Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra matematiky

Statistická analýza sportovních výsledků

BAKALÁŘSKÁ PRÁCE

Plzeň, 2013 **Kateřina Šedivá**

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracovala samostatně a použila jsem pouze podklady (literatura, software apod.) uvedené v přiloženém seznamu.

V Plzni dne _____ _____

podpis

Poděkování

Na tomto místě bych chtěla poděkovat vedoucímu mé bakalářské práce RNDr. Petrovi Stehlíkovi, Ph.D. za jeho cenné rady a čas, který mi při tvorbě této práce věnoval. Dále bych ráda poděkovala Danielu Novákovi za poskytnutí dat pro zpracování. Především bych však chtěla poděkovat členům mé rodiny a mým nejbližším za bezmeznou podporu po celou dobu mého studia.

Abstrakt

Cílem této práce je zpracovat a analyzovat shromážděná data házenkářských soutěží. Pro zpracování byla shromážděna data z nejvyšší mužské házenkářské soutěže v Německu a České republice za posledních 10 sezón, v letech 2002 - 2012. V první fázi praktické části práce byly určeny základní statistiky a v následující části práce bylo provedeno testování tří hypotéz pomocí vhodných statistických testů. Konkrétně se jednalo o hypotézy o normalitě vstřelených gólů, o dynamice v soutěži a o dynamice mezi oběma soutěžemi. V dalších kapitolách praktické části byl proveden výpočet rankingů pro jednotlivé týmy samostatně za celé sledované období a samostatně pro jednotlivé sezóny. Předpovědi výsledků zápasů získané na základě porovnání rankingů byly porovnány se skutečnými výsledky.

Klíčová slova: házená, regresní modely, ranking, Colley Matrix Model, Keener Ranking Model, Offense-Defense Model

Abstract

The aim of this thesis is to collect and analyze results of handball matches. Assembling matches from 2002-2012, we construct a dataset consisting of 4674 matches from top men's handball leagues in Germany and the Czech Republic. First, we present a concise overview of selected statistical methods for hypothesis testing and three ranking algorithms. Next, we provide basic descriptive statistics of the dataset and study hypotheses regarding the normality of the scoring, its dynamics within the game and within the season. Finally, rankings are computed and their prediction rates are studied.

Key words: handball, regression model, ranking, Colley Matrix Model, Keener Ranking Model, Offense-Defense Model

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Výsledky sportovních zápasů jsou velmi zajímavou oblastí pro statistickou analýzu dat. Výhodou analýzy těchto dat je dostupnost velkého množství dat ve snadno interpretovatelné formě.

Tato práce je zaměřena na výsledky zápasů z mužské házené, konkrétně z německé a české nejvyšší soutěže. Házená byla vybrána proto, že se jedná o dynamický sport, kde v zápasech padá velké množství gólů a je tedy možno předpokládat normalitu získaných dat. V našem případě se jedná o data za posledních deset let v obou soutěžích, konkrétně 4674 odehraných zápasů, proto tedy lze říci, že máme k dispozici slušné množství dat. Podrobnému popisu datového souboru, který byl v práci zpracováván a základním statistikám bude věnována Kapitola 1.

Po krátkém představení používaných dat budou následovat Kapitoly 2 a 3, ve kterých je uvedena teoretická část k použitým statistickým metodám na testování hypotéz a metodám pro výpočet rankingů. Budou zde tedy tyto metody matematicky formulovány a popsány, tak jak budou dále použity při praktickém zpracování dat.

V Kapitole 4 je možno nalézt praktické zpracování získaných dat. Nejprve byla testována hypotéza o tom, zda data mají normální rozdělení. Dále byly testovány tři následující hypotézy: zda existuje lineární závislost mezi počtem gólů a počtem odehraných kol v sezóně, zda konečný výsledek závisí na poločasovém výsledku a zda je německá házenkářská liga dynamičtější než česká házenkářská liga.

V poslední Kapitole 5 budou určeny rankingy jednotlivých týmů dle tří zvolených modelů. Jedná se o Colley Matrix Model, Keener Ranking Model a Offense-Defense Model. Rankingy budou počítány pro týmy z německé házenkářské soutěže za celé sledované období a také pro jednotlivé sezóny. Hodnoty rankingů jednotlivých týmů budou využity k porovnání s reálnými výsledky zápasů. Na základě aktuálních rankingů budou předpovězeny výsledky zápasů a následně porovnány s reálnými výsledky.

V závěru práce budou shrnuty zajímavé poznatky a výsledky této práce.

Kapitola 1

Data z házenkářských soutěží

1.1 Popis dat

Získaná data pro statistickou analýzu sportovních výsledků jsou výsledky zápasů v házené z nejvyšší soutěže mužů v České republice a v Německu z let 2002-2012. Výsledky všech zápasu jsou získána od pana Daniela Nováka, administrátora internetového serveru The Czech Handball Server [6].

