Západočeská univerzita v Plzni

Fakulta aplikovaných věd
Katedra matematiky

Bakalářská práce

Modelování a odhadování výsledků ledního hokeje

Plzeň 2018 Pavlína Hellusová

Místo této strany bude zadání práce.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 22. května 2018

Pavlína Hellusová

Poděkování

Tímto bych chtěla poděkovat vedoucímu bakalářské práce Ing. Patrici Markovi, Ph.D. za jeho odborné rady, vstřícný přístup a čas věnovaný při konzultacích během vytváření této práce. Dále bych ráda poděkovala své rodině a blízkým za bezmeznou podporu během studia.

Abstract

This bachelor thesis focuses on modeling and prediction of ice hockey match results, specificaly on the regular seasons of the highest-level Czech league, the *Extraliga*, between 2011 and 2017. The first section of the thesis comprises a list of basic models used in sport matches modeling and their short descriptions. The central part is dedicated to the double Poisson distribution model, which is also the basis for the presented innovation. The main goal of the changes made is to create a model taking into account the effect of the so called home team advantage on individual teams and to test if the predictive ability of the model increased. In the last section, the quality of predictions of both models is checked by fictive betting against a bookmaker.

Key words: sport results prediction, ice hockey, home team advantage, Poisson distribution

Abstrakt

Tato bakalářská práce se zabývá modelováním a odhadováním výsledků ledního hokeje, konkrétně základní části české Extraligy mezi lety 2011–2017. Součástí práce je seznam základních modelů využívaných v oblasti modelování sportovních utkání a jejich krátký popis. Hlavním zpracovaným modelem je dvojitý Poissonův model, na který navazuje i představená inovace. Hlavním cílem úpravy je vytvořit model zohledňující rozdílný vliv tzv. domácí výhody na jednotlivé týmy a otestovat, zda tato změna vylepší predikční schopnosti modelu. V závěru práce je kvalita predikcí obou modelů ověřena na imaginárním sázení proti sázkové kanceláři.

Klíčová slova: odhad sportovních výsledků, lední hokej, výhoda domácího týmu, Poissonovo rozdělení

Obsah

Seznam tabulek

Seznam obrázků

1 Úvod

Modelování a odhadování výsledků zápasů z různých sportovních odvětví je statisticky velmi zajímavá disciplína, kterou se zabývá velké množství autorů z celého světa. Hlavním cílem této práce je představit upravený model využívající dvojité Poissonovo rozdělení a porovnat jeho predikční schopnosti s původním modelem, představeným v článku autorů Marka, Šedivé a Ťoupala (2014) a to vše na datech ze zápasů v ledním hokeji.

V minulosti již byla problematika modelování a odhadování výsledků sportovních utkání několikrát zpracovávána, ale každým rokem se objevují nové články zabývající se touto tématikou. Některé mají základ v již klasických modelech, některé přichází se zcela novými přístupy. V dalších několika odstavcích budou v krátkosti představeny základní články a přístupy úzce spjaté s obsahem bakalářské práce.

Maher (1982) se ve svém článku zabývá modelováním výsledků fotbalových utkání pro anglickou ligu a tento článek lze považovat za naprosto základní zdroj, což se potvrzuje i tím, že většina dále popisovaných článků z něj vychází. Autor nejprve zkoumá model využívající dvě nezávislá Poissonova rozdělení. V zápase mezi týmy i a j předpokládá, že náhodná proměnná popisující počet gólů vstřelených domácím týmem X_{ij} se řídí Poissonovým rozdělením s parametrem λ_{ij}, pro který platí $\lambda_{ij} = \alpha_i \cdot \beta_j$, kde α_i reprezentuje sílu útoku domácího týmu a β_j slabost obrany týmu j na cizím hřišti. Y_{ij}, tedy počet gólů vstřelených hostujícím týmem, se opět řídí Poissonovým rozdělením, tentokrát s parametrem $\mu_{ij} = \gamma_i \cdot \delta_j$, kde γ_i je slabost obrany týmu i na domácím hřišti a δ_j je síla útoku týmu j venku.

Vzhledem k velkému počtu parametrů, které bylo pro tento model nutné odhadovat, autor článku zkoumal, jestli jsou všechny tyto parametry přínosné. Na základě testování byl vybrán jako nejlepší model, ve kterém má každý tým vlastní parametr pro sílu obrany i útoku a navíc je zde i parametr k odrážející vliv domácího prostředí. Problémem tohoto modelu bylo mírné podceňování vzniku remíz a přeceňování extrémů (skóre jednoho z týmů rovné nule nebo vyšší než čtyři góly).

Maher ve svém článku dále používá model využívající dvourozměrné Poissonovo rozdělení. Tento přístup se autorovi jevil výhodnější a přesnější

než použití dvou Poissonových rozdělení, ale nepovoluje zápornou korelaci. Základní myšlenkou je využití výsledku utkání jako celku. Je tedy stanoven bodový rozdíl ve skóre jako $Z_{ij} = X_{ij} - Y_{ij}$. Problém předcházejícího modelu, tedy podceňování remíz ($Z_{ij} = 0$), mohl být způsoben jistou korelací mezi počty vstřelených branek. Pro dvourozměrný Poissonovský model byly tedy jednotlivé výsledky brány jako realizace náhodné proměnné $X_{ij} = U_{ij} + W_{ij}$ a $Y_{ij} = V_{ij} + W_{ij}$, kde se nezávislé náhodné proměnné U_{ij}, V_{ij} a W_{ij} řídí Poissonovým rozdělením s parametry $(\mu_{ij} - \eta_{ij}), (\lambda_{ij} - \eta_{ij})$ a kde $\eta_{ij} = \rho\sqrt{\mu_{ij}\lambda_{ij}}$ a ρ je korelace mezi počty gólů. Tato korelace se v uvedeném článku pohybuje okolo 0,2. Výsledkem byly mnohem přesnější odhady, a model využívající dvourozměrné Poissonovo rozdělení se tedy dá považovat za rozumnou možnost pro předpověď výsledků fotbalových utkání.

Další velmi podstatný článek napsali Dixon a Coles (1997). I v tomto článku je zpracováno téma odhadování výsledků fotbalových utkání. Popisovaný model má základ v předcházejícím článku (Maher (1982)) a to v modelu využívajícím dvě nezávislá Poissonova rozdělení. Tento model je ještě modifikován novým parametrem, který upravuje závislost pro výsledky 0:0, 1:0, 0:1 a 1:1. Tato úprava vyhovuje fotbalovým výsledkům, ale například pro hokej by nebyla dostačující a vzhledem k vyšším hodnotám skóre v hokejových utkáních by bylo třeba upravit i další možné výsledky. Další změnou je zde proměnnost parametrů v čase. Autoři zde dávají novějším výsledkům větší váhu než výsledkům starším.

Karlis a Ntzoufras (2003) také napsali článek se základem v Maherově modelu. Jako možné vylepšení zde autoři představili inflační faktor pro prvky na diagonále, tedy remízy a také parametr korelace mezi počty gólů. Zvýhodnění prvků na diagonále připomíná model Dixona a Colese (1997), ale vzhledem k upravování celé diagonály se mnohem více hodí na hokejová utkání.

V poslední době je díky rozvoji výpočetní techniky možné využívat výpočetně náročnější postupy. Buttrey (2016) se ve článku zabývá přímo hokejovou tématikou. Je v něm představen model, který je založen opět na původním Maherově modelu, ale pracuje i s daty o vyloučení hráčů a obsahuje další tři parametry (pro regulérní stav hráčů 5:5, nebo přesilovou hru 4:5 respektive 5:4). Další část článku popisuje jiný možný přístup za pomoci simulování zápasů. Tento způsob má lepší výsledky, ale je velmi náročný na detailnost a rozsah dat, díky čemuž se stává pro českou extraligu prakticky nepoužitelný.

V této práci bude především čerpáno z článku od autorů Marka, Šedivé a Ťoupala (2014), na který bakalářská práce navazuje. Autoři se zabývají

odhadováním výsledků ledního hokeje pro českou extraligu. Představují čtyři modely, zčásti známé pro fotbalová utkání, upravené pro rozdílnou strukturu výsledků v ledním hokeji. Do modelů zavádí možnost negativní korelace mezi výsledky domácích a hostů. Další modifikací je skutečnost, že v těchto modelech jsou odhady závislé na čase, proto je zde zavedena funkce přidělující jednotlivým výsledkům rozdílnou váhu. Tato váha je závislá přímo na datu, ve kterém byl příslušný zápas odehrán. Z těchto modifikací bude v bakalářské práci vycházeno před představením vlastních inovací.

2 Data a testování předpokladů

Pro bakalářskou práci byla vybrána data z české extraligy ledního hokeje a to sezóny 2011–2012 až 2016–2017. Prvních pět sezón je využito ke zpřesnění odhadů a poslední sezóna je ponechána pro testování predikčních schopností modelů. Tato data byla získána ze zdrojů BetExplorer.com (2018), Sfstats.net (2018) a ověřena na stránce SPORT.CZ (2018). Kurzy využité v závěrečné části pro testování modelů pomocí imaginárního sázení jsou přejaty ze stránky Sfstats.net (2018).

