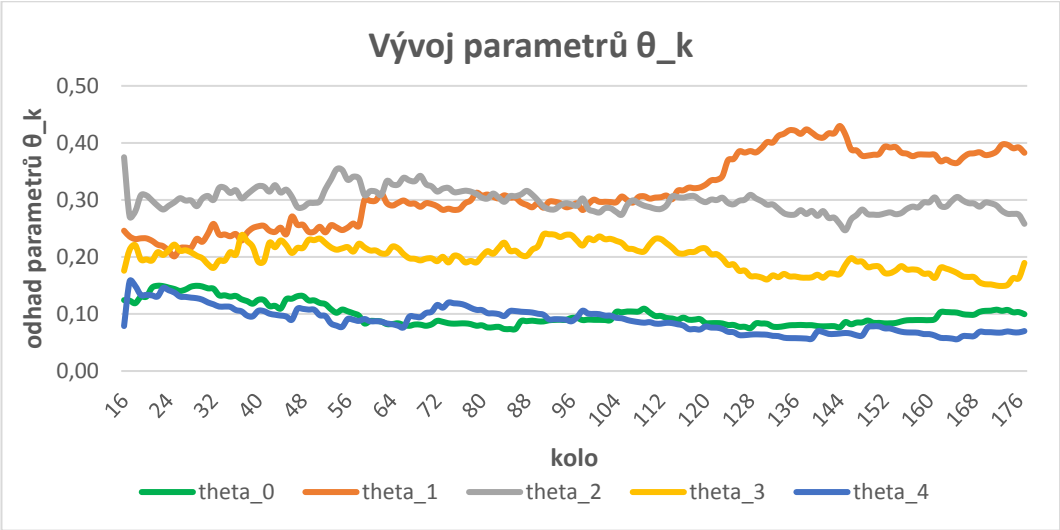
Odhady parametru výhody domácího prostředí γ a parametr μ se téměř nemění během jednotlivých kol, protože popisují celou ligu, kterou tvoří v jedné sezóně stejná mužstva. Průměrná hodnota odhadu parametru γ je 1,091. Průměrná hodnota odhadu parametru μ je 2,533, což lze chápat jako průměrný počet gólů hostujícího týmu. Průměrný počet gólů domácího týmu je 2,764. Průměrná hodnota odhadu parametru λ je -1,275, tudíž korelační koeficient ρ je záporný, tj. X_{ij} a Y_{ij} jsou závislé náhodné veličiny.

Obrázek 17 zobrazuje vývoj odhadu parametru p, tj. mixovacího parametru použitého v rovnici (6.9). Odhad tohoto parametru se mění během jednotlivých kol, ale ve všech případech je mezi 0,069 až 0,095. V tomto případě, vysoká hodnota parametru p značí vhodnost diagonálního rozšíření.



Obrázek 17: Vývoj mixovacího parametru *p*

Poslední skupina parametrů je θ_k, pro k = 0,1,…,5, tyto parametry jsou použité pouze v diagonálně rozšířených modelech. Nejvyšší hodnotu pro 16. až 91. kolo má odhad parametru θ_2, to značí, že parametr se použil ke zvýšení hodnoty pravděpodobnosti remízy se dvěma góly na každé straně. Téměř ve všech zbývajících kolech (92. až 177.) má nejvyšší hodnotu parametr θ_1. Odhady parametrů $\theta_{0,\dots,4}$ jsou ukázány na obrázku 19. Při porovnání s dvourozměrným Poissonovo modelem (BP), odhady p a θ_k znamenají, že největší změny jsou v pravděpodobnostech remízových utkání s žádným, jedním, dvěma a třemi góly na každé straně. Menší změny jsou vytvořeny pro případ remízy se čtyřmi góly, naopak žádné změny nejsou vytvořeny pro remízy s pěti góly, proto tento parametr není zahrnut v následujícím grafu.



Obrázek 18: Vývoj parametrů $\theta_{0,\dots,4}$

Odhad výsledků zápasů je vytvořen na listu *BP-DI sezóna 2015-2016*.

6.9 Návrh „vlastního“ modelu pro českou ligu

Na základě dat z Extraligy bylo zjištěno, že všechny týmy by neměly mít stejnou hodnotu parametru γ vyjadřujícího výhodu domácího prostředí, protože se u některých mužstev může stát, že jim domácí prostředí nesvědčí a naopak hrají venku lépe než doma. Z tohoto důvodu jsme navrhli úpravu modelů z kapitoly 6 a odhadujeme parametr γ pro každý tým i.

Předchozí tvrzení je podložené následující tabulkou.

Tým	Góly doma	Góly venku	Poměr gólů doma a venku
Hradec Králové	76	57	1,33
Karlovy Vary	66	42	1,57
Kometa Brno	76	74	1,03
Liberec	65	54	1,20
Litvínov	90	71	1,27
Mladá Boleslav	83	73	1,14
Olomouc	74	47	1,57
Pardubice	80	62	1,29
Plzeň	83	53	1,57
Slavia Praha	56	47	1,19
Sparta Praha	89	99	0,90
Třinec	93	86	1,08
Vítkovice	68	64	1,06
Zlín	78	62	1,26

Tabulka 19: Počet gólů domácích/hostů pro jednotlivé týmy v sezóně 2014/2015

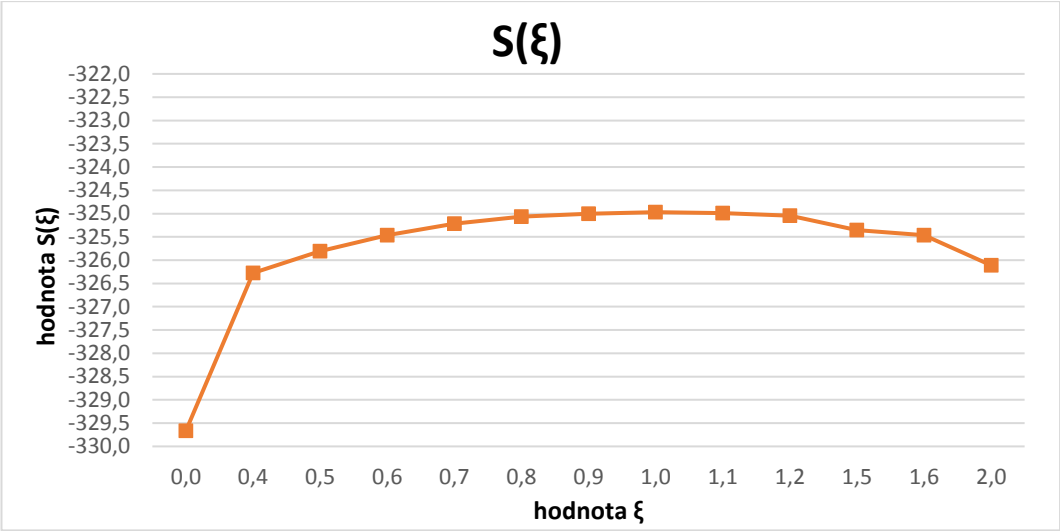
Parametr γ vychází u týmu Sparta Praha menší než 1, proto lze říci, že dává více gólů jako hostující tým, a proto pro tento tým není domácí prostředí výhodou.

6.9.1 Parametr ξ pro sezónu 2014/2015

Od sezóny 2012/2013 až do 2015/2016 nehrál Extraligu tým České Budějovice, proto byl z odhadovaných týmů vyřazen. Veškeré odhady jsou provedené pouze pro 16 týmů.

Stejně jako pro ostatní ligy, bylo potřeba nejprve vhodně zvolit parametr ξ na základě odhadů ze sezóny 2014/2015, detail lze nalézt v souboru *CZE_VM_odhad ksi.xlsm*.

Na obrázku 20 jsou hodnoty $S(\xi)$ pro různé hodnoty ξ. Porovnáme-li všechny modely, tak DP model poskytoval maximální hodnotu $S(\xi)$. Optimální volba parametru ξ pro tento model je rovna 1. V modelu BP-DI a DP-DI je optimální hodnota ξ také 1, pouze u modelu BP tato hodnota vychází 1,3. Pro vybraný model je váha výsledků zápasů starých jeden rok 36,81 %, kdežto stejných výsledků před dvěma lety pouze 13,55 %.