Z německé nejvyšší mužské soutěže je k dispozici 3060 výsledků zápasů a z české nejvyšší mužské soutěže 1614 výsledků, dohromady tedy 4674 výsledků zápasů. Pro každý zápas byly k dispozici tyto informace: datum a čas utkání, názvy soupeřů, určení domácího a hostujícího celku, výsledek v poločase a na konci utkání, jména rozhodčích a identifikační číslo zápasu. Z těchto údajů bylo určeno několik dalších charakteristik utkání. Například rozdíl výsledku utkání v prvním poločase, druhém poločase a také na konci zápasu. Dále bylo ke každému výsledku přiřazeno v poločase a také na konci písmeno V nebo P nebo R, které určují, zda domácí celek v dané fázi zápasu vyhrál, prohrál nebo remizoval [6].

1.1.1 Nedokonalá data

Získaná data byla po stažení ze serveru nedokonalá, proto bylo potřeba data upravit. V několika případech se v datech objevil problém se špatným přiřazením u některých výsledků zápasů ke správnému hracímu kolu. Tento problém se podařilo odhalit v okamžiku, kdy bylo kontrolováno, zda v každém kole byl sehrán odpovídající počet zápasů.

Dalším problémem byl zápas v německé soutěži, ke kterému byl přiřazen výsledek 0:0. Tento zápas byl proto z dalšího zpracování dat vyřazen.

Vážnějším problémem však bylo časté přejmenovávání jednotlivých týmů v soutěžních ročnících. Tento problémový jev, který byl s největší pravděpodobností způsobem častým měněním sponzorů, se projevuje v české nejvyšší soutěži, naopak v německé nejvyšší soutěži zůstávají názvy celků stejné ve všech ročnících. Veškeré změny názvů českých týmů jsou zobrazeny v Tabulce 1.1.

Tabulka 1.1: Názvy týmů používané v české lize

Město	**Název týmu**
Frýdek Místek	Radegast SKP Frýdek-Místek , HC Frýdek-Místek, SKP Frýdek-Místek, SKP ARCIMPEX Frýdek-Místek
Plzeň	HSC Plzeň, TJ Lokomotiva Plzeň, SSK Talent M.A.T. Plzeň
Brno	Házená Brno, Házená Brno s.r.o., KP Brno
Karviná	HC Baník Karviná, HC Baník Karviná s.r.o., HCB OKD Karviná
Zubří	HC Zubří HC Gumárny Zubří
Hranice	Cement Hranice, TJ Cement Hranice
Jičín	CS CARGO HBC Jičín, HBC RONAL Jičín
Prešov	ŠK Farmakol Tatran Prešov, Tatran Prešov
Lovosice	HK Město Lovosice HK.A.S.A Město Lovosice
Třeboň	TJ Jiskra Třeboň, TJ Jiskra Lázně Třeboň
Košice	1. MHK Košice
Dvůr králové nad Labem	1. HK Dvůr Králové nad Labem
Dukla Praha	HC Dukla Praha
Allrisk Praha	Allrisk CAC Praha
Hustopeče	Házená Legata Hustopeče
Topolčany	HC THP TOPVAR Topolčany
Kopřivnice	KH Kopřivnice
Bardejov	MHC Bardejov
Považská Bystrica	MŠK Považská Bystrica
Louny	SHK Knauf-Chemiko Louny
Přerov	Sokol HC Přerov
Kostelec na Hané	Sokol Kostelec na Hané
Bratislava	ŠKP Bratislava
Sečovce	ŠKP Sečovce
Bystřice pod Hostýnem	TJ Bystřice pod Hostýnem
Hlohovec	TJ Drot'ovna Hlohovec
Napajedla	TJ Fatra Slavia Napajedla

1.2 Základní statistické zpracování dat

V další fázi statistického zpracování dat byly určeny základní statistické charakteristiky. Pro jednotlivé týmy obou sledovaných soutěží byly spočteny počty všech utkání, které za sledované období jednotlivý tým odehrál. Dále byly podrobně sledovány počty výher, počty proher a počty remíz a byl určen jejich podíl z celkového počtu utkání. Jako doplňkové informace byly spočteny průměrné počty gólů, které tým vstřelil na domácí a na hostující palubovce. Poslední informací, která je obsažena v Tabulce 1.2 a 1.3, je počet sezón, který daný tým odehrál a průměrné pořadí, kterého za dobu svého působení v soutěži dosáhl.