Z každé sezóny byla použita pouze základní část, kterou hraje vždy 14 týmů. Ještě konkrétněji: byly použity výsledky po uplynutí základní hrací doby a to proto, že různě dlouhá prodloužení či výsledky po samostatných nájezdech by bylo složité zakomponovat do modelu. Systém základní části ukládá každému z týmů sehrát zápas se všemi ostatními dvakrát na domácí půdě a dvakrát v roli hosta. Základní část se tedy dělí na 52 kol. Díky tomuto systému mají data základní části vhodnou strukturu pro statistické zpracovávání. Jsou k dispozici stejná data pro všechny týmy a zároveň jsou vzájemně dobře provázána.

Díky prostupnosti extraligy a první ligy se ve zkoumaných sezónách v extralize vystřídalo 17 týmů.

V sezóně 2011–2012 bylo nutné upravit výsledek kontumovaného zápasu mezi Třincem a Plzní z 0:5 na původních 3:0. Zápas byl regulérně odehrán, ale došlo k následné kontumaci, protože jeden z hráčů vítězného týmu (Třinec) neměl zaregistrovanou hráčskou smlouvu. Vstupní data po zpracování (viz obrázek 2.1) se nachází v souboru *Data a Poisson.xlsx*.

2.1 Testování předpokladů modelů

Zkoumaná data bylo nutné otestovat. Jelikož modely, které budeme používat, předpokládají Poissonovo rozdělení vstupních dat, byla nejprve ověřována tato vlastnost a to dvěma způsoby.

2015–2016							
Datum	Domácí	Skóre		Hosté	1	x	2
4.3.2016	**Hradec Králové**	4	1	Pardubice	**1,66**	4,46	3,93
4.3.2016	**Litvínov**	1	1	Vítkovice	**2,00**	4,05	3,01
4.3.2016	**Sparta Praha**	6	4	Mladá Boleslav	**1,52**	4,91	4,66
4.3.2016	**Plzeň**	3	1	Chomutov	**1,63**	4,38	4,14
4.3.2016	Kometa Brno	2	4	**Zlín**	2,00	4,13	**2,96**
4.3.2016	**Liberec**	6	2	Karlovy Vary	**1,30**	6,16	6,94
4.3.2016	Třinec	0	2	**Olomouc**	1,74	4,26	**3,79**
1.3.2016	Chomutov	1	3	**Sparta Praha**	2,40	4,03	**2,41**
1.3.2016	**Zlín**	6	4	Plzeň	**2,20**	4,08	2,63
1.3.2016	Vítkovice	5	6	**Liberec**	2,23	4,19	**2,58**
1.3.2016	Pardubice	3	3	**Kometa Brno**	2,40	4,03	**2,43**
1.3.2016	**Olomouc**	4	1	Litvínov	**1,84**	4,10	3,41
1.3.2016	**Mladá Boleslav**	0	0	Třinec	**2,04**	4,08	2,90
1.3.2016	Karlovy Vary	2	4	**Hradec Králové**	2,63	4,22	**2,16**

Obrázek 2.1: Ukázka zpracovaných vstupních dat

2.1.1 Chí-kvadrát test dobré shody

Princip této metody spočívá v porovnání naměřených četností výskytu jednotlivých jevů s četnostmi očekávanými od dat řídících se Poissonovým rozdělením pravděpodobnosti. Více o metodě se lze dočíst například v knize autorů Hátleho a Likeše (1974, s. 340).

V našem případě byly testovány dvě složené hypotézy na hladině významnosti $\alpha = 5\,\%$. První z nich je H_0 : Počty gólů vstřelených domácími týmy v dané sezóně se řídí Poissonovým rozdělením pravděpodobnosti proti alternativní hypotéze H_1 : Počty gólů vstřelených domácími se Poissonovým rozdělením pravděpodobnosti neřídí. Druhá má pak tvar H_0 : Počty gólů vstřelených hostujícími týmy v dané sezóně se řídí Poissonovým rozdělením pravděpodobnosti proti alternativní hypotéze H_1 : Počty gólů vstřelených hosty se Poissonovým rozdělením pravděpodobnosti neřídí.

Naměřené četnosti jednotlivých skóre domácích a hostujících týmů byly porovnávány s četnostmi předpokládanými. Po dopočtení hodnot očekávaných četností a jejich úpravě byly vypočteny jednotlivé testové statistiky. Úprava očekávaných četností byla prováděna dle pravidel uvedených v knize Reif (2004), tedy 80 % hodnot musí být větší než 5 a všechny větší než 1. Testové

statistiky byly následně porovnány s kvantilem χ^2 rozdělení upraveným Bonferroniho korekcí. Tato korekce se používá při zkoumání složené hypotézy. Kdybychom u každého z týmů použili 95% kvantil, zvýšila by se pravděpodobnost, že zamítneme hypotézu H_0 i přesto, že H_0 platí, tedy pravděpodobnost chyby prvního druhu. Z tohoto důvodu je při výpočtu upravena hladina významnosti α na α/n, kde n je počet týmů v sezóně. Více o metodě je uvedeno například ve článku Abdi (2007).

Poznámka. *Myšlenku Bonferroniho korekce lze jednoduše demonstrovat pomocí následujícího příkladu: Mějme složenou hypotézu H_0 skládající se ze dvou podhypotéz, které platí. U každé z nich je pak šance* 95 %*, že nebudou zamítnuty. U složené hypotézy tedy platí, že šance nezamítnutní je celkem* $0,95 \cdot 0,95 = 0,9025$.

Problémem úpravy je snižování síly testu s rostoucím počtem podhypotéz. V případě, že by se hypotéza H_0 skládala z dvaceti podhypotéz, byla by upravená hladina významnosti už jen $\alpha = 0,05/20 = 0,0025$, rostla by hodnota chyby druhého druhu β a tím pádem i klesala síla testu.

Další korekce, které by bylo možné použít, jako například Šidákovu či Tukeyho metodu, je možné najít například v článku Abdi (2007).

V každé podhypotéze se tedy porovnává příslušné testové kritérium s kvantilem χ^2 rozdělení, tj.

$$\sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} > \chi^2_{1-\alpha/n}(v), \tag{2.1}$$

kde je

n	počet týmů,
k	počet hodnot po sloučení,
n_i	naměřená četnost,
o_i	očekávaná četnost,
$\chi^2_{1-\alpha/n}$	kvantil χ^2 rozdělení,
v	počet stupňů volnosti

Pokud není ani jedna podhypotéza zamítnuta, dojde k nezamítnutí složené hypotézy H_0. V opačném případě je přijata alternativní hypotéza H_1.

Tímto způsobem byla otestována data ze všech zpracovávaných sezón a pro každou z nich byl výsledek testu nezamítnutí hypotézy H_0. Na základě chíkvadrát testu dobré shody se tedy nedá vyloučit, že počty vstřelených gólů

domácími a hosty se v každé ze sezón řídí Poissonovým rozdělením pravděpodobnosti.

Tabulka s jednotlivými p-hodnotami testu pro všechny sezóny (viz obr. 2.2) je v souboru *Data a Poisson.xlsx* na listu *Souhrn*. P-hodnotou testu se rozumí nejmenší hladina významnosti, na níž ještě zamítáme nulovou hypotézu.

P-hodnoty												
Tým	2016–2017		2015–2016		2014–2015		2013–2014		2012–2013		2011–2012	
	doma	venku	doma	venku	doma	venku	doma	venku	doma	venku	doma	venku
České Budějovice									0,658	0,865	0,717	0,808
Hradec Králové	0,278	0,968	0,568	0,468	0,730	0,263	0,756	0,321				
Chomutov	0,751	0,236	0,577	0,311			0,750	0,697	0,344	0,606		
Karlovy Vary	0,541	0,198	0,190	0,362	0,007	0,254	0,974	0,512	0,056	0,369	0,611	0,498
Kladno							0,363	0,537	0,881	0,433	0,154	0,596
Kometa Brno	0,250	0,192	0,297	0,818	0,636	0,711	0,307	0,222	0,135	0,098	0,048	0,530
Liberec	0,287	0,701	0,104	0,548	0,227	0,386	0,132	0,456	0,677	0,586	0,102	0,630
Litvínov	0,023	0,371	0,213	0,157	0,521	0,425	0,047	0,109	0,189	0,219	0,731	0,265
Mladá Boleslav	0,264	0,106	0,370	0,073	0,047	0,455					0,471	0,360
Olomouc	0,418	0,896	0,621	0,656	0,142	0,425						
Pardubice	0,275	0,469	0,458	0,517	0,125	0,205	0,358	0,094	0,500	0,385	0,619	0,893
Plzeň	0,112	0,369	0,767	0,545	0,898	0,188	0,287	0,517	0,700	0,291	0,125	0,095
Slavia Praha					0,495	0,660	0,582	0,171	0,240	0,413	0,448	0,715
Sparta Praha	0,070	0,927	0,230	0,289	0,415	0,126	0,500	0,029	0,242	0,451	0,344	0,172
Třinec	0,299	0,524	0,436	0,369	0,085	0,653	0,100	0,818	0,417	0,822	0,068	0,677
Vítkovice	0,308	0,750	0,021	0,973	0,111	0,081	0,014	0,742	0,304	0,991	0,301	0,186
Zlín	0,172	0,171	0,736	0,368	0,290	0,009	0,258	0,188	0,148	0,492	0,429	0,227

Obrázek 2.2: p-hodnoty chí-kvadrát testu v jednotlivých sezónách

Testové tabulky jsou také k nahlédnutí v přiloženém souboru *Data a Poisson.xlsx* na listech s označením sezóny a dodatkem „*test*“.