Obrázek 19: Hodnoty $S(\xi)$ proti ξ pro DP model

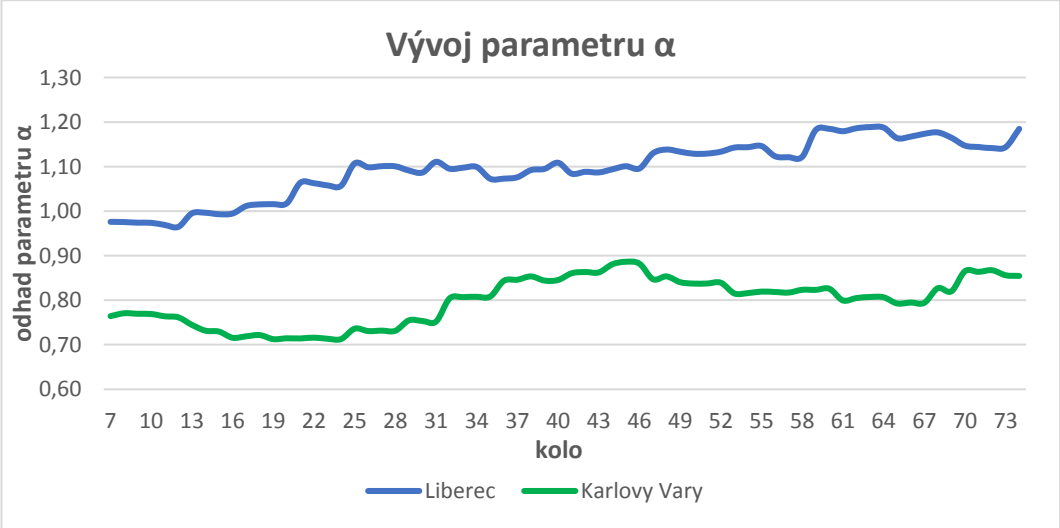
V následující tabulce je přehled hodnot $S(\xi)$ jednotlivých modelů pro různé hodnoty parametru ξ. Mezi jednotlivými modely jsou malé rozdíly v hodnotách statistiky $S(\xi)$ v optimálním bodě ξ, a jsou označené žlutě. Na základě maximalizace $S(\xi)$ byl vybrán model DP, který vyšel s hodnotou -324,970, tj. vyšší oproti ostatním modelům, v tabulce je zvýrazněn zeleně.

ksí	0,0	0,8	1,0	1,1	1,2	1,3	1,4	1,5	2,0
S(ksí) - DP	-329,663	-325,066	-324,970	-324,993	-325,046	-325,130	-325,216	-325,352	-326,112
S(ksí) - BP	-336,676	-330,060	-329,807	-329,741	-329,711	-329,682	-329,685	-329,694	-330,903
S(ksí) - DP-DI	-331,788	-326,171	-325,634	-326,076	-326,136	-326,188	-326,225	-326,507	-327,173
S(ksí) - BP-DI	-331,154	-325,828	-325,442	-325,614	-325,692	-325,763	-325,776	-326,059	-327,459

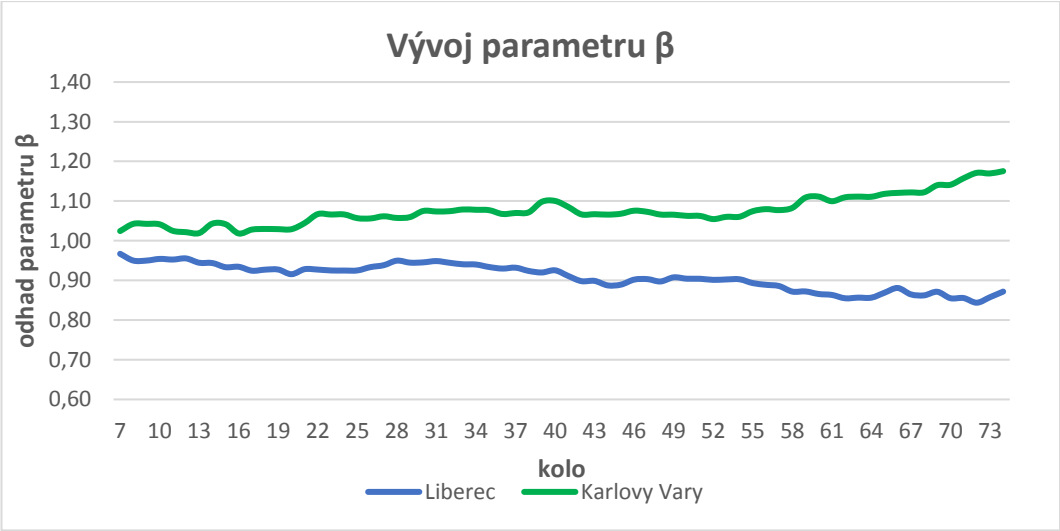
Tabulka 20: Hodnoty $S(\xi)$ pro jednotlivé modely

6.9.2 Odhad parametrů DP model

K maximalizaci se používá opět řešitel, ve kterém je navíc nastavena změna buněk *D4:D19*, tj. odhad parametru γ_i. Na následujících obrázcích je vývoj parametru útoku α a obrany β během sezóny 2015/2016 pro nejlepší a nejhorší tým. Máme k dispozici časovou řadu 68 odhadů pro každý parametr.



Obrázek 20: Vývoj parametru útoku α – vlastní model

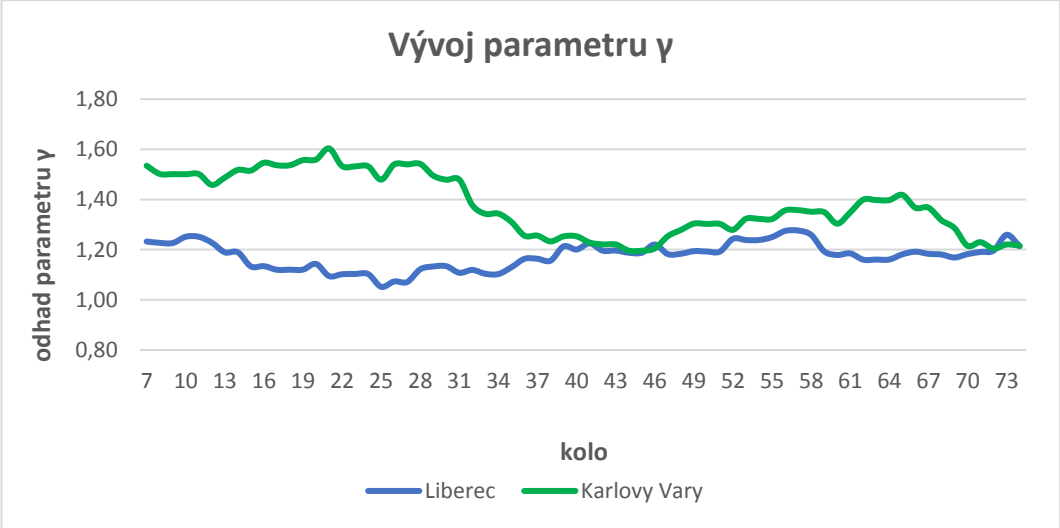


Obrázek 21: Vývoj parametru obrany β – vlastní model

Z grafů jsou patrné rozdíly mezi oběma týmy, jak v útoku, tak i obraně. Oba parametry vycházejí podprůměrně pro tým Karlovy Vary, což odpovídá skutečnosti, že tento tým byl nejhorším týmem v sezóně.

Vývoj těchto parametrů je odlišný od původního modelu pro českou ligu.

V tomto upraveném modelu nepovažujeme parametr γ za globální, ale je odhadován pro každý tým zvlášť, jak již bylo uvedeno výše. Změny tohoto parametru v průběhu sezóny jsou na Obrázek 22 pro tým Liberec a Karlovy Vary.



Obrázek 22: Vývoj parametru γ – vlastní model

Další možnou úpravou modelu by mohlo být vytvoření nové váhové funkce nebo pouze její úprava, například bychom nebrali v úvahu váhu zápasu menší než 0,05, protože takto malá hodnota model téměř neovlivní.