Tabulka 1.2: Základní statistické zpracování dat (německá liga)

Tým	Počet utkání	Počet výher	Podíl výher	Počet proher	Podíl proher	Počet remíz	Podíl remíz	Prům.poč. gólů venku	Prům.poč. gólů doma	Počet sezón	Průměr. pořadí
Bergischer HC	34	8	24%	25	74%	1	3%	27,118	27,059	1	16,000
Concordia Delitzsch	34	4	12%	28	82%	2	6%	24,118	25,529	1	18,000
DHC Rheinland	34	8	24%	26	76%	0	0%	24,294	25,588	1	16,000
Eintracht Hildesheim	68	7	10%	60	88%	1	1%	25,471	27,588	2	18,000
FA Göppingen	340	158	46%	153	45%	29	9%	26,918	29,412	10	8,800
Füchse Berlin	170	95	56%	62	36%	13	8%	28,482	29,447	5	7,200
GWD Minden	272	68	25%	181	67%	23	8%	25,493	27,507	8	14,750
HBW Balingen-Weilstetten	204	55	27%	133	65%	16	8%	25,804	27,392	6	14,333
HSG Ahlen-Hamm	34	6	18%	25	74%	3	9%	26,882	27,412	1	17,000
HSG D/M Wetzlar	272	77	28%	170	63%	25	9%	25,934	27,051	8	13,250
HSG Düsseldorf	136	31	23%	96	71%	9	7%	25,603	26,868	4	15,500
HSG Nordhorn	238	137	58%	84	35%	17	7%	29,437	31,437	7	6,571
HSG Wetzlar	68	23	34%	39	57%	6	9%	25,000	25,735	2	12,000
HSV Hamburg	340	235	69%	79	23%	26	8%	29,488	31,571	10	4,300
MT Melsungen	238	80	34%	140	59%	18	8%	28,353	29,462	7	11,429
Rhein-Neckar-Löwen	170	121	71%	39	23%	10	6%	31,165	32,541	5	4,000
SC Magdeburg	340	211	62%	111	33%	18	5%	29,200	31,929	10	5,900
SG Flensburg-Handewitt	340	259	76%	65	19%	16	5%	30,559	34,018	10	2,900
SG Kronau/Östringen	102	46	45%	50	49%	6	6%	27,392	28,980	3	10,000
SG Wallau-Massenheim	102	41	40%	48	47%	13	13%	28,275	30,725	3	9,333
SG Willstätt/Schutterwald	34	7	21%	25	74%	2	6%	26,000	28,118	1	18,000
Stralsunder HV	68	10	15%	54	79%	4	6%	22,382	24,235	2	17,500
SV Post Schwerin	34	4	12%	29	85%	1	3%	25,647	26,706	1	18,000
TBV Lemgo	340	217	64%	100	29%	23	7%	29,988	31,865	10	5,300
ThSV Eisenach	68	18	26%	45	66%	5	7%	24,794	26,706	2	15,500
THW Kiel	340	287	84%	38	11%	15	4%	32,800	34,182	10	1,700
TSG Ludwigshafen-Friesenheim	34	4	12%	27	79%	3	9%	26,529	28,294	1	18,000
TSV Dormagen	68	13	19%	48	71%	7	10%	25,588	26,235	2	16,000
TSV Hannover-Burgdorf	102	28	27%	66	65%	8	8%	26,922	27,275	3	14,000
TuS N-Lübbecke	272	80	29%	168	62%	24	9%	27,007	28,801	8	12,625
TUSEM Essen	170	66	39%	87	51%	17	10%	26,424	28,353	5	10,600
TV 05/07 Hüttenberg	34	6	18%	23	68%	5	15%	25,824	26,118	1	17,000
TV Großwallstadt	340	136	40%	169	50%	35	10%	26,024	27,994	10	10,400
VfL Gummersbach	340	198	58%	112	33%	30	9%	29,924	30,865	10	6,500
VfL Pfullingen	136	27	20%	94	69%	15	11%	24,574	27,603	4	16,250
Wilhelmshavener HV	204	55	27%	127	62%	22	11%	25,294	27,216	6	13,833
CELKEM	**6120**	**2826**	**46%**	**2826**	**46%**	**468**	**8%**	**26,964**	**28,551**		

Tabulka 1.3: Základní statistické zpracování dat (česká liga)