2.1.2 Cramér von Mises

Vzhledem k tomu, že chí-kvadrát test dobré shody je velmi obecný, byl proveden i test, který se přímo specializuje na Poissonovo rozdělení. Jak uvádí autoři článku Spinelli a Stephens (1997), alespoň jeden z testových parametrů uvedených v článku má vždy mnohem větší sílu než Pearsonova statistika, tedy chí-kvadrát test dobré shody.

Stanovme nejdřív n jako počet pozorování, n_i jako naměřené četnosti, o_i jako očekávané četnosti a p_i jako pravděpodobnost danou Poissonovým rozdělením s parametrem λ, který byl odhadnut pomocí výběrového průměru. Dále mějme $Z_j = \sum_{i=0}^{j}(n_i - o_i)$ a $H_j = \sum_{i=0}^{j} p_i$. V článku jsou navrhnuty tři

testové statistiky, které je možné k ověření hypotézy použít.

$$W^2 = n^{-1} \sum_{j=0}^{\infty} Z_j^2 p_j, \tag{2.2}$$

$$A^2 = n^{-1} \sum_{j=0}^{\infty} \frac{Z_j^2 p_j}{H_j(1 - H_j)}, \tag{2.3}$$

$$W_m^2 = n^{-1} \sum_{j=0}^{\infty} Z_j^2. \tag{2.4}$$

Výsledné hodnoty statistik jsou porovnávány s kritickými hodnotami danými tabulkou z uvedeného článku. Tabulka bohužel neobsahuje hodnoty přesně pro λ, které je třeba testovat, ale tyto hodnoty je možné získat aproximací. V práci byly tedy aproximovány nejbližší body, pro které jsou hodnoty dostupné z tabulky. Tato aproximace je ukázána v souboru *Cramer Von Mises.xlsx* na listu *Kritické hodnoty*. V okolí bodu λ, tedy přibližně u hodnoty 3, má funkce prokládající kritické hodnoty konvexní tvar. Při následné aproximaci přímkou tedy dojde k mírnému zvýšení kritické hodnoty a tím pádem i k zmírnění testu, nicméně hodnoty statistik jsou natolik odlišné od aproximovaných kritických hodnot, že tento fakt nemá na výsledek testu vliv. Příklad aproximovaných hodnot a jednotlivých testových statistik je uveden v tabulce 2.1. Tabulka s původními kritickými hodnotami přejatá z článku je obsažena v příloze A.1.

Tabulka 2.1: Testové statistiky a aproximované kritické hodnoty - část sezóny 2015–2016

	Testové statistiky			**Aprox. krit. hodn.**		
Tým	W^2	**A**	W_m^2	W^2	**A**	W_m^2
Hradec Králové	0,040	0,348	0,332	0,178	1,137	1,059
Chomutov	0,023	0,184	0,157	0,179	1,140	1,022
Karlovy Vary	0,022	0,148	0,114	0,181	1,148	0,918

Výsledek testu potvrzuje předchozí tvrzení, tedy nezamítá hypotézy H_0 a to pro všechny sezóny.

2.1.3 Pearsonův chí-kvadrát test nezávislosti

Dalším předpokladem nejjednodušších modelů, který bylo nutné ověřit, byla nezávislost veličin X_{ij} a Y_{ij}. Tedy, že počty vstřelených gólů domácích (X_{ij})

a hostů (Y_{ij}) jsou nezávislé náhodné veličiny. K tomuto účelu byl vybrán chí-kvadrát test dobré shody. Pro každou sezónu byla testována hypotéza H_0: Náhodné veličiny X_{ij} a Y_{ij} jsou nezávislé, oproti hypotéze H_1: Náhodné veličiny X_{ij} a Y_{ij} jsou závislé.

Po vypočtení reálných a očekávaných četností a po sloučení kategorií s ohledem na dříve zmíněné pravidlo (Reif (2004)) byla dopočtena hodnota testové statistiky. Ta byla následně porovnána s odpovídajícím kvantilem chí-kvadrát rozdělení.

Výsledkem testu bylo zamítnutí hypotézy H_0 a to ve všech sezónách kromě sezóny 2014–2015 viz tab. 2.2.

Tabulka 2.2: Výsledky testování nezávislosti v jednotlivých sezónách

Sezóna	**Hodnota statistiky**	**Kritická hodnota**	**p-hodnota**
2011–2012	53,810	37,652	0,001
2012–2013	44,087	37,652	0,011
2013–2014	39,920	37,652	0,030
2014–2015	20,221	37,652	0,735
2015–2016	43,002	37,652	0,014

Je tedy možné předpokládat, že počty gólů vstřelených jednotlivými týmy v utkání jsou navzájem závislé.

Další možností by bylo testovaní složené hypotézy H_0: Počty gólů vstřelené domácími týmy a počty gólů vstřelené hostujícími týmy jsou nezávislé, oproti hypotéze H_1: Počty gólů vstřelené domácími týmy a počty gólů vstřelené hostujícími týmy jsou závislé.

V tomto případě by bylo nutné opět použít Bonferroniho korekci a výsledkem testu by bylo zamítnutí hypotézy H_0, vzhledem k zamítnutní jedné z pohypotéz (sezóna 2011–2012).

3 Použité modely a odhad parametrů

V následující sekci je převážně čerpáno z článku autorů Marka, Šedivé a Ťoupala (2014). Nejprve bude podrobněji popsán model využívající dvojité Poissonovo rozdělení a následně bude představena i na něj navazující úprava.

3.1 Dvojitý Poissonův model

Ačkoliv v článku byl jako nejlepší vyhodnocen dvojitý Poissonův model s úpravou prvků na diagonále, pro práci byl vybrán model bez úpravy, kvůli jeho relativní jednoduchosti a téměř stejným predikčním schopnostem.

Předpokladem pro tento model je, že se počty vstřelených gólů domácími a hostujícími týmy řídí Poissonovým rozdělením a jsou navzájem nezávislé. Pak

$$X_{ij} \sim \mathrm{Po}(\lambda_H = \mu\alpha_i\beta_j\gamma), \tag{3.1}$$

$$Y_{ij} \sim \mathrm{Po}(\lambda_A = \mu\alpha_j\beta_i), \tag{3.2}$$

kde

α_i	je parametr popisující útok (čím vyšší hodnota, tím lepší),
β_i	je parametr popisující obranu (čím menší hodnota, tím lepší),
μ	je parametr měřítka,
γ	zachycuje takzvaný efekt domácího týmu.

Parametry μ a γ jsou stejné pro všechny týmy, zatímco individuální parametry α_i a β_i splňují $\sum_i \alpha_i = N$ a $\sum_i \beta_i = N$, kde N je počet týmů.

3.2 Upravený dvojitý Poissonův model

Hlavní myšlenkou návrhu je variování parametru γ, který určuje vliv domácího prostředí. Tento vliv v utkáních hraje podstatnou roli, nicméně je otázkou, zda je vhodnější tento parametr považovat za globální, nebo jestli je domácím prostředím každý z týmů ovlivněn odlišnou měrou.

Problematikou rozdílnosti vlivu domácího prostředí pro individuální týmy se zabývají i autoři Marek a Vávra (2017). Ti v článku zkoumali data z fotbalových utkání a mimo jiné ukázali, že vliv domácího prostředí není stejný pro všechny týmy, ale je specifickou vlastností každého týmu.

Na základě informací z tohoto článku byl navržen následující model, který globální parametr γ upravuje na individuální a zároveň je vhodný pro hokejová data.

$$X_{ij} \sim \mathrm{Po}(\lambda_H = \mu\alpha_i\gamma_i\gamma\beta_j), \tag{3.3}$$

$$Y_{ij} \sim \mathrm{Po}(\lambda_A = \mu\alpha_j\beta_i\delta_i\delta) \tag{3.4}$$

V modelu se objevují nové parametry γ_i a δ_i. Parametr γ_i slouží k individualizaci globálního vlivu domácího prostředí a tím pádem dochází k úpravě síly útoku domácího týmu (parametru α_i). Parametr δ_i pak upravuje slabost obrany domácího týmu, tedy β_i. Globální parametry γ a δ ve spojení s podmínkou $\sum \alpha_i = \sum \gamma_i = \sum \beta_i = \sum \delta_i = N$, kde je N počet týmů, zajišťují relativnost individuálních parametrů. Podmínky na součet zároveň zaručují jednoznačnou identifikaci parametrů.

Tento přístup využívá k odhadování čtyři parametry, stejně jako jeden z modelů zkoumaný Maherem (1982). Nicméně propojení parametrů je v tomto případě jiné. Model poskytuje možnost odlišit hru na domácím hřišti, ale zároveň zachovává jednoznačnou identifikaci týmů pomocí parametrů α_i a β_i, které jsou při hře na domácí půdě upravovány (u Mahera docházelo k rozdělení týmů na dva bez zachování jakékoliv souvislosti).

Vzhledem k výskytu dvou konstantních parametrů v každé rovnici a z toho plynoucích problémů s jednoznačností modelu byly tyto dvojice parametrů sloučeny do jednoho.

Poznámka. *Mějme například dvojici parametrů* $\mu = 2$ *a* $\gamma = 1,5$. *Tato dvojice v modelu dává stejnou hodnotu, jako kdyby parametry byly například* $\mu = 1$ *a* $\gamma = 3$.