7 Předvídací schopnost a sázecí strategie

V této kapitole je čerpáno z [3].

Aby výsledky demonstrovaly předvídací schopnost vybraného modelu, užijeme je proti sázkovým kancelářím. Existuje mnoho strategií jakým způsobem sázet. Zde bude použita strategie Flat betting. V této strategii se vkládá na každou sázku stejný vklad, v našem případě se jedná o 10 Kč. Výhoda tohoto modelu spočívá v tom, že sázející nemusí řešit kolik má na každý zápas vsadit, protože sází pořád stejnou částku.

Pro každý zápas lze říci, že pravděpodobnosti jsou odhadnuty užitím modelu, a podle očekávané hodnoty zisku (jednoduše spočítaného jako součin odhadů pravděpodobností a daných kurzů) zvolíme, zda-li máme sázet či nikoliv[7]. Přirozeným požadavkem je, aby se sázelo jen v případech, kdy očekávaná hodnota zisku převyšuje hodnotu 1. V našich simulacích zkoušíme také hodnoty větší než 1, to znamená, sázení pouze v těch případech, v nichž při zápase *m* očekávaná hodnota zisku je minimálně rovna *L*, *L > 1*. Pravidlo je možné napsat takto:

$$p_m^R \cdot o_m^R \geq L, R \in \{H, D, A\}, \tag{7.1}$$

kde p_m^R je odhad pravděpodobnosti předpokládaný modelem pro jednotlivé varianty *R* (*R* nabývá hodnoty *H* pro vítězství domácího týmu, *D* pro remízu a *A* pro vítězství hostujícího týmu) a o_m^R jsou kurzy pro danou variantu. Protože jsou tři možné výsledky utkání *{H, D, A}*, lze obdržet dva případy, kde je splněna podmínka daná rovnicí (7.1). V těchto případech se provede sázení pouze pro variantu s nejvyšším očekávaným ziskem.

Parametr *L* lze zvolit např. 1,0; 1,1; 1,2; atd. Parametr *L* menší než 1 nemá smysl volit, protože tento parametr udává minimální střední hodnotu výhry při sázce 1, pokud by pravděpodobnosti p_m^R odhadnuté modelem byly shodné se skutečnými neznámými pravděpodobnostmi p_i. Teoreticky by bylo nejlepší volit co největší *L* (1,6 a vyšší), ale takových zápasů je za sezónu jen velmi málo.

Pro testování se použily kurzy z [A], které jsou získané jako průměr z pěti vybraných sázkových kanceláří, více v kapitole 2.

7.1 Sázení pro Extraligu (CZE)

V kapitole 6.6 byl na základě sezóny 2014/2015 vybrán BP model pro výpočet pravděpodobností výher domácího týmu, remízy a vítězství hostujícího mužstva.

Při sázení na 91 % zápasů by na všechny zápasy bylo potřeba 2 940 Kč abychom v nejhorším případě skončili po sezóně s 0 Kč. Zvolili jsme 91 %, protože když by se uvažovalo sázení dle modelu pro parametr *L* = 1, tak bychom vsadili na 294 zápasů z 324 možných.

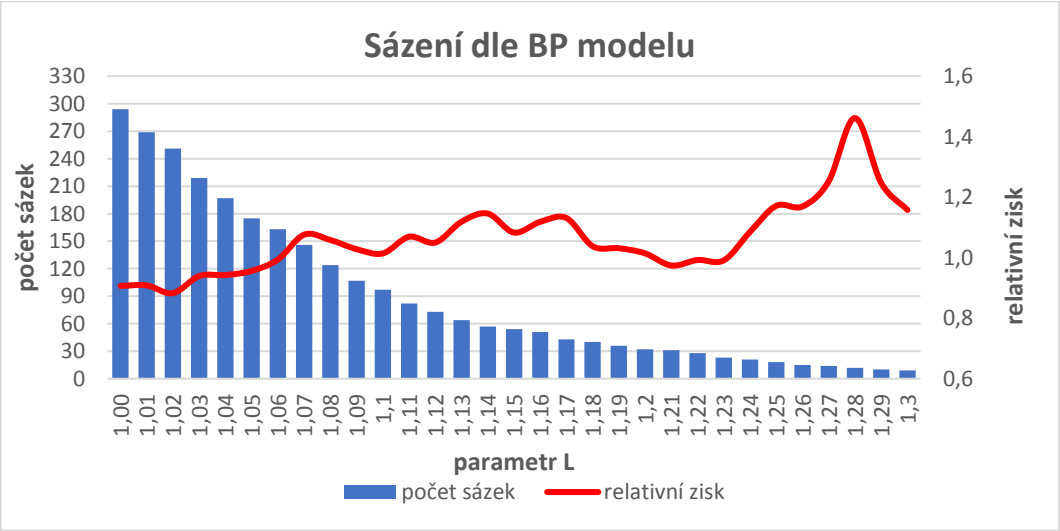
Při vyšších hodnotách parametru *L* bychom sázeli na méně procent zápasů, což lze nalézt v Tabulka 21.

[7] Tato podmínka je převzata z článku [3].

	L												
	1,00	1,03	1,06	1,07	1,08	1,09	1,12	1,15	1,18	1,21	1,24	1,27	1,3
počet sázek	294	219	163	146	124	107	73	54	40	31	21	14	9
celkem vsazeno	2 940	2 190	1 630	1 460	1 240	1 070	730	540	400	310	210	140	90
počet vítězných sázek	121	91	69	66	56	47	31	24	16	11	8	6	3
celkem čistý zisk	-274,20	-132,80	-10,10	111,20	72,60	30,20	36,60	44,90	15,20	-8,00	18,00	35,40	14,20
relativní zisk	0,91	0,94	0,99	1,08	1,06	1,03	1,05	1,08	1,04	0,97	1,09	1,25	1,16
na kolik % zápasů by se vsadilo	91%	68%	50%	45%	38%	33%	23%	17%	12%	10%	6%	4%	3%

Tabulka 21: Souhrnná tabulka pro sázení dle BP modelu pro různé hodnoty *L* – Extraliga (CZE)

Relativní zisk pro průměrné sázkové kurzy je znázorněn na Obrázek 23. Rostoucí parametr *L* snižuje počet zápasů, kde proběhlo sázení, a zde jsou znázorněny pouze výsledky, které byly sázeny alespoň třikrát za sezónu. Pro průměrné kurzy je model ziskový, již pro hodnoty parametru *L* 1,07 a vyšší, to znamená, že pro *L = 1,07* by bylo 66 vítězných sázek ze 146 vsazených (ze 324 odhadovaných zápasů). Počet vsazených a vítězných zápasů pro různé hodnoty parametru *L* jsou uvedeny v předchozí tabulce.



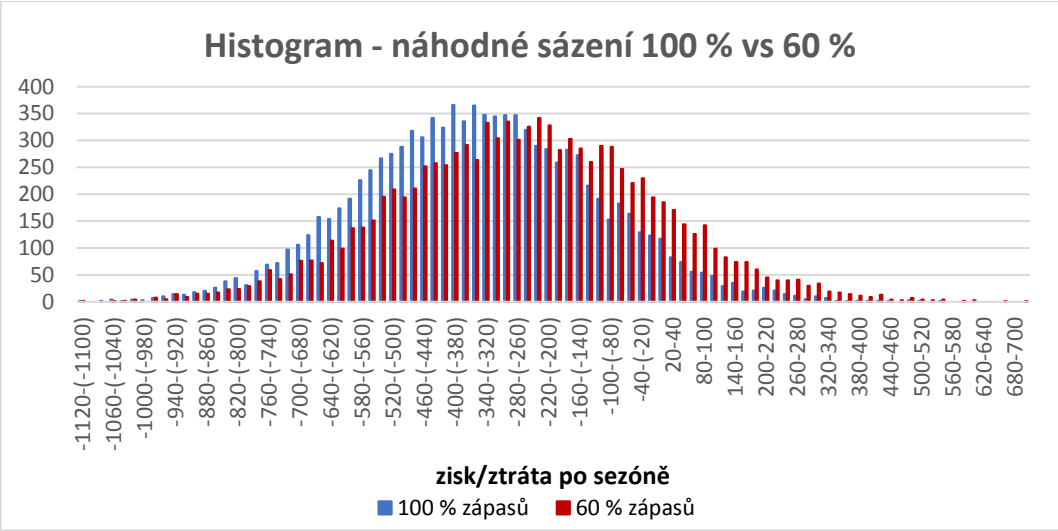
Obrázek 23: Relativní zisk v porovnání s parametrem *L* - Extraliga

7.1.1 Typy sázení

Model jsme porovnali s triviálními strategiemi sázení, abychom ukázali, že je pro naše data vhodnější použít model. Veškeré výpočty jsou provedené v sešitě *CZE_BP 15-16.xlsm* na listech *sázení*.