Tým	Počet utkání	Počet výher	Podíl výher	Počet proher	Podíl proher	Počet remíz	Podíl remíz	Prům.poč. gólů venku	Prům.poč. gólů doma	Počet sezón	Průměr. pořadí
Košice	86	43	50%	39	45%	4	5%	26,395	28,977	3	7,667
Dvůr Králové nad Labem	22	3	14%	19	86%	0	0%	22,636	23,364	1	12,000
Allrisk Praha	58	31	53%	22	38%	5	9%	23,655	27,207	2	6,500
Hranice	224	95	42%	114	51%	15	7%	26,304	29,027	9	7,625
Brno	124	17	14%	101	81%	6	5%	23,903	25,387	5	11,750
Hustopeče	48	10	21%	37	77%	1	2%	24,875	26,042	2	11,500
Jičín	168	79	47%	70	42%	19	11%	23,869	27,429	7	6,167
Karviná	252	192	76%	47	19%	13	5%	29,119	32,262	10	2,444
Dukla Praha	252	175	69%	56	22%	21	8%	27,690	29,730	10	3,222
Zubří	252	155	62%	78	31%	19	8%	26,857	31,532	10	4,333
Lovosice	168	101	60%	60	36%	7	4%	26,929	28,952	7	4,500
Trnava	30	15	50%	15	50%	0	0%	25,733	30,467	1	9,000
Kopřivnice	168	52	31%	106	63%	10	6%	24,560	26,464	7	9,167
Bardejov	26	3	12%	22	85%	1	4%	20,385	24,385	1	14,000
Považská Bystrica	86	56	65%	26	30%	4	5%	25,465	31,558	3	4,000
Louny	30	6	20%	21	70%	3	10%	24,200	28,667	1	13,000
Frýdek-Místek	252	111	44%	127	50%	14	6%	27,183	29,651	10	7,111
Přerov	66	20	30%	44	67%	2	3%	26,485	28,788	3	10,000
Kostelec na Hané	56	21	38%	33	59%	2	4%	26,857	29,857	2	10,000
Bratislava	86	24	28%	55	64%	7	8%	23,977	27,116	3	10,667
Sečovce	86	15	17%	68	79%	3	3%	22,163	25,302	3	12,667
Prešov	134	102	76%	26	19%	6	4%	30,433	35,328	5	2,000
Bystřice pod Hostýnem	80	19	24%	57	71%	4	5%	24,375	25,900	3	10,667
Hlohovec	30	5	17%	24	80%	1	3%	20,667	22,533	1	15,000
Napajedla	22	3	14%	18	82%	1	5%	20,091	24,000	1	11,000
Třeboň	168	49	29%	113	67%	6	4%	24,333	25,310	7	9,167
Plzeň	168	72	43%	86	51%	10	6%	25,393	27,702	7	7,833
Nové Zámky	86	46	53%	36	42%	4	5%	26,116	29,814	3	7,000
CELKEM	**3228**	**1520**	**47%**	**1520**	**47%**	**188**	**6%**	**25,023**	**27,955**		

Kapitola 2

Teoretická část k použitým testům pro testování hypotéz

2.1 Testy normality dat

Většina běžných statistických metod předpokládá, že data, která jsou zpracovávána, pochází ze základního souboru s normálním rozdělením. V případě, že není normalita potvrzena, je třeba zvýšené opatrnosti při interpretaci výsledků statistických testů. U některých testů je tento předpoklad testů velmi důležitý, u jiných však mírné odchylky od normality ovlivní závěry statistických testů jen nepatrně.

Na porušení normality může upozornit například skutečnost, že medián je vzdálený od hodnot průměrných.

Rychlou představu o normalitě dat mohou dále poskytnou grafické metody. Mezi základní grafické metody patří vykreslení histogramů relativních četností, empirické distribuční funkce, q-q grafy, p-p grafy, box-ploty a další.

Existuje celá řada testů normality, které se liší principem otestování normality. Všechny testy normality jsou založeny na nulové hypotéze H_0 : náhodný výběr pochází z normálního rozdělení $N(\mu, \sigma^2)$ oproti alternativní hypotéze H_1 : náhodný výběr pochází z jiného rozdělení. Některé testy vychází z porovnávání koeficientů šikmosti a špičatosti (např. Jarque-Bera test, Shapiro-Wilk test), jiné pracují s funkcí hustoty (např. χ^2 test dobré shody) nebo s distribuční funkcí (např. Kolmogorov-Smirnov test, Lilliefors test) [5].

V této práci byly z grafických metod využity boxploty, histogramy relativních četností a empirické distribuční funkce a z testů byly využity Jarque-Bera test a Lilliefors test.

2.2 Lineární regrese

Jedním z možných způsobů testování hypotéz je test o parametrech lineárního regresního modelu. Regresní analýza nám umožňuje zkoumat a testovat závislost vysvětlované proměnné (např. počet gólů v jednotlivých zápasech) a vysvětlující proměnné (např. kolo sezony).

Uvažuje následující lineární regresní model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1, 2, 3, ..., n)$$

kde

- $-\infty < \beta_0, \beta_1 < +\infty$ jsou neznámé parametry modelu;
- x_i jsou pozorované hodnoty vysvětlující proměnné;
- y_i jsou pozorované hodnoty vysvětlované proměnné;
- ϵ_i jsou neznáme náhodné odchylky.

Obvykle se pracuje s následujícími předpoklady na náhodné odchylky ϵ_i [2, str. 40]

(P1) střední hodnota odchylek je nulová, $E(\epsilon_i) = 0$ **pro všechna** $i = 1, 2, ..., n$**;**

(P2) rozptyl odchylek je shodný, $D(\epsilon_i) = \sigma^2$ **pro všechna** $i = 1, 2, ..., n$**;**

(P3) veličiny ϵ_i **jsou navzájem nezávislé;**

případně

(P4) složky ϵ_i **jsou nezávislé náhodné veličiny a mají normální rozdělení,** .

Předpoklad **(P4)** zahrnuje všechny předpoklady **(P1)-(P3)**.

Jak je uvedeno např. v [7, str. 97], pokud jsou splněny výše uvedené předpoklady (P1)-(P4), pak odhady parametrů β_0, β_1 získané metodou nejmenších čtverců mají normální rozdělení a lze testovat jejich nenulovost.