Model má po spojení tvar:

$$X_{ij} \sim \mathrm{Po}(\lambda_H = \gamma\alpha_i\gamma_i\beta_j), \tag{3.5}$$

$$Y_{ij} \sim \mathrm{Po}(\lambda_A = \delta\alpha_j\beta_i\delta_i). \tag{3.6}$$

3.3 Odhad parametrů

U obou představených modelů bylo po zpracování vstupních dat nutné odhadnout příslušné parametry.

Jako první byl zpracován dvojitý Poissonův model. Nechť má sdružená pravděpodobnostní funkce výsledku zápasu mezi domácím týmem i a hostujícím týmem j tvar:

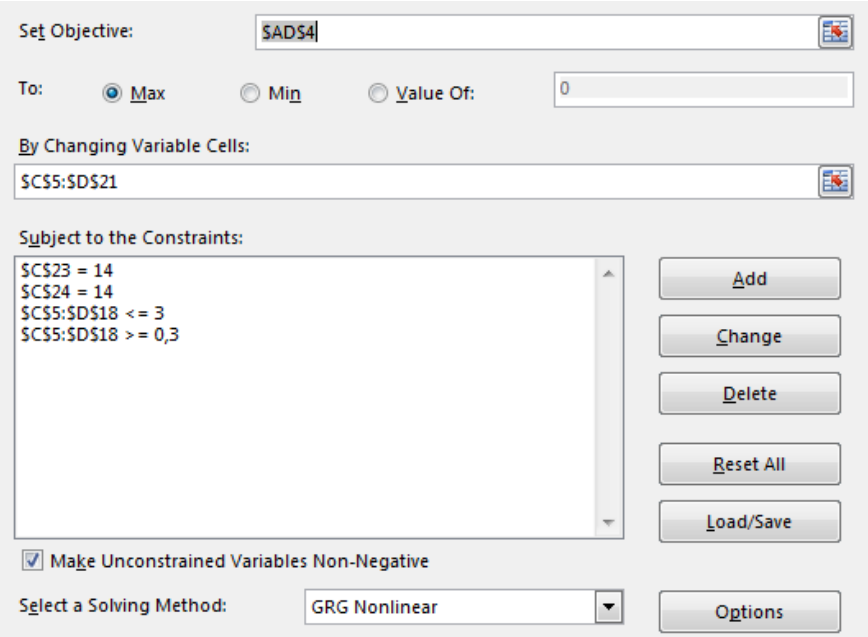
$$P(X_{i,j}, Y_{i,j} = y) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \frac{\lambda_A^y e^{-\lambda_A}}{y!}. \tag{3.7}$$

Pak je pomocí věrohodnostní funkce ve tvaru

$$V(\alpha_i, \beta_i, \gamma, \mu, i = 1, ..., N) = \prod_{m=1}^{M} P(x_m, y_m) \tag{3.8}$$

možné provést odhad používaných parametrů. V rovnici je použit index pořadí zápasů m, kde $m = 1$ je označen nejstarší zápas a $m = M$ zápas nejnovější. Při odhadování parametrů v jednotlivých sezónách bylo postupováno následovně:

Po vystavění vazeb modelu v programu Microsoft Excel, sešit *Dvojitý Poissonův model.xlsm*, proběhlo nastavení doplňku „Řešitel“ viz obr. 3.1.

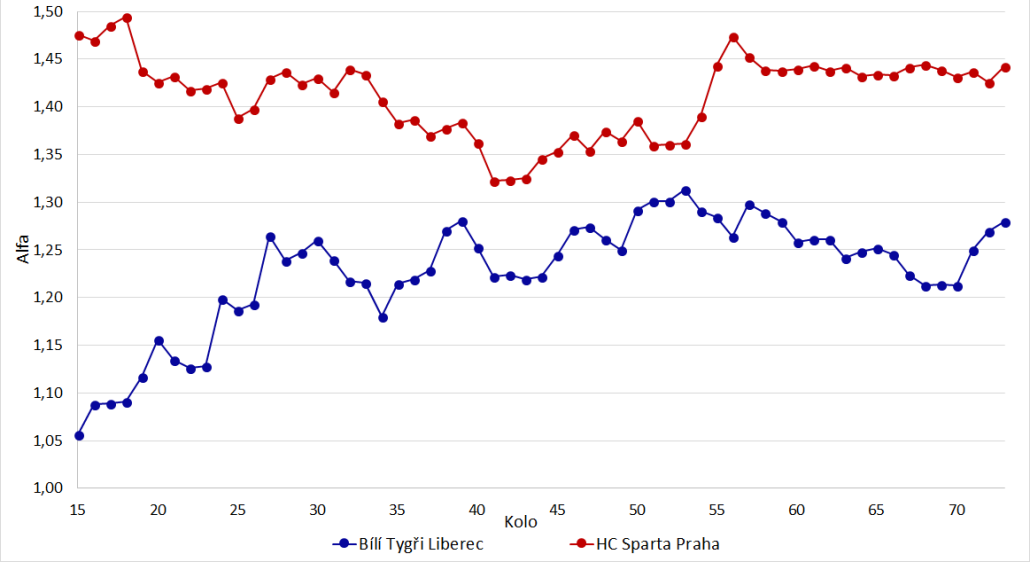


Obrázek 3.1: Ukázka nastavení doplňku řešitel

Byla tedy maximalizována hodnota logaritmické věrohodnostní funkce (buňka $AD4$) pomocí změn parametrů α, β, γ a δ. Pro tyto parametry byla zadána podmínka $0,3 < parametr < 3$ a to především z důvodu, že hodnoty mimo udaný interval mohou způsobit selhání výpočtu. Dané meze jsou zároveň dostatečně vzdálené, aby bylo prakticky nemožné je při odhadu překročit. Další podmínkou bylo dodržení $\sum_i \alpha_i = 14$ a $\sum_i \beta_i = 14$.

Vzhledem k tomu, že k dalšímu postupu bylo nutné odhadovat parametry vždy po odehrání nového kola (v práci byla hodnota kola zvýšena vždy s novým datem zápasu), bylo pro zrychlení těchto odhadů vytvořeno makro (kód uveden v příloze A.2). Hlavní funkcí je spuštění řešitele a přenesení odhadnutých hodnot parametrů z každého kola na listy „*výsledky*“ vždy s doplněním o určení roku sezóny, tedy například *výsledky 2011–2012*. Na listu „*ksi*“ (opět s doplněním o určení sezóny) je pak rovnou tvořena tabulka s hodnotami parametrů z předcházejícího kola, tedy například pro zápasy ve dvacátém kole jsou uvedeny hodnoty parametrů odhadnuté v kole devatenáctém.

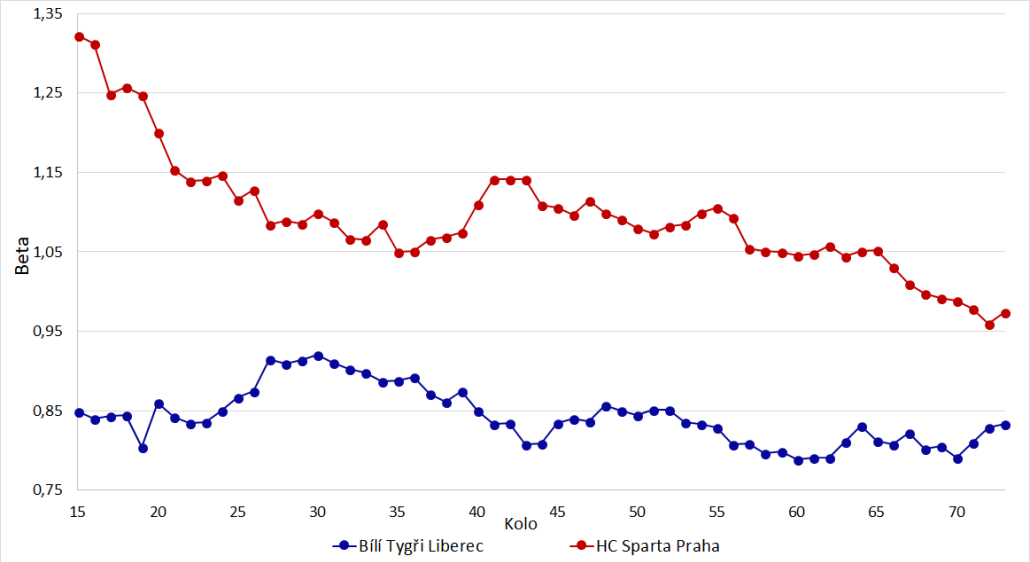
Na obrázcích 3.2, 3.3 a 3.4 je zanesen vývoj odhadnutých hodnot parametrů s narůstajícím počtem odehraných kol. Jedná se o odhady ze sezóny 2015–2016, tedy nejnovější, kterou v práci považujeme za „známou“.