Předpokládali jsme dostupné finanční prostředky na začátku pouze ve výši 1 000 Kč, aby mohla nastat situace, že sázející zbankrotuje, což se ale nakonec nestalo u žádné varianty sázení v této lize.

V každé lize jsme porovnávali model náhodného sázení na 100 % zápasů a na 60 % zápasů, pro Extraligu je srovnání v následujícím histogramu.



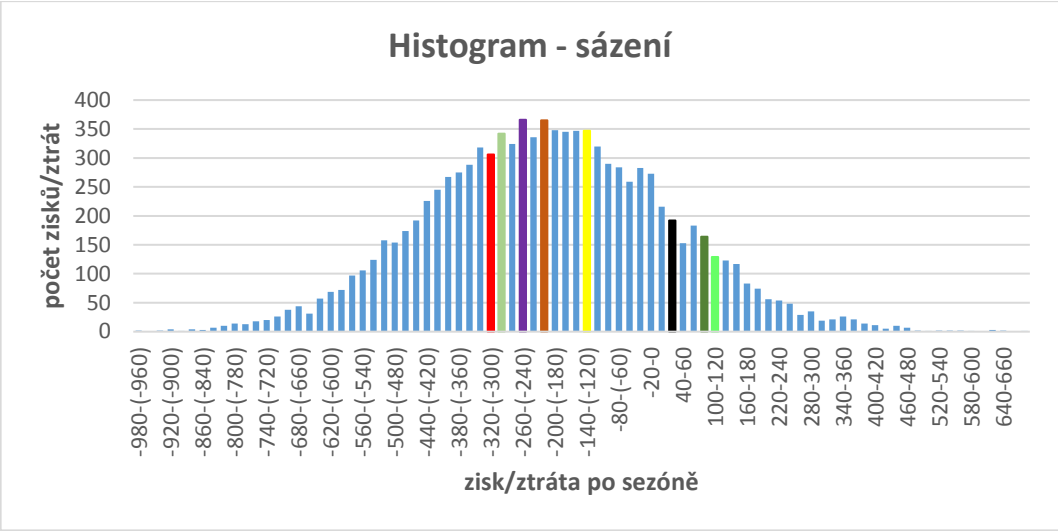
Obrázek 24: Porovnání náhodného sázení na 100 % a 60 % zápasů – Extraliga (CZE)

Náhodné sázení je zpracováno i pro třetí variantu, a to pro sázení na 80 % zápasů, kterou využíváme k porovnání s vybraným modelem a s triviálními schématy sázení (na listu *porovnání sázení*).

Při sestavování náhodného sázení na 80 % (resp. 60 %) zápasů jsme postupovali v následujících krocích:

- vygenerovali jsme si v Microsoft Excel funkcí *randbetween* náhodná čísla od 0 do 2 podle konkrétního výsledku sázení (1 = sázka na výhru domácích, 0 = sázka na remízu, 2 = sázka na výhru hostů),
- abychom nesázeli na všechny zápasy, přidali jsme podmínku, že v případě vygenerování čísla menšího než 0,2 (resp. 0,4), nesázíme na tuto variantu, v opačném případě vsadíme,
- po každém zápasu jsme si uchovávali hodnoty čistého zisku nebo ztráty a připočítaly je k předchozímu stavu, abychom na konci měli celkovou hodnotu čistého zisku/ztráty,
- náhodné generování jsme provedli přes makro ve Visual Basicu, které pouze aktualizuje hodnoty (F9) a uchovává čistý zisk/ztrátu po celé sezóně, a tento postup se opakuje v cyklu 10 000 krát. Makro se spouští přes vytvořené tlačítko „Aktualizace“, ukázka viz *Příloha 4*,
- na závěr z těchto ukládaných čistých zisků/ztrát po sezóně jsme si vytvořili histogram.

Do histogramu náhodného sázení pro 80 % zápasů jsme přidali triviální schémata sázení, a to konkrétně sázení na výhru domácích, remízu, výhru hostů, na minimální a maximální kurzy, a také sázení dle BP modelu pro tři různé hodnoty parametru L ($L = 1{,}02$; $1{,}07$ a $1{,}14$). Tyto možnosti jsou barevně vyznačené v následujícím obrázku.



Obrázek 25: Porovnání různých typů sázení – Extraliga (CZE)

Odstíny zelené barvy v histogramu jsou použité pro variantu sázení dle BP modelu s vybranými třemi hodnotami parametru *L*. Zelený sloupec zprava značí model s parametrem $L = 1{,}07$, kdy při této hodnotě tento model vycházel nejvíce ziskový, sloupec vlevo od této hodnoty je pro model s $L = 1{,}14$ a světle zelený pro $L = 1{,}02$. Černý sloupec je pro variantu sázení na max. kurzy, žlutý pro min. kurzy, hnědý pro remízu, fialový pro výhru domácích a červený pro výhru hostů.

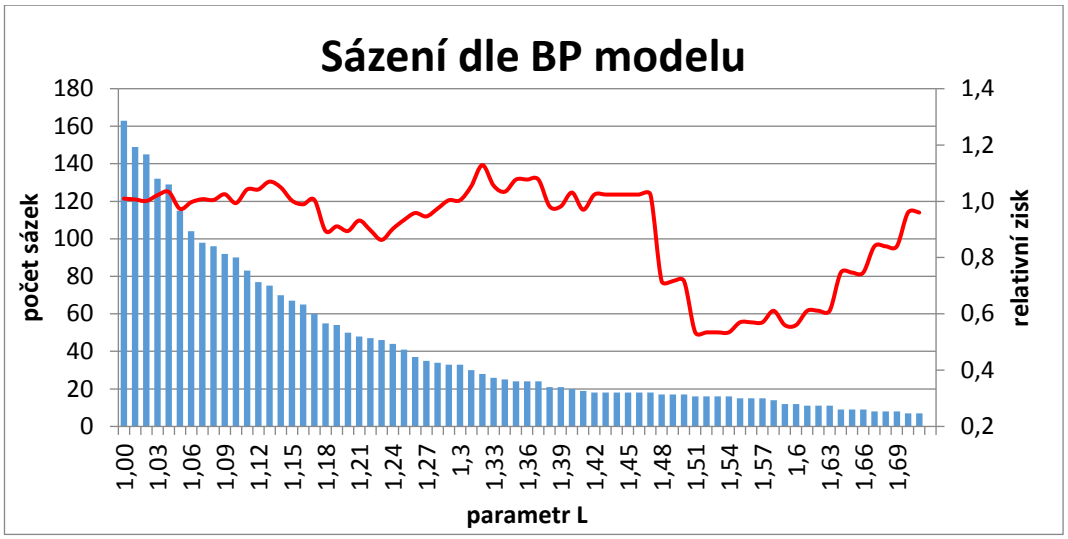
Hodnoty znázorněné v předchozím histogramu v pravé části zobrazují vyšší zisk. Z tohoto srovnání je patrné, že model je nejvhodnější variantou, dokonce ziskovou i při menších hodnotách parametru *L*. Druhou nejvýhodnější variantou se ukázalo sázení na max. kurzy, což ovšem mohlo být způsobeno i tím, že máme k dipozici průměrné kurzy z pěti sázkových kanceláří.

7.2 Sázení pro Ekstraligu (POL)

Pro polskou ligu jsme v kapitole 6.7 zvolili jako nejvhodnější model BP, celé zpracování tohoto modelu včetně sázení je vytvořeno v souboru *POL_BP 15-16.xlsx*.