Pro ověření kladné lineární závislosti mezi vysvětlující a vysvětlovanou proměnnou lze vycházet z jednostranné hypotézy: [2, str. 60]

$$H_0 : \beta_1 = \beta_1^0 \text{ a } H_1 : \beta_1 > \beta_1^0 \ .$$

Při platnosti předpokladů uvedených výše použijeme testovací statistiku:

$$t_1 = \frac{b_1 - \beta_1^0}{SE(b_1)}$$

kde

- b_1 je odhad parametru β_1 získaný metodou nejmenších čtverců;
- β_1^0 je hodnota testovaného parametru, v našem případě 0;
- $SE(b_1)$ je odhad směrodatné odchylky b_1 definován vztahem:

$$SE(b_1) = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}};$$

- $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2}$ je odhad parametru σ^2.

Hypotézu $H_0 : \beta_1 = 0$ zamítáme na hladině významnosti α při jednostranné alternativě $\beta_1 > 0$, jestliže pro výše uvedenou statistiku t_1 po dosazení $\beta_1^0 = 0$ platí: $t_1 > t_{1-\alpha}(n-2)$, kde $t_{1-\alpha}(n-2)$ je kvantil studentova rozdělení se stupni volnosti $(n-2)$. Zamítneme-li hypotézu H_0 na zvolené hladině významnosti, říkáme, že koeficient b_1 (odhad hodnoty β_1) se významně liší od nuly a je kladný.

2.3 Kontingenční tabulky a χ^2 - test nezávislosti a homogenity

Pro testování nezávislosti dvou náhodných veličin se často používají testy nezávislosti založené na dvourozměrných kontingenčních tabulkách. Nulová hypotéza H_0 je v tomto případě formulována takto: náhodná veličina X a náhodná veličina Y jsou nezávislé. Předpokladem pro vytvoření kontingenčních tabulek je skutečnost, že náhodná veličina X nabývá konečně mnoha hodnot (např. výhry, prohry, remízy) a náhodná veličina Y nabývá také konečně mnoha hodnot (např. 1, 0, -1). V i - tém řádku a j - sloupci tabulky je počet realizací, ve kterých náhodná veličina X nabývá hodnoty i a náhodná veličina Y nabývá hodnoty j. Řádkové součty jsou označeny $n_{i\cdot}$ a sloupcové $n_{\cdot j}$.

Očekávané četnosti jsou označeny $o_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$. Testové kritérium má tvar: $\sum\sum \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$, kde se sčítá přes všechny prvky kontingenční tabulky.

Při platnosti hypotézy H_0 se testové kritérium asymptoticky blíží rozdělení χ^2 se stupni volnosti $\nu = (r-1)(c-1)$, kde r je počet řádků kontingenční tabulky a c je počet sloupců kontingenční tabulky. Obvykle tedy požadujeme, aby očekávané četnosti $o_{ij} \geq 5$. Pokud hodnota testového kritéria je větší než $\chi^2_{\nu=(r-1)(c-1)}(1-\alpha)$, zamítáme hypotézu o nezávislosti veličin X a Y na hladině významnosti α.

Někdy se může stát, že řádkové četnosti $n_{i\cdot}$ jsou předem stanoveny. Pak i - tý řádek tabulky má multinomické rozdělení s parametry $n_{i\cdot}$, $q_{i1}, \ldots, q_{ic}$, kde $q_{i1}, \ldots, q_{ic}$ jsou nějaké pravděpodobnosti splňující podmínku $q_{i1}+, \cdots + q_{ic} = 1$. Většinou je pak třeba testovat *hypotézu homogenity*, která říká, že pravděpodobnosti $q_{i1}, \ldots, q_{ic}$ nezávisejí na řádkovém indexu i, takže všechny řádky matice (q_{ij}) jsou stejné. Jak je uvedeno v [1, str. 282] lze i v tomto případě počítat testové kriterium χ^2 podle stejného vzorce jako u testu nezávislosti. Hypotéza homogenity se zamítá v případě, že $\chi^2 \geq \chi^2_{\nu=(r-1)(c-1)}(1-\alpha)$.

Kapitola 3

Teoretická část k rankingům

Rankingové modely vychází z potřeby určit relativní sílu sledovaného objektu vzhledem k ostatním objektům. Tyto modely mají široké využití, v praxi se s těmito modely setkáváme např. při určování síly sportovních týmů, při určování pořadí internetových odkazů, ve kterém se každému uživateli objeví po zadání klíčových slov do webových vyhledávačů a podobně.

V rámci této práce se rankingové modely využijí pro určení predikování výsledků zápasů za využití určené síly jednotlivých týmů. Pokud je k dispozici informace o aktuální síle obou týmu, které spolu sehrají konkrétní zápas, je možno predikovat výsledek tohoto zápasu. Pro určení síly týmů je nutno vycházet z výsledků všech předchozích zápasů. Pokud by se určovala síla jen ze zápasů, které odehrály dva konkrétní týmy proti sobě, byla by tato síla značně nepřesná, neboť v házené hrají proti sobě konkrétní týmy pouze dvakrát za sezónu.