Obrázek 3.2: Vývoj parametru α u týmů Bílí Tygři Liberec a HC Sparta Praha

Na prvním obrázku je pro porovnání uveden vývoj parametru α pro dva nejúspěšnější týmy základní části sezóny 2015–2016 (Bílí Tygři Liberec a HC Sparta Praha). Je zřejmé, že odhadnutá síla útoku klubu Sparta Praha

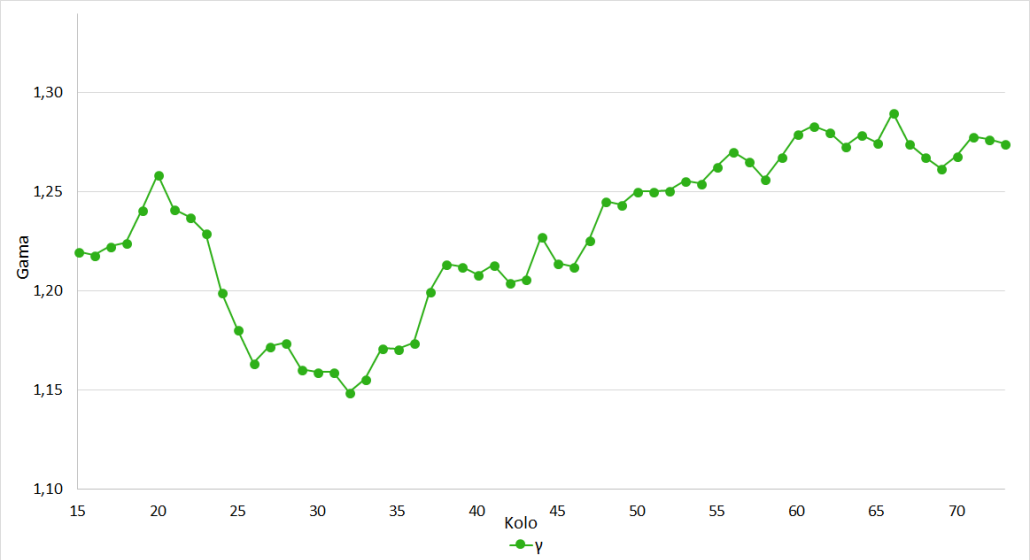
je v průběhu celé sezóny vyšší, než síla útoku Libereckého týmu. Na první pohled také u vývoje odhadů pro pražský tým upoutá propad u 40. kola, který je způsoben dvěma prohrami za sebou (1:5 proti Olomouci a 0:4 proti Zlínu). Další značný skok u 56. kola je způsoben nezvykle vysokou výhrou Sparty nad Chomutovem 7:1. Počáteční růst parametru α libereckého týmu je následkem několika výher s větším gólovým rozdílem, přerušených jen jednogólovými prohrami či remízami. Zpomalení růstu a následný propad od 60. kola je způsoben několika remízami a následnou prohrou 1:5 proti týmu z Hradce Králové.



Obrázek 3.3: Vývoj parametru β u týmů Bílí Tygři Liberec a HC Sparta Praha

Zatímco nižší hodnota síly útoku u týmu Bílí Tygři Liberec byla u vítězného týmu sezóny poněkud překvapivá, nižší hodnotu parametru β, tedy slabosti obrany již lze očekávat. Hodnoty parametru libereckého týmu se během sezóny příliš neliší, dá se tedy říci, že odhadnuté obranné schopnosti týmu byly v sezóně stálé. Zajímavé jsou klesající hodnoty parametru u týmu HC Sparta Praha, tento jev je možné vyložit jako zlepšující se obranné schopnosti týmu v průběhu sezóny.

Pro tuto bakalářskou práci je velmi podstatný i parametr γ, tedy výhoda domácího prostředí. Na obrázku 3.4 je znázorněno, jak se tento parametr vyvíjel u dvojitého Poissonova modelu, ve kterém byl brán jako globální. Po prvních dvaceti kolech sezóny dochází v odhadech parametru γ k poklesu zřejmě způsobenému výraznějšími prohrami domácích týmů.
Od 32. kola už odhadnuté hodnoty parametru rostou až na výsledných 1,275.



Obrázek 3.4: Vývoj parametru γ u dvojitého Poissonova modelu

Po odhadnutí všech parametrů sezóny 2011–2012 bylo třeba nalézt optimální hodnotu parametru ξ. K tomuto účelu byla využita tabulka na listu *ksi 2011–2012* ze souboru *Dvojitý Poissonův model.xlsm*. Parametr ξ, taktéž představený ve článku Dixona a Colese (1997), se vyskytuje ve funkci $\tau(t_m)$ přidělující starším výsledkům nižší váhu. Zahrnutí funkce $\tau(t_m)$ do modelu zajistí proměnlivost parametrů v čase, tím pádem bude možné vzít v úvahu rozdíly ve hře týmů například na začátku a na konci sezóny. V této práci je čas t počítán ve dnech, stejně jako v článku Marka, Šedivé a Ťoupala (2014).

$$\tau(t_m) = \begin{cases} 0 & , \text{pro } t_m \geq \mathrm{T} \\ \mathrm{e}^{-\xi(T-t_m)/365{,}25} & , \text{pro } t_m < \mathrm{T} \end{cases} \tag{3.9}$$

Protože při použití funkce $\tau(t_m)$ by odhadování pomocí metody maximální věrohodnosti vedlo k výsledku $\xi \to +\infty$, byl přejat přístup z článku autorů Dixona a Colese (1997). Ti pro odhadnutí parametru ξ definovali hodnotící funkci:

$$S(\xi) = \sum_{m=1}^{M} (\delta_m^H \ln p_m^H + \delta_m^D \ln p_m^D + \delta_m^A \ln p_m^A), \tag{3.10}$$

kde p_m^D, p_m^H a p_m^A jsou pravděpodobnosti remízy, výhry domácích a výhry hostů vypočtené podle modelu a δ_m je funkce, nabývající hodnoty 1 nebo 0

- $\delta_m^H = 1$, $\delta_m^D = 0$, $\delta_m^A = 0$ u výhry domácího týmu
- $\delta_m^H = 0$, $\delta_m^D = 1$, $\delta_m^A = 0$ v případě, že dojde k remíze
- $\delta_m^H = 0$, $\delta_m^D = 0$, $\delta_m^A = 1$ u výhry hostů

Vzhledem k tomu, že ve funkci $S(\xi)$ jsou vždy sčítány logaritmy z hodnot nižších než 1, nabývá tato funkce záporných hodnot. Cílem optimalizace je maximalizovat výslednou hodnotu, tedy přiblížit se co nejvíce k nule.

Pro každou sezónu, tou nejstarší počínaje, bylo tedy zvoleno několik hodnot ξ, které byly testovány. Pro každou z nich byl vždy proveden odhad všech individuálních parametrů a navíc parametrů globálních (tedy 30 parametrů u původního modelu a 58 u modelu upraveného) a to po každém kole, kterých bylo v sezónách přibližně 80. Z odhadnutých hodnot (zaznamenaných makrem v tabulce na listech *ksi*) byla následně dopočítána hodnota funkce $S(\xi)$. Tyto výpočty jsou k nahlédnutí v souborech „*Odhad ksi*“ s označením sezóny. Po dopočítání hodnoty funkce $S(\xi)$ pro dostatečné množství hodnot parametru již bylo možné odhadnout, při které hodnotě nabývá funkce maxima.

S informací o optimální výši ξ v sezóně 2011–2012 bylo pokročeno k odhadům pro sezónu 2012–2013. Nicméně i zde bylo nutné ozkoušet značné množství hodnot, protože se optimální ξ velmi liší. Tento postup byl poté postupně aplikován i na zbývající sezóny. Tabulka 3.1 obsahuje získané optimální hodnoty pro původní model.

Tabulka 3.1: Optimální parametr ξ v jednotlivých sezónách

Sezóna	ξ	**S(ξ)**
2011–2012	0	-325,50
2012–2013	5	-330,46
2013–2014	2,5	-281,64
2014–2015	1,2	-268,36
2015–2016	0,3	-302,67

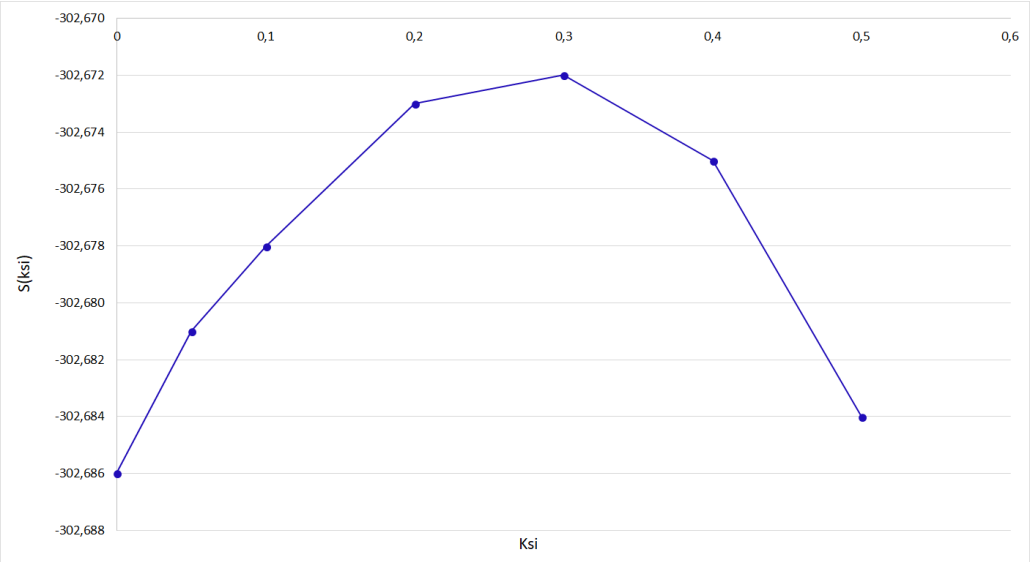
Nápadně vysokou hodnotu v sezóně 2012–2013 si lze vysvětlit jako snahu modelu „zapomenout“ první část dat způsobenou velkými odlišnostmi ve výsledcích první části sezóny. V tabulce 3.2 je pro porovnání uvedeno pořadí týmů po odehrání části zápasů (získané z článku na stránkách iDnes.cz (2012)) a na konci základní části. V dalších sezónách pak s rostoucí stabilitou výkonů, podávaných extraligovými týmy, klesá optimální hodnota parametru ξ přibližně na polovinu oproti minulé sezóně, a to až na hodnotu 0,3 pro sezónu 2015–2016.