Relativní zisk pro průměrné sázkové kurzy pro polskou ligu je znázorněn na Obrázek 26. Rostoucí parametr *L* opět snižuje počet zápasů, kde proběhlo sázení, a zde jsou znázorněny pouze výsledky, které byly sázeny alespoň pětkrát za sezónu. Pro naše data je model ziskový, již pro hodnotu parametru $L = 1{,}0$. Pro danou hodnotu by v tomto případě bylo 82 vítězných sázek z celkově vsazených 163 zápasů. Počet vítězných zápasů a relativní zisky jsou uvedeny pro různé hodnoty parametru *L* jsou uvedeny v následující tabulce.

Tento model pro polskou ligu vychází zajímavě, z důvodu malých ztrát i při menších hodnotách parametru *L*. Při vyšších hodnotách parametru *L* ale vychází s malým ziskem.



Obrázek 26: Relativní zisk v porovnání s parametrem *L* – Ekstraliga (POL)

	L													
	1,00	**1,03**	**1,06**	**1,07**	**1,08**	**1,09**	**1,12**	**1,15**	**1,18**	**1,31**	**1,32**	**1,33**	**1,37**	**1,38**
počet sázek	163	132	104	98	96	92	77	67	55	30	28	26	24	21
celkem vsazeno	1 630	1 320	1 040	980	960	920	770	670	550	300	280	260	240	210
počet vítězných sázek	82	60	45	41	39	38	31	26	20	10	10	8	7	5
celkem čistý zisk	5,20	28,70	-3,60	7,30	4,40	23,00	32,40	1,60	-58,10	16,00	36,00	14,50	18,50	-3,80
relativní zisk	1,01	1,02	1,00	1,01	1,00	1,03	1,04	1,00	0,89	1,05	1,13	1,06	1,08	0,98
na kolik % zápasů by se vsadilo	77%	62%	49%	46%	45%	43%	36%	32%	26%	14%	13%	12%	11%	10%

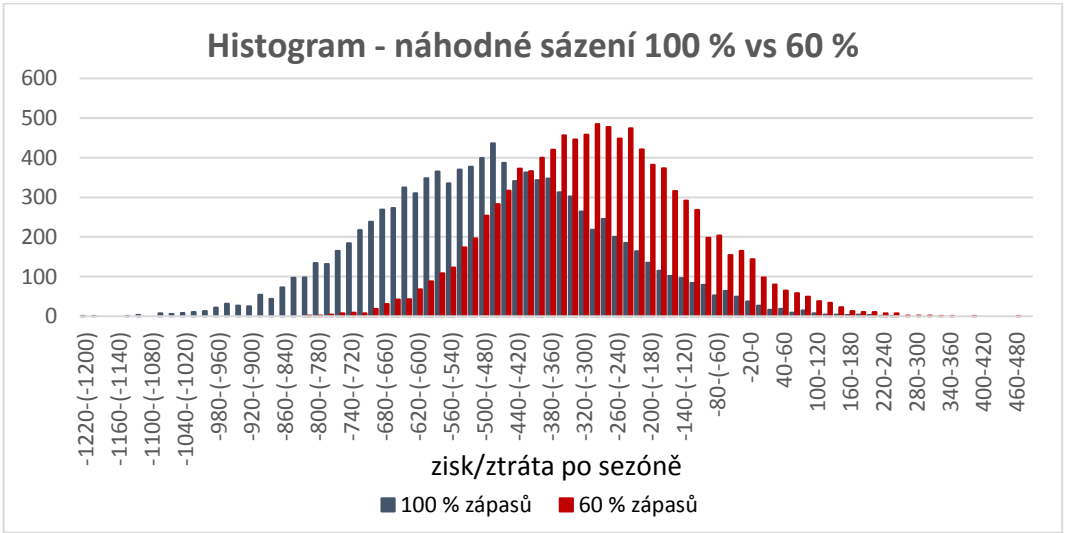
Tabulka 22: Souhrn pro sázení dle BP modelu – Ekstraliga (POL)

Počet vsazených a vyhraných zápasů při různých hodnotách parametru *L* jsou uvedené v předchozí tabulce, kde zeleně vyznačené buňky označují zisk po konci celé sezóny při konkrétní hodnotě parametru *L*.

V předchozí tabulce je uvedena hodnota relativního zisku, která vyjadřuje kolikanásobek vsazených prostředků byl po konci sezóny.

7.2.1 Typy sázení

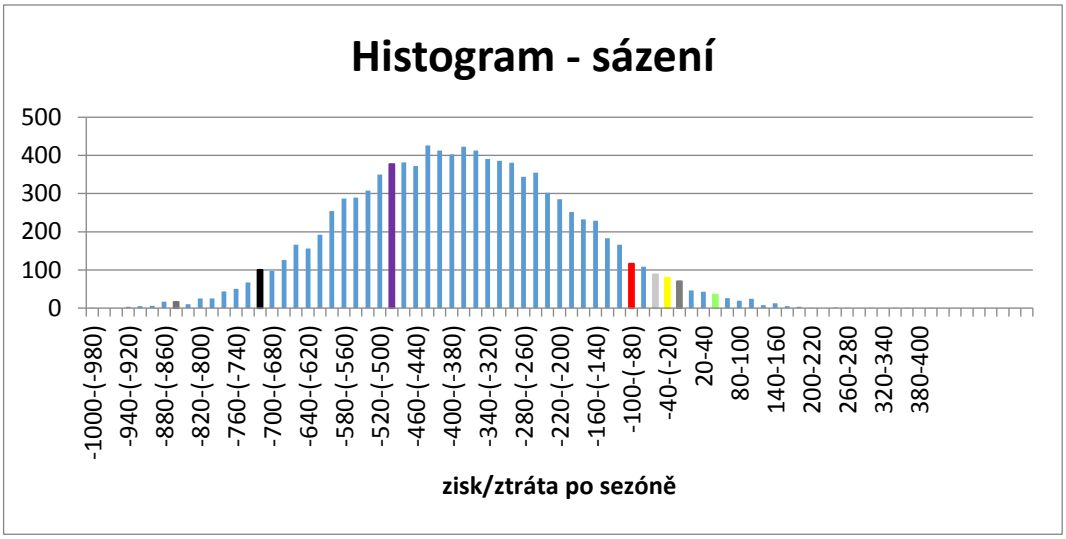
Náhodné sázení na všechny zápasy jsme porovnali s náhodným sázením na 60 % zápasů, které je zobrazené na histogramu.



Obrázek 27: Porovnání náhodného sázení na 100 % a 60 % zápasů – Ekstraliga (POL)

Pro polskou ligu jsme stejným způsobem porovnali zvolený BP model s triviálními schématy sázení, jak je vidět na následujícím histogramu. Nejvhodnější a zároveň nejvíce ziskovou variantou je pro naše data BP model s hodnotou parametru *L = 1,04*.

Vysvětlivky k použitým barvám v následujícím grafu jsou uvedené v kapitole 7.1.1.



Obrázek 28: Porovnání různých typů sázení – Ekstraliga (POL)

7.3 Sázení pro NHL

V kapitole 6.8 jsme zvolili jako nejvhodnější model BP-DI pro NHL ligu a výsledky sázení jsou zpracovány v souboru *NHL_BP-DI 15-16.xlsx* na listu *sázení dle modelu*.