3.1 Colley Matrix Model

Tento model vytvořil Wesley Colley v roce 2002 a poprvé byl použit pro určení síly týmů amerického fotbalu [3, str. 15]. V modelu jsou využity informace o počtu výher a počtu proher týmu. Předpokládá se, že rozlosování týmů do soutěže je náhodné. Jako n_i je označen počet zápasů odehraných i - tým týmem a w_i je počet vyhraných zápasů tohoto týmu. Potom pravděpodobnost, že tým i v $n_i + 1$ zápase vyhraje je určena podílem $\frac{w_i + 1}{n_i + 2}$. Pokud neexistují žádné další závislosti mezi w_i a n_i je ranking daného týmu určen právě tímto podílem jako $r_i = \frac{w_i + 1}{n_i + 2}$. Tvůrce modelu upravil w_i do tvaru $w_i = \frac{w_i + n_i - l_i}{2} = \frac{w_i - l_i}{2} + \frac{n_i}{2}$, kde l_i je počet proher daného týmu a dále předpokládal, že $\frac{n_i}{2}$ zahrnuje síly protihrajících týmů. Pro zjednodušení lze předpokládat, že tato hodnota je pro všechny týmy na počátku soutěžního ročníku stejná, tedy $\frac{n_i}{2} = \sum_{j=1}^{n_i} \frac{1}{2}$ a v průběhu sezóny se mění podle aktuálního vývoje, tedy $\frac{n_i}{2} = \sum_{j=1}^{n_i} r_j^{(i)}$, kde $r_j^{(i)}$ je síla j - tého protihrajícícho týmu. Po dosazení těchto vztahů do rovnice $r_i = \frac{w_i + 1}{n_i + 2}$ se získá finální vztah pro i - tý tým

$$(n_i + 2)r_i - \sum_{j=1}^{n_i} r_j^{(i)} = 1 + \frac{(w_i - l_i)}{2}.$$

Uvedený vztah platí pro všechny týmy, proto lze získat soustavu n rovnic, kde n je počet týmů. Maticově zapsáno

$$\boldsymbol{C}_{n\times n}\boldsymbol{r} = \boldsymbol{b},$$

kde

$$C_{ij} = \begin{cases} -n_{ji} & pro\ i \neq j, \\ 2 + n_i & pro\ i = j \end{cases}$$

a

$$b_i = 1 + \frac{(w_i - l_i)}{2},$$

kde n_{ji} je počet zápasů mezi i - tým a j - tým týmem. Matici C_{ij} nazýváme Colleyho maticí. Lze ukázat, že vektor $\boldsymbol{r}$, který je řešením rovnice $\boldsymbol{Cr} = \boldsymbol{b}$, vždy existuje, je určen jednoznačně a je nezáporný [3].

Při výpočtu aktuálního rankingu týmu dle tohoto modelu se tedy postupuje v následujících krocích:

1. Sestavení Colleyho matice $\boldsymbol{C}$

$$C_{ij} = \begin{cases} -n_{ji} & pro\ i \neq j, \\ 2 + n_i & pro\ i = j \end{cases}.$$

2. Sestavení vektoru $\boldsymbol{b}$

$$b_i = 1 + \frac{(w_i - l_i)}{2}.$$

3. Vyřešení rovnice $\boldsymbol{Cr} = \boldsymbol{b}$, kde vektor $\boldsymbol{r}$ obsahuje hodnotu rankingové síly každého týmu.

3.2 Keener Ranking Model

Tento model vytvořil James P. Keener v roce 1993 a poprvé byl použit pro odvození rankingů týmů amerického fotbalu. Keener Ranking Model je zobecněním tzv. přímé rankingové metody, která vychází z následujících pravidel. Nechť r_j jsou kladná čísla rankingů vyjadřující sílu j - tého týmu, pak skóre s_i definované vztahem

$$s_i = \frac{1}{n_i}\sum_{j=1}^{n} a_{ij}r_j,$$

kde a_{ij} jsou nezáporná čísla závisející na síle i - tého a j - tého týmu (např. počet vstřelených gólů), n_i je počet zápasů sehraný i - tým týmem a n je počet týmů v soutěži, vyjadřuje aktuální sílu i - tého týmu. Matice $\boldsymbol{A}$ s prvky a_{ij}/n_i se nazývá preferenční matice a lze předpokládat, že skóre s_i je úměrné rankingům týmu, tedy $\boldsymbol{Ar} = \lambda\boldsymbol{r}$, neboli $\boldsymbol{r}$ je vlastním vektorem matice $\boldsymbol{A}$.

V předchozím jednodušším Colley Matrix Modelu byla brána v potaz jen výhra a prohra týmu a nezáleželo na počtu vstřelených a obdržených gólů. Keener Ranking Model je založen na speciálním tvaru preferenční matice, který zohledňuje počet vstřelených a obdržených gólů.