Při bližším pohledu na vývoj hodnot funkce $S(\xi)$ v jednotlivých sezónách je možné nalézt určitou podobnost.

Tabulka 3.2: Pořadí týmů v průběhu na na konci základní části 2012–2013

30.10.2012		**26.2.3013**	
Tým	**Body**	**Tým**	**Body**
Škoda Plzeň	37	Zlín	94
Kometa Brno	33	Slavia Praha	94
Rytíři Kladno	32	Škoda Plzeň	89
Oceláři Třinec	31	Oceláři Třinec	86
Zlín	30	**Sparta Praha**	86
Slavia Praha	28	Verva Litvínov	83
ČSOB Pojišťovna Pardubice	26	Rytíři Kladno	77
Verva Litvínov	25	Mountfield České Budějovice	76
Piráti Chomutov	22	Vítkovice Steel	75
Vítkovice Steel	21	ČSOB Pojišťovna Pardubice	73
Mountfield České Budějovice	21	**Kometa Brno**	72
Energie Karlovy Vary	19	Energie Karlovy Vary	67
Bílí Tygři Liberec	18	Bílí Tygři Liberec	63
Sparta Praha	14	Piráti Chomutov	57

Při změnách parametru ξ dochází jen k malým změnám v hodnotách fuknce $S(\xi)$ (viz obr. 3.5), model je tedy na tyto změny málo citlivý. Rychlejší změny můžeme pozorovat v hodnotách vyšších, než je optimum (tedy napravo od něj). S ohledem na tyto skutečnosti bude pro predikci výsledků v sezóně 2016–2017 zvolena hodnota $\xi = 0,2$.



Obrázek 3.5: Vývoj funkce $S(\xi)$ pro sezónu 2015–2016

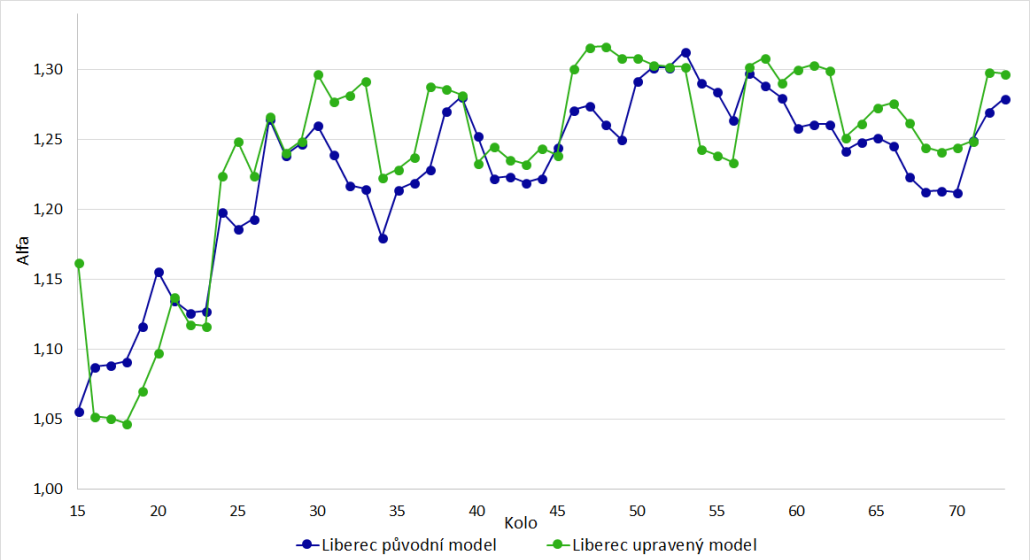
Při odhadování parametrů u upraveného dvojitého Poissonova modelu bylo postupováno stejným způsobem. Nejprve byly testovány optimální hodnoty ξ z původního modelu, ale bylo opět nutné vyzkoušet několik dalších hodnot, protože optimální hodnoty ξ (viz tabulka 3.3) se mírně liší. Nicméně i zde lze pozorovat postupné snižování hodnot až na $\xi = 0,5$ pro sezónu 2015–2016. Vzhledem k tomu byla pro predikci zvolena hodnota $\xi = 0,3$. Počet testovaných hodnot parametru ξ je pro upravený model nižší i proto, že přidáním dvaceti osmi parametrů k odhadnutí se zvýšila složitost výpočtu a tím i časová náročnost celého procesu odhadování.

Tabulka 3.3: Optimální parametr ξ v jednotlivých sezónách pro upravený i původní model

	Původní model		**Upravený model**	
Sezóna	ξ	**S(ξ)**	ξ	**S(ξ)**
2011–2012	0	-325,50	0	-334,46
2012–2013	5	-330,46	3	-338,68
2013–2014	2,5	-281,64	2	-281,99
2014–2015	1,2	-268,36	1	-276,60
2015–2016	0,3	-302,67	0,5	-306,18

Vyšší hodnoty funkce $S(\xi)$ oproti předchozímu modelu napovídají, že úprava modelu mohla bohužel způsobit zhoršení jeho predikčních schopností. Možné vysvětlení tohoto jevu spočívá v „přeparametrizování“ modelu. Toto riziko vyplynulo už i z jiných článků, například složitější modely představené Maherem (1982) se nakonec ukázaly jako horší z důvodu příliš vysokého počtu parametrů.

Pro porovnání ještě uveďme graf vývoje odhadů parametru α u původního modelu a u modelu upraveného (obr. 3.6). Zdá se, že hodnoty odhadnuté pomocí upraveného dvojitého Poissonova modelu mají výraznější extrémy, tedy více odráží změny ve vstupních datech.



Obrázek 3.6: Vývoj odhadů parametru α pro sezónu 2015–2016 pomocí původního a upraveného modelu

4 Srovnání modelů

Modely byly porovnávány dvěma způsoby. Prvním z nich bylo vypočítání několika kritérií, díky nimž bylo možné určit kvalitu modelů, a druhým bylo ověření účinnosti modelů na fiktivním sázení proti sázkové kanceláři.

4.1 Srovnání dle kritérií

Po odhadnutí všech potřebných parametrů a volbě vhodného ξ již bylo možné provést pomocí obou modelů odhad výsledků utkání ze sezóny 2016–2017.

Na začátku každé sezóny bylo vyčleněno prvních patnáct kol (datumů) k ustálení odhadů. Volba patnácti kol zaručuje dostatečný počet dat, aby bez problémů proběhl výpočet a zároveň umožňuje ustálení odhadů. Poté byl proveden odhad výsledků v šestnáctém kole (na základě dat do patnáctého kola), přehodnoceny odhady a provedena předpověď na kolo sedmnácté. Tento postup byl opakován až k poslednímu kolu. V každém z kol byla následně vypočítána pravděpodobnost výhry domácích, výhry hostů a remízy.

Kvalita predikcí obou modelů byla hodnocena na základě několika kritérií: Prvním z nich je již zmíněná funkce $S(\xi)$, která ukazuje míru chyby v předpovědích v celé sezóně, případně lze použít její úpravu (*LogLoss* function). Podle tohoto kritéria vychází jako lepší původní dvojitý Poissonův model (viz tabulka 3.3).

Logaritmická ztrátová funkce (*LogLoss* function) je definována jako

$$LogLoss = -\frac{1}{M}\sum_{m=1}^{M}(\delta_m^H \ln p_m^H + \delta_m^D \ln p_m^D + \delta_m^A \ln p_m^A). \tag{4.1}$$

Jde tedy o funkci $S(\xi)$ vynásobenou hodnotou -1 a vydělenou počtem zápasů. Jak je zmíněno výše, p_m^D, p_m^H a p_m^A jsou pravděpodobnosti remízy, výhry domácích a výhry hostů vypočtené podle modelu a δ_m je funkce, nabývající hodnoty 1 nebo 0

- $\delta_m^H = 1$, $\delta_m^D = 0$, $\delta_m^A = 0$ u výhry domácího týmu
- $\delta_m^H = 0$, $\delta_m^D = 1$, $\delta_m^A = 0$ v případě, že dojde k remíze

- $\delta_m^H = 0$, $\delta_m^D = 0$, $\delta_m^A = 1$ u výhry hostů

Výsledky funkce $S(\xi)$ i $LogLoss$ funkce v tomto případě dávají stejnou informaci, jelikož počty zápasů jsou u obou modelů stejné. V případě, že by se hodnoty M lišily, byla by $LogLoss$ funkce vhodná pro porovnávání.

Dalším kritériem, které lze použít, je takzvaná Kalibrace (Cal) popsaná v článku Kovalchiková (2015) na výsledcích tenisových zápasů. V Kalibraci je porovnáván součet pravděpodobností výhry favorita vypočítaných určeným modelem s počtem zápasů, ve kterých favorit opravdu vyhrál. Cal tedy nabývá kladných hodnot - čím blíže jedné, tím lépe kalibrovaný model je. Vzorec pro vypočtení kritéria Cal má tedy tvar:

$$Cal = \frac{\sum_{m=1}^{M} max(p_m^H, p_m^D, p_m^A)}{\sum_{m=1}^{M} \sigma_m}, \tag{4.2}$$

kde M je počet zápasů, p_m^D, p_m^H a p_m^A jsou již výše zmíněné pravděpodobnosti remízy, výhry domácích a výhry hostů a σ_m je funkce, která nabývá hodnoty 1 v případě, že m-tý zápas skončí výhrou favorita a 0 jinak.