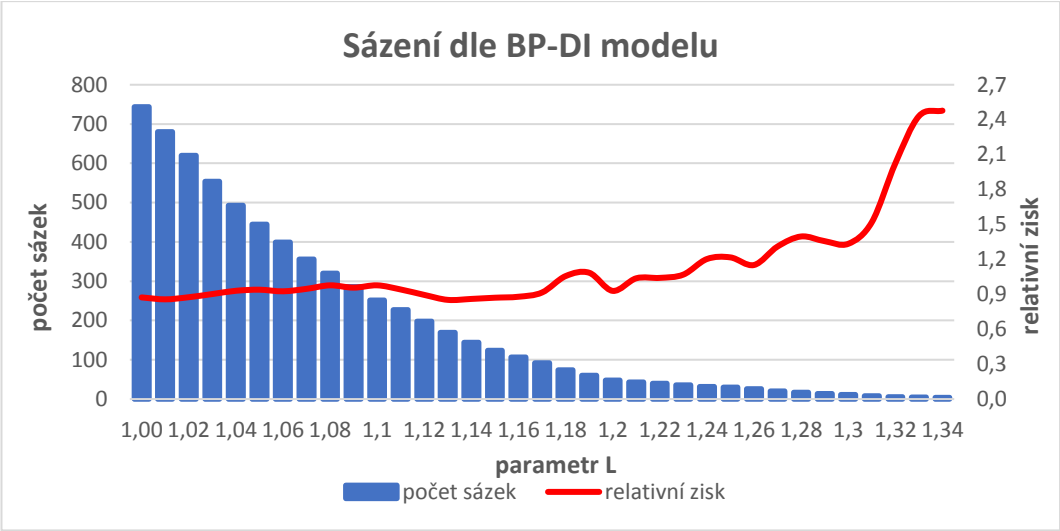
Abychom měli prostředky k sázení na 80 % zápasů při velikosti sázky 10 Kč, celkově bychom tedy potřebovali částku ve výši 9 000 Kč. Po započtení zisku a ztrát v průběhu sezóny se dostaneme na potřebnou počáteční částku pouze 2 000 Kč (zaokrouhleně na tisíce). Při této počáteční částce se po sezóně nedostáváme do bankrotu.

Pro průměrné kurzy je model opět ziskový, ale tentokrát pouze pro vyšší hodnoty parametru *L* (1,18 a vyšší), což by při této hodnotě znamenalo vsadit pouze na 74 zápasů z 1 131 možných (z toho by bylo vítězných zápasů pouze 25).

	L											
	1,00	1,03	1,06	1,09	1,12	1,15	1,18	1,21	1,24	1,27	1,3	1,33
počet sázek	744	554	399	286	198	124	74	43	32	20	11	5
celkem vsazeno	7 440	5 540	3 990	2 860	1 980	1 240	740	430	320	200	110	50
počet vítězných sázek	240	179	128	94	60	36	25	14	12	8	5	4
celkem čistý zisk	-943,80	-542,20	-296,30	-123,10	-208,20	-160,30	41,60	16,10	64,90	62,10	36,70	71,50
relativní zisk	0,87	0,90	0,93	0,96	0,89	0,87	1,06	1,04	1,20	1,31	1,33	2,43
na kolik % zápasů by se vsadilo	65,8%	49,0%	35,3%	25,3%	17,5%	11,0%	6,5%	3,8%	2,8%	1,8%	1,0%	0,4%

Tabulka 23: Souhrn pro sázení dle BP-DI modelu - NHL

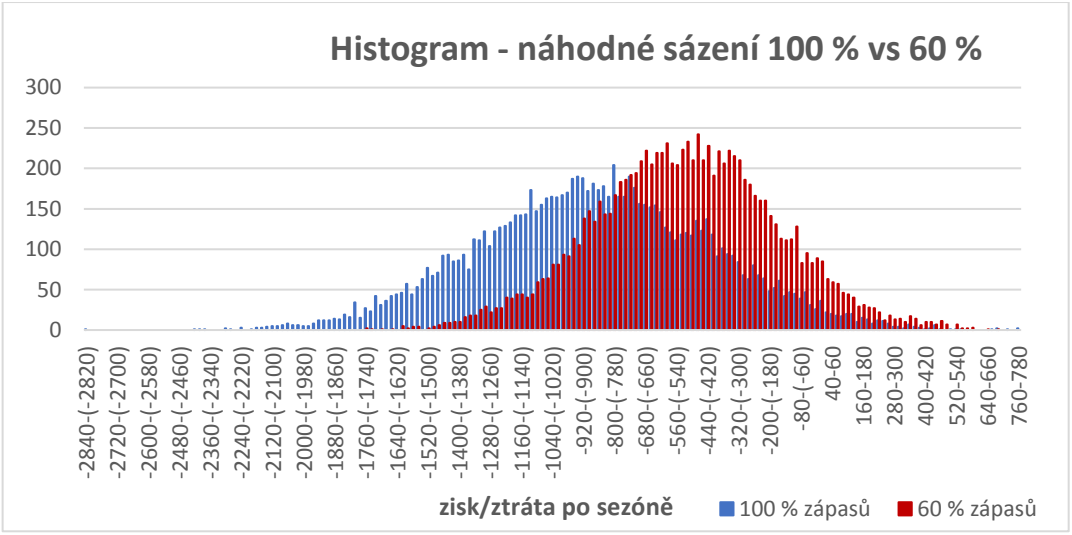
V následujícím obrázku je ukázán graf počtu vsazených zápasů při různých hodnotách *L* pro NHL ligu.



Obrázek 29: Relativní zisk v porovnání s parametrem *L* – NHL

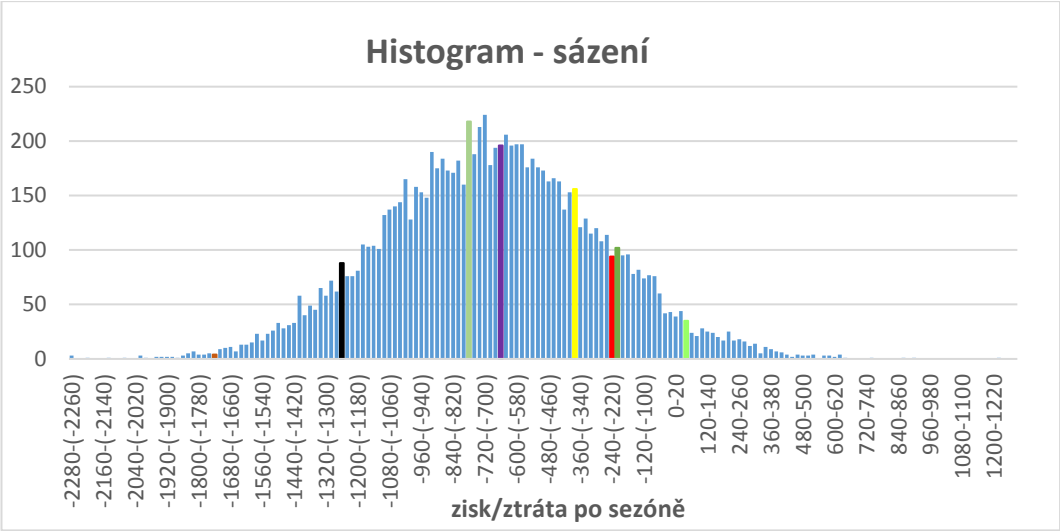
7.3.1 Typy sázení

Na Obrázek 30 jsou zobrazeny četnosti celkových zisků/ztrát po celé sezóně při sázení na 100 % a 60 %zápasů v NHL lize. Jsou zde opět vidět rozdíly, varianta sázení na 60 % zápasů by mohla přinést vyšší zisk oproti sázení na 100 % zápasů, což by mohlo odpovídat skutečnosti při náhodném sázení.



Obrázek 30: Porovnání náhodného sázení na 100 % a 60 % zápasů – NHL

Stejně jako u předchozích lig, i zde je srovnání BP-DI modelu s ostatními variantami sázení na základě následujícího histogramu. Jako nejméně výhodná se ukázala varianta sázení na remízu, kde bychom měli po celé sezóně ztrátu -1 734 Kč.