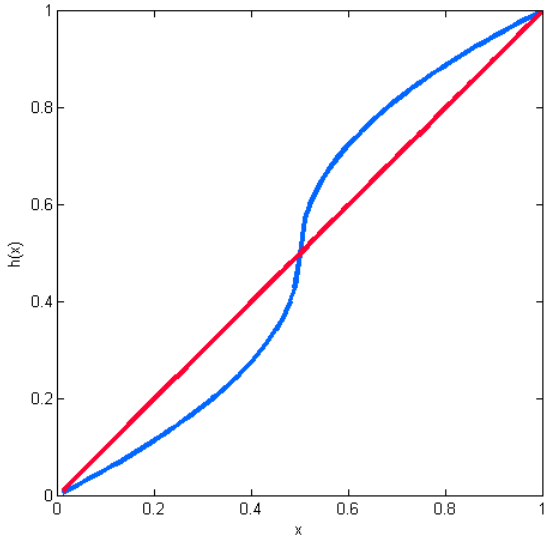
Pokud S_{ij} označuje počet gólů, které i - tý tým vstřelil j - témů týmu, pak prvek preferenční matice Keener Ranking Modelu vychází z poměru $\frac{S_{ij}}{S_{ij}+S_{ji}}$.

Model je založen na Keenerovo matici

$$K_{ij} = \begin{cases} h\left(\frac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right) & \text{pokud tým } T_i \text{ sehrál již sehrál zápas s týmem } T_j \\ 0 & \text{jinak} \end{cases},$$

kde h je „vyhlazovací funkce“ ve tvaru $h(x) = \frac{1}{2} + \frac{1}{2}\text{sgn}\left(x - \frac{1}{2}\right)\sqrt{|2x-1|}$.

Obrázek 3.1: Vyhlazovací funkce $h(x)$



Na Obrázku 3.1 je uveden průběh vyhlazpovací funkce $h(x)$, který ukazuje, že použití této funkce v modelu přidělí větší váhu těsným výhrám než je váha, která by odpovídala použití poměru bez vyhlazovací funkce. V článku [4] je dokázáno, že vlastní vektor $\boldsymbol{r}$ příslušící největšímu vlastnímu číslu matice $\boldsymbol{K}$ je vektor hledaných rankingů.

3.3 Offense-Defense Model (ODM)

Model ODM využívá dvou rankingových vektorů, jednoho vytvořeného jako ranking obrany týmu a jeden jako ranking útoku týmu. [3, str. 25] Pro souhrnný ranking týmu je brán podíl těchto dvou rankingů. Tento model vychází z matice skóre $\boldsymbol{S}$, jejíchž prvky jsou tvořeny S_{ij}, které také vyjadřují počet gólů, které tým T_i vstřelil týmu T_j. Matice S je upravena o malou chybu $0 < \epsilon < 1$ takto

$$P_{ij} = \begin{cases} S_{ij} + \epsilon & \text{pokud tým } T_i \text{ již sehrál zápas s týmem } T_j \\ \epsilon & \text{jinak} \end{cases}.$$

Rankingy obrany a útoku týmu jsou získány iteračním procesem podle vztah;

$$\begin{aligned}
\boldsymbol{d}^{(0)} &= [1, 1, \ldots, 1] \\
\boldsymbol{o}^{(k)} &= \boldsymbol{P}^T \cdot \left[\frac{1}{d_1^{(k-1)}} \frac{1}{d_2^{(k-1)}}, \ldots, \frac{1}{d_n^{(k-1)}} \right] \\
\boldsymbol{d}^{(k)} &= \boldsymbol{P} \cdot \left[\frac{1}{o_1^{(k)}}, \frac{1}{o_2^{(k)}}, \ldots, \frac{1}{o_n^{(k)}} \right]
\end{aligned}$$

Tento iterační proces je ukončen v okamžiku, kdy rankingové vektory $\boldsymbol{d}^{(k)}$ a $\boldsymbol{o}^{(k)}$ se již nemění. Výsledky rankingový vektor týmu je získán jako podíl útočného a obranného rankingu, tedy

$$\boldsymbol{r} = \left[\frac{o_1}{d_1}, \frac{o_2}{d_2}, \ldots, \frac{o_n}{d_n} \right].$$

Při praktických výpočtech se využívají následující nastavení $\epsilon = 0,00001$ a zastavovací podmínka iterace $||\boldsymbol{d}^{(k)} - \boldsymbol{d}^{(k-1)}|| < 0,01$ [3, str. 59].

3.4 Porovnání hodnot rankingů s reálnými výsledky zápasů

Jednou z možností jak využít hodnoty rankingů jednotlivých týmů je odhadování výsledků zápasů. Vzhledem k tomu, že jednotlivé rankingy nejsou normovány, je třeba je převést na pravděpodobnosti. Předpokládá se, že jsou k dispozici hodnoty rankingů r_i pro tým i a hodnoty rankingů r_j pro tým j, pak pravděpodobnost π_{ij}, že tým i zvítězí nad týmem j je dána vztahem $\pi_{ij} = \dfrac{r_i}{r_i + r_j}$.