Dalším z používaných kritérií je tvz. Přesnost (Ac) modelu ve tvaru

$$Ac = \frac{\sum_{m=1}^{M} \sigma_m}{M}, \tag{4.3}$$

kde je M opět počet zápasů a σ_m funkce použitá již při výpočtu kritéria Cal (nabývá hodnoty 1 v případě, že m-tý zápas skončí výhrou favorita a 0 jinak).

Výsledné hodnoty k porovnání jsou v tabulce 4.1

Tabulka 4.1: Výsledné hodnoty jednotlivých kritérií pro ověřované modely

Kritérium	**Původní model**	**Upravený model**
$S(\xi)$	-315,050	-324,005
$LogLoss$	1,090	1,121
Cal	1,214	1,212
Ac	0,481	0,484

Zatímco při pohledu na výsledky funkce $S(\xi)$, potažmo kritéria $LogLoss$, vychází původní Dvojitý Poissonův model jako kvalitnější, podle kritérií Cal a Ac, ve kterých je zohledněn počet správně určených vítězů, jsou modely

srovnatelné a upravený model je dokonce o trochu přesnější. Výpočty a výsledky jsou obsaženy v souboru *Ověření.xlsm* a *Ověření (nový model).xlsm* na listech *Odhad výsledků*.

4.2 Srovnání dle sázení

Poslední možností, jak ověřit přesnost predikcí modelu, která bude v této práci představena, je imaginární sázení proti sázkové kanceláři (základy viz Cover, Thomas (2006)).

Předpokládejme „neomezený“ bank (maximálně $počet zápasů \cdot h$) a konstantní výši sázky $h = 10\,Kč$. Přístup k sázení je přejatý z článku autorů Marka, Šedivé a Ťoupala (2014). Nejprve je definováno kritérium pro vložení sázky φ jako

$$\varphi = p_m^R \cdot o_m^R \tag{4.4}$$

kde p_m^R, $R \in \{H, D, A\}$ je opět pravděpodobnost výhry domácích, remízy a výhry hostů a o_m^R je kurz vypsaný sázkovou kanceláří na příslušný výsledek. Pokud je hodnota tohoto kritéria vyšší než stanovená mez L, pak dojde k vložení sázky. Hodnota L musí být vždy vyšší nebo rovna jedné, aby byla sázka výhodná.

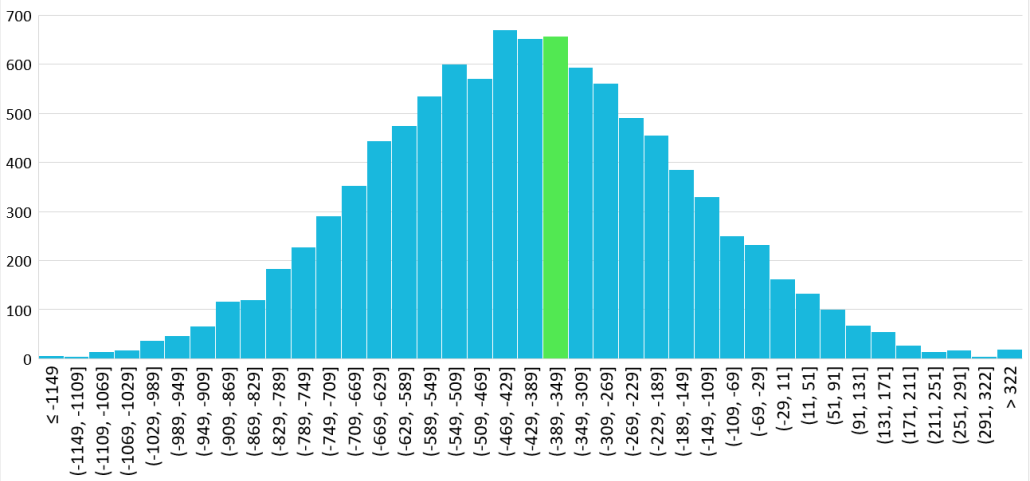
V souboru *Ověření.xlsm* a *Ověření (nový model).xlsm* na listech *Odhad výsledků* je vypočítána výše výhry/prohry při stanoveném L. Tyto výsledky jsou shrnuty i v tabulce 4.2.

Tabulka 4.2: Vývoj výher při různém L

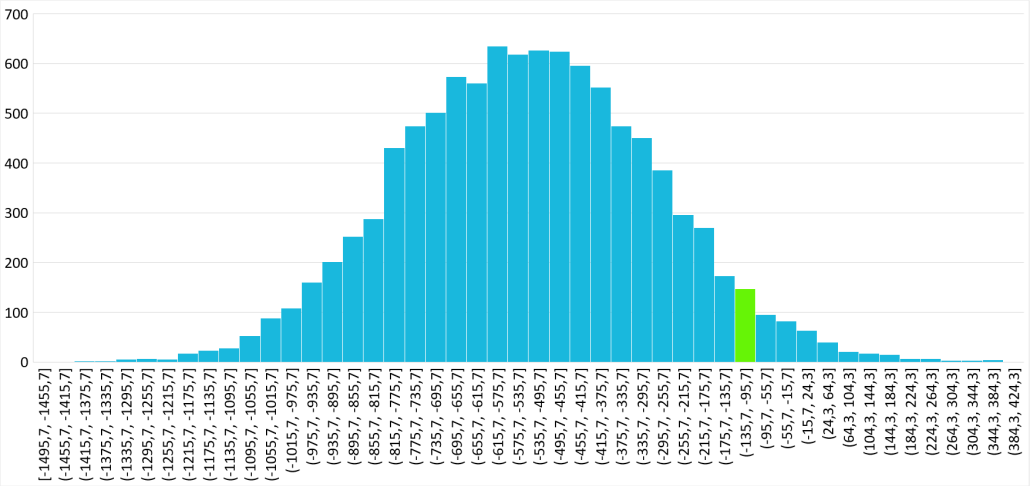
L	**Původní model**	**Upravený model**
1,00	-359,9	-126,4
1,05	12,9	-77,5
1,25	70,5	-181,3
1,50	-30,0	-42,8
1,70	-10,0	-17,6
1,80	0,0	-20,0

Se zvyšujícím se L klesá počet sázek a výše výhry se tím pádem stává spíše dílem náhody, než hodnotou vypovídající o kvalitě modelu.

Mějme L stanoveno například jako nejnižší možné, tedy 1. U původního modelu by v tomto případě bylo vsazeno na $85,43\,\%$ zápasů, zatímco u upraveného modelu by to bylo $86,09\,\%$. Pro porovnání výsledků s naivními metodami sázení bylo nutné vytvořit simulaci s náhodným sázením. To spočívá ve vsazení stejného kapitálu na stejně procent náhodně vybraných zápasů. Výběr výsledku, na který bude v tomto případě vsazeno, byl také náhodný (použitím funkce „Náhčíslo“ a „Randbetween“ v programu Microsoft Excel). Výsledky simulací viz obr. 4.1 a 4.2 obsahují i vyznačený sloupec, do kterého spadá hodnota výhry při sázení podle vybraného modelu.



Obrázek 4.1: Histogram četností výher/proher v porovnání s původním modelem



Obrázek 4.2: Histogram četností výher/proher v porovnání s upraveným modelem

Další naivní metody sázení, jako například sázení na domácí, hosty, nebo na „outsidera“ (nejvyšší kurz) není možné přímo porovnávat s hodnotami získanými použitými modely, protože u těchto metod je vsazeno na každý ze zápasů. Určité porovnání je možné po vytvoření „výhry/prohry za vsazený zápas“ (tabulka 4.3).

Tabulka 4.3: Porovnání výher/proher

Metoda	**Celková výhra**	**Výhra za zápas**
Domácí	-284,6	-0,985
Hosté	-670,4	-2,319
Outsider	114,7	0,395
Původní model $L = 1,00$	-359,9	-1,395
Upravený model $L = 1,00$	-126,4	-0,486
Původní model $L = 1,05$	12,9	0,066
Upravený model $L = 1,05$	-77,5	-0,359

V tabulce jsou uvedeny výsledky naivních metod spolu s výsledky představených modelů při hodnotě $L = 1,00$, tedy se sázkou na nejvyšší možný počet zápasů a dále při $L = 1,05$, která v tomto případě znamená sázku přibližně na dvě třetiny zápasů. Výše sázky zůstává konstantní, a to $h = 10\,K$č. Z naivních metod sázení dává pro odhadovanou sezónu nejlepší výsledky metoda sázení na outsidera, tedy na nejvyšší kurz. Tato metoda zároveň vychází jako nejlepší i v porovnání s představenými modely. Nejhorší výsledky naopak poskytuje sázení striktně na hostující tým. Zajímavý je i výsledek původního dvojitého Poissonova modelu s $L = 1,05$, který je už mírně v kladných hodnotách. Při využití modelů pro sázení by bylo vhodné se zaměřit na nejvíce výdělečné varianty a jejich kombinace, tedy představené modely s hodnotou $L > 1,05$ a metodu sázení na outsidera.