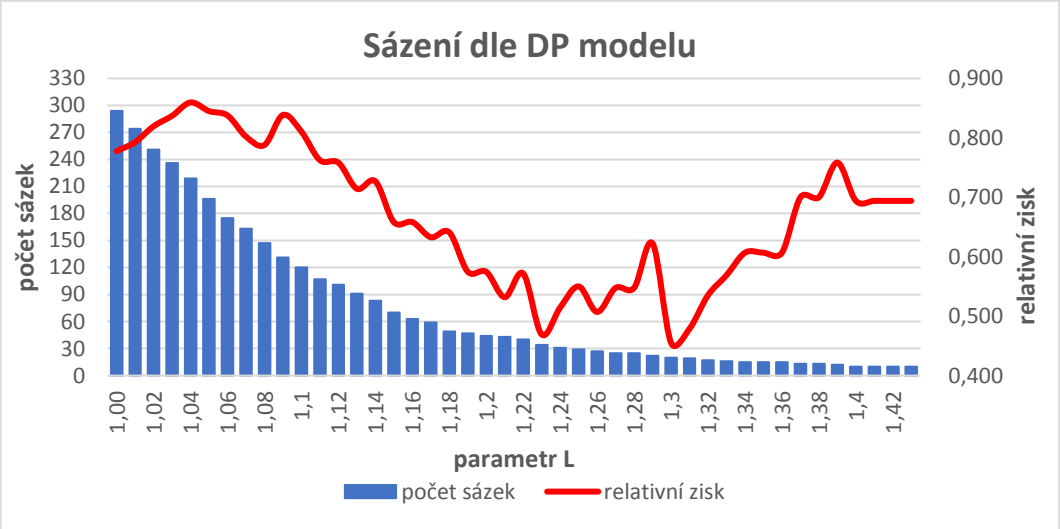


Obrázek 31: Porovnání různých typů sázení – NHL

Vysvětlivky k použitým barvám v grafu jsou uvedené v kapitole 7.1.1.

7.4 Vlastní model

Stejným způsobem je zpracováno i sázení dle upraveného vlastního modelu, jehož výsledky jsou v souboru *CZE_VM_DP 15-16.xlsm*.



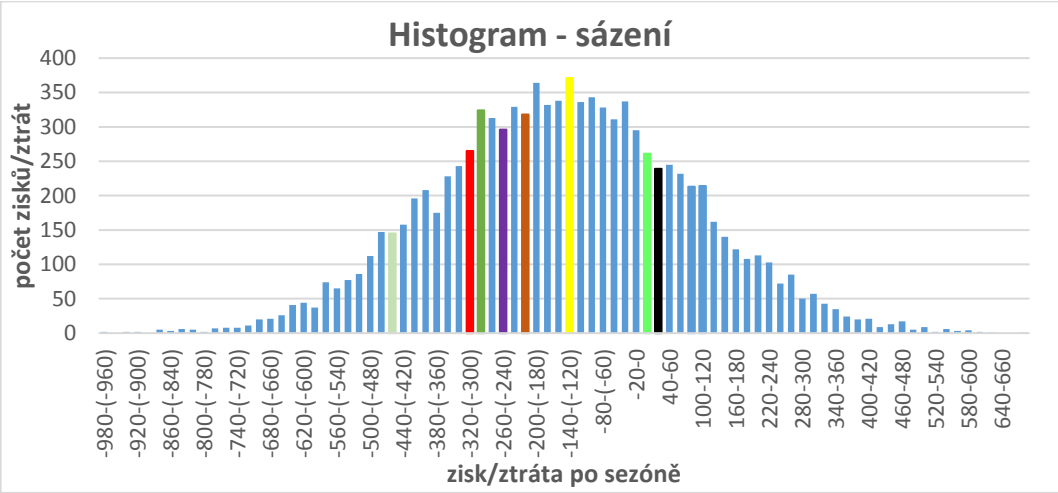
Obrázek 32: Relativní zisk v porovnání s parametrem *L* – vlastní model

	L												
	1,00	**1,03**	**1,06**	**1,07**	**1,08**	**1,09**	**1,12**	**1,15**	**1,18**	**1,21**	**1,24**	**1,27**	**1,3**
počet sázek	294	236	175	163	147	131	101	70	49	43	31	25	20
celkem vsazeno	2 940	2 360	1 750	1 630	1 470	1 310	1 010	700	490	430	310	250	200
počet vítězných sázek	103	87	63	57	50	47	32	19	13	9	6	5	3
celkem čistý zisk	-653,50	-384,60	-284,20	-322,60	-311,80	-211,60	-244,20	-239,30	-175,90	-201,30	-150,50	-113,10	-109,00
relativní zisk	0,78	0,84	0,84	0,80	0,79	0,84	0,76	0,66	0,64	0,53	0,51	0,55	0,46
na kolik % zápasů by se vsadilo	91%	73%	54%	50%	45%	40%	31%	22%	15%	13%	10%	8%	6%

Tabulka 24: Souhrn pro sázení dle DP modelu – vlastní model

7.4.1 Typy sázení

Vlastní model vychází v porovnání s původním modelem pro českou ligu horší co se týče sázení. Navržený model by byl sice výhodnější než ostatní triviální strategie sázení, ale až při vyšších hodnotách parametru L. Oproti původnímu modelu použitém pro data z české ligy z kapitoly 7.1 je tento navržený model téměř pro všechny hodnoty L neziskový. Pro tento model vycházela nejvhodnější varianta sázení na maximální kurzy.



Obrázek 33: Porovnání různých typů sázení – vlastní model

8 Závěr

Na začátku této práce byly popsány modely používané na fotbalová data, které používal M. J. Maher ve svém článku [1], poté byl zpracován vybraný model 2 pro hokejová data. Otestovali jsme předpoklady pro Maherovy modely, a to shodu s Poissonovo rozdělením a nezávislost dat. Při testování zda se data řídí Poissonovo rozdělením jsme u všech třech vybraných lig (česká, polská a národní hokejová) došli k závěru, že hypotézu H_0 nezamítáme, tj. můžeme předpokládat, že se data řídí tímto rozdělením. V případě testování nezávislosti náhodných veličin X_{ij} (počet gólů domácích) a Y_{ij} (počet gólů hostů) nebyl výsledek ve všech sezónách až tak jednoznačný. Pro některé sezóny vycházelo, že se jedná o nezávislé náhodné veličiny, u jiných naopak o závislé náhodné veličiny. Z tohoto důvodu jsme zjistili, že model 2 dle Mahera není příliš vhodný pro naše hokejová data. Díky tomuto zjištění jsme tento model pro ostatní ligy již nepoužili.

V další části jsme se zaměřili na modely definované ve článku [3], které jsou již přizpůsobené hokejovým datům. Tyto modely zahrnují informace z historických výsledků zápasů a za použití váhové funkce snižují informace obsažené ve výsledcích každého utkání, tj. zohledňují při odhadech parametrů jak týmy hrají v průběhu sezóny. Pro remízy se užívají upravené modely definované ve článku [3], tzv. diagonálně rozšířené Poissonovy modely.

Všechny modely pro českou ligu vycházely téměř shodně z pohledu funkce $S(\xi)$, přičemž BP model vycházel nejlépe. Tento model jsme použili pro odhad parametrů v sezóně 2015/2016 a poté k porovnání s jednoduššími strategiemi sázení (sázení výhru domácích, remízu, hostů, min. kurz a max. kurz). Model se ukázal jako nejziskovější. Při aplikování tohoto modelu pro sezónu 2015/2016 bychom došli po celé sezóně k zisku 111 Kč při sázení částky 10 Kč na 146 zápasů, z nichž 66 bylo pro nás vítězných.

Pro polskou ligu byly modely také téměř shodné, ale nejvhodněji vycházel BP model, který se následně použil pro odhad sezóny 2015/2016 a poté opět k porovnání s jednoduchými strategiemi sázení. Model byl opět ziskový. Když bychom vsadili 10 Kč na 77 zápasů, z nichž pro nás vítězných by bylo 31 a získali bychom tak zisk po celé sezóně ve výši 32 Kč.

Pro NHL ligu byl zvolen model BP-DI, pro který na rozdíl od ostatních lig vycházela hodnota $S(\xi)$ oproti ostatním modelům významně vyšší. Tento model vycházel více ziskový oproti sázení na polskou ligu.

V práci byl také použit vlastní model na základě úpravy DP modelu pro českou ligu, který se však ukázal jako méně vhodný oproti zvolenému modelu použitého pro data z české ligy.

Seznam obrázků

Seznam tabulek

Literatura a zdroje

[1] MAHER, M. J. (1982). Modelling association fotball scores. *Statistica Neerlandica*, 36, pp. 109 - 118

[2] DIXON, M. J., & COLES, S. G. (1997). Modelling Association Fotball Scores and Inefficiencies in the Fotball Betting Market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 46(2), pp. 265 - 280.

[3] MAREK, P., ŠEDIVÁ, B., ŤOUPAL, T. (2014). Modeling and prediction of ice hockey match results. *Journal of Quantitative Analysis in Sports*, 10(3), pp. 357 – 365.

[4] KARLIS, D., NTZOUFRAS, I. (2003). Analysis of sports data by using bivariate Poisson models. The Statistician, 52(3), pp. 381 - 393

[5] FAMOYE, F. (2010). A new bivariate generalized Poisson distribution. *Statistica Neerlandica*, 64(1), pp. 112 - 124

[6] HÁTLE, J., & LIKEŠ, J. (1974). Základy počtu pravděpodobnosti a matematické statistiky. Praha: SNTL

[7] REIF, J. *Metody matematické statistiky*. Plzeň: Západočeská univerzita, 2004, s. 61 – 63. ISBN 80-7043-302-7

[8] Pravděpodobnost a statistika hypertextově. P-hodnota. [online]. 2014 [cit. 2016-05-20]. Dostupné z: http://home.zcu.cz/~friesl/hpsb/phodn.html

[9] ŠEDIVÁ, B. Přednášky MSM. [online]. 2014 [cit. 2016-05-20]. Dostupné z: http://home.zcu.cz/~sediva/msm/slideMSM-07.pdf

[10] MAREK, P. Přednášky ZTI. [online]. 2014 [cit. 2016-05-20]. Dostupné z: http://home.zcu.cz/~patrke/WWW-KMA/ZTI/

[11] Microsoft. *GRG Algorithm.* [online]. © 2016 [cit. 2016-05-20]. Dostupné z: https://support.microsoft.com/en-us/kb/82890

[12] J. Špaček. Modelování a odhadování výsledků sportovních utkání. Plzeň, 2015. Bakalářská práce. ZČU Plzeň

Zdroje dat

[A] Sfstats. [online]. © 2006 - 2012 [cit. 2016-20-05]. Dostupné z: http://www.sfstats.net/

[B] Odds portal. [online]. © 2008 - 2016 [cit. 2016-20-05]. Dostupné z: http://www.oddsportal.com/

[C] BetExplorer. [online] © 2003 - 2016 [cit. 2016-20-05]. Dostupné z: http://www.betexplorer.com/hockey/

Přílohy

Přílohy obsažené na CD:

DP_Gabrišková.pdf - soubor s kompletním textem diplomové práce v elektronické podobě,

data.xlsx – soubor obsahující data pro všechny tři ligy,

chybějící kurzy.xlsx – soubor s chybějícími kurzy a zápasy, které bylo nutné dohledávat,

CZE – složka obsahující veškeré soubory s testy, odhady a sázení pro českou ligu

NHL – složka obsahující veškeré soubory s testy, odhady a sázení pro polskou ligu

POL - složka obsahující veškeré soubory s testy, odhady a sázení pro NHL ligu

vlastní model - složka obsahující veškeré soubory s testy, odhady a sázení pro upravený „vlastní“ model

Přílohy tištěné

Příloha 1: kód vytvořeného makra pro odhad parametrů Maherova modelu v sezóně 2014/2015

```
Sub Odhad_6()

' Hledání optimálního odhadu parametrů alpha a beta pomocí řešitele (5 iterací)

' Spuštění řešitele pro sezónu 2014/2015
For i = 1 To 5
 SolverOk SetCell:="$P$18", MaxMinVal:=3, ValueOf:=1077, ByChange:="$L$4:$M$17" _
        , Engine:=1, EngineDesc:="GRG Nonlinear"

 SolverAdd CellRef:=Range("N18"), Relation:=2, FormulaText:="O18"

 SolverSolve UserFinish:=True

If i = 5 Then
    Range("N4:O17").Select
    Selection.Copy
    Range("Q4").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False

Else
    Range("N4:O17").Select
    Selection.Copy
    Range("L4").Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False

End If

Next

End Sub
```

Příloha 2: ukázka kódu vytvořeného makra pro odhad parametrů BP modelu pro českou ligu v sezóně 2015/2016

```
Sub Odhad_BP_model()

Application.ScreenUpdating = False
Application.Calculation = xlAutomatic

Sheets("BP model").Select
Start = ActiveSheet.Range("B26").Value
Max = Start + ActiveSheet.Range("B27").Value
If Max > ActiveSheet.Range("C27").Value Then
Max = ActiveSheet.Range("C27").Value
End If

For i = Start To Max
Application.Calculation = xlAutomatic
Sheets("BP model").Select
ActiveSheet.Range("B28").Offset(i, 0).Select
Application.CutCopyMode = False
Selection.Copy
ActiveSheet.Range("B25").Select
Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False


' start řešitel maximalizace

SolverReset

  SolverOk SetCell:="$B$24", MaxMinVal:=1, ValueOf:=0, ByChange:= _
        "$B$4:$C$20,$F$4:$H$4"
  SolverOptions MaxTime:=99999, Iterations:=32000, Scaling:=True, Precision:=0.00001
  SolverAdd CellRef:=Range("B21"), Relation:=2, FormulaText:="17"
  SolverAdd CellRef:=Range("C21"), Relation:=2, FormulaText:="17"
  SolverAdd CellRef:=Range("B4:C20"), Relation:=3, FormulaText:="0,1"
  SolverAdd CellRef:=Range("H4"), Relation:=3, FormulaText:="1"

  SolverSolve UserFinish:=True


' uložení spočtených parametrů

    Sheets("BP model").Select
    Range("B4:B20").Select
    Application.CutCopyMode = False
    Selection.Copy
    Sheets("BP model_parametry").Select
    Range("XFD5").Select
    Selection.End(xlToLeft).Offset(0, 1).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False

    Sheets("BP model").Select
    Range("C4:C20").Select
    Application.CutCopyMode = False
    Selection.Copy
    Sheets("BP model_parametry").Select
    Range("XFD26").Select
```

Příloha 3: ukázka kódu vytvořeného makra pro odhad parametrů BP-DI modelu pro NHL ligu v sezóně 2015/2016

```
Sub Odhad_BP_DI_model()

Application.ScreenUpdating = False
Application.Calculation = xlAutomatic

Sheets("BP-DI model").Select
Start = ActiveSheet.Range("B39").Value
Max = Start + ActiveSheet.Range("B40").Value
If Max > ActiveSheet.Range("C40").Value Then
Max = ActiveSheet.Range("C40").Value
End If

For i = Start To Max
Application.Calculation = xlAutomatic
Sheets("BP-DI model").Select
ActiveSheet.Range("B41").Offset(i, 0).Select
Application.CutCopyMode = False
Selection.Copy
ActiveSheet.Range("B38").Select
Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False


' start řešitel maximalizace

  SolverReset

  SolverOk SetCell:="$B$37", MaxMinVal:=1, ValueOf:=0, ByChange:= _
        "$B$4:$C$33,$F$4:$N$4"
  SolverOptions MaxTime:=9999, Iterations:=32000, Scaling:=True, Precision:=0.00001
  SolverAdd CellRef:=Range("B34"), Relation:=2, FormulaText:="30"
  SolverAdd CellRef:=Range("C34"), Relation:=2, FormulaText:="30"
  SolverAdd CellRef:=Range("B4:C33"), Relation:=3, FormulaText:="0,1"
  SolverAdd CellRef:=Range("H4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("P4"), Relation:=2, FormulaText:="1,000"
  SolverAdd CellRef:=Range("I4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("J4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("K4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("L4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("M4"), Relation:=3, FormulaText:="0"
  SolverAdd CellRef:=Range("N4"), Relation:=3, FormulaText:="0"

  SolverSolve UserFinish:=True

' uložení spočtených parametrů

    Sheets("BP-DI model").Select
    Range("B4:B33").Select
    Application.CutCopyMode = False
    Selection.Copy
    Sheets("BP-DI model_parametry").Select
    Range("XFD5").Select
    Selection.End(xlToLeft).Offset(0, 1).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
```

Příloha 4: ukázka kódu makra použité při náhodném sázení

```
Sub Aktualizace()

' Aktualizace (F9) a uchování celkového zisku/ztráty po sezóně

Application.ScreenUpdating = False

For i = 1 To 10000

    Calculate
    Range("Q366").Select
    Selection.Copy
    ActiveSheet.Range("U41").Offset(i, 0).Select
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    Application.CutCopyMode = False

Next

End Sub
```