5 Závěr

Cílem bakalářské práce bylo zjistit, zda je možné pomocí matematických modelů odhadovat výsledky utkání v ledním hokeji. Nejprve byl představen dvojitý Poissonův model z článku od autorů Marka, Šedivé a Ťoupala (2014) a následně jeho inovace, tedy upravený dvojitý Poissonův model. Úprava modelu spočívala v novém návrhu provázání jednotlivých výsledků, díky kterému bylo možné považovat vliv domácího prostředí na jednotlivé týmy za individuální vlastnost. Oba modely mají základ v původním Maherově návrhu z roku 1982, který vytvořil model popisující výsledky fotbalových utkání.

Oba představené přístupy byly následně aplikovány na data z české hokejové extraligy mezi sezónami 2011–2012 a 2015–2016. Výsledky z těchto sezón byly využity pro predikci výsledků v sezóně 2016–2017. Kvalita předpovědí byla poté ověřena pomocí několika kritérií, jako například $LogLoss$ funkce či kalibrace (Cal). Podle těchto výpočtů se jeví původní model jako o něco přesnější, ačkoliv kritéria se základem v počtu správně určených vítězů zápasů ukazují na mírnou výhodu modelu upraveného. V závěru práce jsou predikce použity na imaginární sázení proti sázkové kanceláři. Výsledky sázení jsou velmi závislé na zvolené výši parametru L, který určuje, od jaké hodnoty kritéria φ bude vsazeno. Celkové výhry/prohry jsou porovnány s naivními metodami sázení, jako například náhodné sázení. Oproti této metodě se jeví oba z modelů jako lepší, ačkoliv jsou ve většině případů (s rozdílnou hodnotou L) ztrátové. Tento fakt je možné odůvodnit marží sázkové kanceláře a také skutečností, že způsob výpočtu pravděpodobností výhry sázkovou kanceláří je velmi efektivní. Při sázení proti průměrnému kurzu z několika sázkových kanceláří, či dokonce proti nejvyšším kurzům napříč trhem se dají očekávat výsledky v kladných hodnotách, což by bylo možné zajímavé rozšíření této práce.

Literatura

Abdi, H., (2007). The Bonferonni and Šidák corrections for Multiple Comparisons. The University of Texas at Dallas [cit. 2018-04-17]. http://www.utdallas.edu/ herve/Abdi-Bonferroni2007-pretty.pdf.

BetExplorer.com, (2018). Hockey - Czech Republic. [cit. 2018-04-17]. http://www.betexplorer.com/hockey/czech-republic/.

Buttrey, S. E., (2016). Beating the market betting on NHL hockey games. *Journal of Quantitative Analysis in Sports*, 12(2), 87-98.

Cover, T. M. a Thomas, J. A., (2006). *Elements of Information Theory.* Hoboken NJ: John Wiley & Sons, Inc.

Dixon, M. J. a Coles, S. G., (1997). Modelling association foootball scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society*, 46(2), 265-280.

Hátle, J. a Likeš, J., (1974). *Základy počtu pravděpodobnosti a matematické statistiky.*

iDNES.CZ, (2012). Týmy ze dna extraligy v dohrávkách zabraly, vyhrály Sparta i Liberec. [cit. 2018-04-17]. https://hokej.idnes.cz/liberec-kometa-sparta-chomutov-dmt-/hokej.aspx?c=A121030_162119_hokej_cig.

Karlis, D. a Ntzoufras, I., (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society,* 381-393.

Kovalchik, A., S., (2016). Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports,* 127-138.

Maher, M. J., (1982). Modelling association football scores. *Statistica Neerlandica,* 36, 109-118.

Marek, P. a Vávra, F., (2017). Home team advantage in english premier league, 244-254, Padova university Press, ISBN 978-88-6938-058-7.

Marek, P., Šedivá, B., a Ťoupal, T., (2014). Modelling and prediction of ice hocey match results. *Journal of Quantitative Analysis in Sports,* 357-365, ISSN: 1559-0410.

Reif, J., (2004). *Metody matematické statistiky.* Západočeská univerzita.

Sfstats.net., (2018). Extraligue. [cit. 2018-04-17].
http://www.sfstats.net/hockey/leagues/2_Extraleague.

Spinelli, J. J. a Stephens, M. A., (1997). Cramér-von mises tests of fit for the poisson distribution. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique,* 257–268.

SPORT.CZ, (2018). Tipsport extraliga. [cit. 2018-04-17].
https://www.sport.cz/hokej/extraliga/#vysledky.

A Přílohy

A.1 Tabulka kritických hodnot pro Cramér von Mises test

Tato tabulka je přejata přímo z článku autorů Spinelli a Stephens (1997).

Tabulka 5.1: Kritické hodnoty

λ	W^2	A	W_m^2
0,1	0,016	0,46	0,031
0,5	0,164	1,104	0,35
1	0,203	1,191	0,624
2	0,182	1,151	0,881
5	0,172	1,112	1,359
10	0,169	1,099	1,906
20	0,167	1,093	2,685
50	0,166	1,089	4,235
100	0,166	1,088	5,984
∞	0,165	1,087	

A.2 Zdrojový kód

Příloha obsahuje zdrojový kód v jazyku Visual Basic for Applications použitý při odhadování parametrů obou modelů.

```
1  Sub OdhadParametru(sezona As String)
2  '
3  ' OdhadParametru Makro
4  ' Odhad alpha a beta
5  '
6  Dim pocetKol As Integer
7  Dim i As Integer
8  Dim data As Worksheet
9  Dim vysledek As Worksheet
10 Dim pomIndex As Integer
11 Dim pozice As Integer
12
```

```
13
14 Application.ScreenUpdating = False
15
16 Set data = Sheets(sezona)
17 Set vysledek = Sheets("vysledky " + sezona)
18 Set ksi = Sheets("ksi " + sezona)
19
20 pocetKol = WorksheetFunction.Max(data.Range("G:G"))
21 pozice = WorksheetFunction.Count(data.Range("G:G")) + 4
22
23 vysledek.Range("B5:B21").Value = data.Range("B5:B21").Value
24 vysledek.Range("B23:B36").Value = data.Range("B5:B18").Value
25
26 For i = 15 To pocetKol
27     pomIndex = i - 12
28     data.Range("AD6").Value = i
29     SolverSolve UserFinish:=True
30
31     vysledek.Cells(4, pomIndex).Value = i
32     vysledek.Range(vysledek.Cells(5, pomIndex), vysledek.Cells
      (21, pomIndex)).Value = data.Range("C5:C21").Value
33     vysledek.Range(vysledek.Cells(23, pomIndex), vysledek.Cells
      (36, pomIndex)).Value = data.Range("D5:D18").Value
34
35     Do While pozice > 4
36         pomIndex = data.Range("G" & pozice).Value
37
38         If pomIndex = i + 1 Then
39             ksi.Range("B" & pozice & ":" & "O" & pozice).Value =
       data.Range("F" & pozice & ":" & "S" & pozice).Value
40         ElseIf pomIndex > i + 1 Then
41             Exit Do
42         End If
43
44         pozice = pozice - 1
45     Loop
46 Next i
47
48 Application.ScreenUpdating = True
49
50 End Sub
```

A.3 Elektronické přílohy na CD-ROM

1. *BP Hellusová.pdf*: Text bakalářské práce.
2. *Cramer von Mises.xlsx*: Cramér von Mises test, zda se data řídí Poissonovým rozdělením.
3. *Data a Poisson.xlsx*: Zdrojová data a chí-kvadrát test, zda se data řídí Poissonovým rozdělením.
4. *Dvojitý Poissonův model.xlsm*: Odhadování parametrů u původního modelu.
5. *Nezávislost.xlsx*: Testování nezávislosti.
6. *Odhad ksi 11–12.xlsx*: Odhadování optimálního parametru ξ pro sezónu 2011–2012.
7. *Odhad ksi 12–13.xlsx*: Odhadování optimálního parametru ξ pro sezónu 2012–2013.
8. *Odhad ksi 13–14.xlsx*: Odhadování optimálního parametru ξ pro sezónu 2013–2014.
9. *Odhad ksi 14–15.xlsx*: Odhadování optimálního parametru ξ pro sezónu 2014–2015.
10. *Odhad ksi 15–16.xlsx*: Odhadování optimálního parametru ξ pro sezónu 2015–2016.
11. *Odhad ksi 11–12(upravený model).xlsx*: Odhadování optimálního parametru ξ u upraveného modelu pro sezónu 2011–2012.
12. *Odhad ksi 12–13(upravený model).xlsx*: Odhadování optimálního parametru ξ u upraveného modelu pro sezónu 2012–2013.
13. *Odhad ksi 13–14(upravený model).xlsx*: Odhadování optimálního parametru ξ u upraveného modelu pro sezónu 2013–2014.
14. *Odhad ksi 14–15(upravený model).xlsx*: Odhadování optimálního parametru ξ u upraveného modelu pro sezónu 2014–2015.
15. *Odhad ksi 15–16(upravený model).xlsx*: Odhadování optimálního parametru ξ u upraveného modelu pro sezónu 2015–2016.
16. *Ověření.xlsm*: Ověřování predikčních schopností původního modelu.
17. *Ověření (upravený model).xlsm*: Ověřování predikčních schopností upraveného modelu.

18. *Upravený dvojitý Poissonův model.xlsm*: Odhadování parametrů u upraveného modelu.