V této práci bude zvolen následující postup srovnávání rankingů s reálnými výsledky zápasů:

1. na základě rankingů získaných podle jednotlivých modelů budou určeny pravděpodobnosti π_{ij};
2. pokud $\pi_{ij} > 0.5$ je predikováno vítězství i - tého týmu;
3. pokud $\pi_{ij} < 0.5$ je predikováno vítězství j - tého týmu;
4. remízové výsledky zápasů budou predikovány pouze v případě kdy $\pi_{ij} = 0.5$, protože tato situace je velmi neobvyklá, je predikovaných remíz velmi málo;
5. predikované výsledky zápasů budou porovnány s reálnými výsledky zápasů.

V další fázi bude prozkoumáno, zda má na predikování výsledků zápasů vliv hodnota 0.5 stanovená jako hraniční bod.

Kapitola 4

Testování hypotéz na reálných datech

4.1 Testy normality vstřelených gólů

4.1.1 Motivace k vytvoření hypotézy

Jak je uvedeno v teoretické části této práce, je pro celou řadu základních statistických metod a postupů potřeba předpoklad, aby data, na kterých se tyto metody využívají, pocházela z normálního rozdělení. Házená patří mezi kolektivní sporty, ve kterých padá hodně gólů, lze předpokládat, že normalita bude splněna častěji než u sportů, ve kterých padá méně gólů jako například ve fotbale nebo hokeji.

4.1.2 Výběr testu pro testování hypotézy

Z celá řady statistických metod a postupů, které slouží k testování normality, byly vybrány dva testy (Jarque-Bera test a Lilliefors test), každý založený na jiném principu. Jarque-Bera test využívá k testování normality koeficienty špičatosti, Lilliefors test je založen na porovnávání distribučních funkcí.

4.1.3 Formulace hypotézy o normalitě dat

Při testování této hypotézy bylo vycházeno ze základní hypotézy H_0 : *data pochází z normálního rozdělení* oproti alternativní hypotéze H_1 : *data pochází z jiného rozdělení.*

4.1.3.1 Výsledky testu

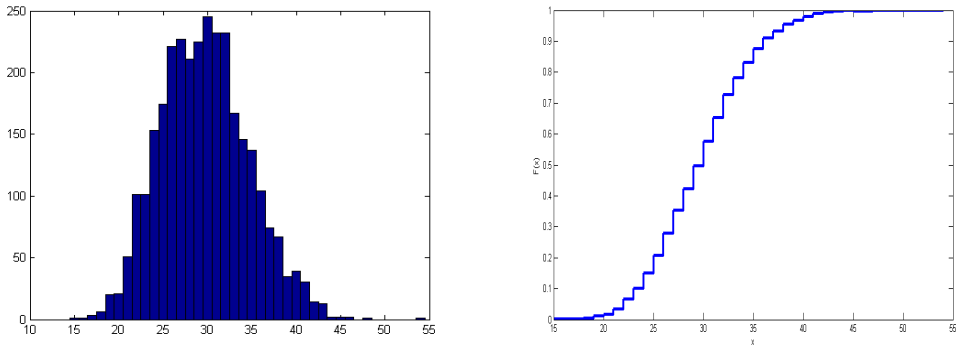
Testy normality byly prováděny samostatně pro góly domácích, góly hostů a také dohromady pro všechny góly, které v zápasech padly.

Vykreslené histogramy relativních četností a distribuční funkce uvedené na Obrázcích 4.1, 4.2, 4.3 naznačují symetričnost dat, ale je možné, že testy zamítnou domněnku o normalitě dat.

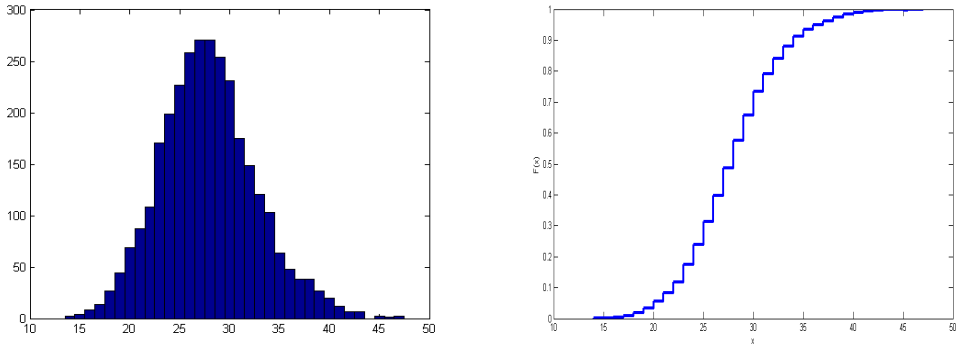
Testována byla normalita jak souhrnně pro všechna data, tak pro jednotlivá soutěžní kola v rámci všech sezón. Všechny testy byly provedeny na hladině významnosti $\alpha = 5\%$. Výsledky testů normality dat jsou uvedeny v Tabulkách 4.1, 4.2, 4.3, 4.4.

Výsledky uvedené v Tabulkách 4.1, 4.2, 4.3, 4.4 potvrzují domněnku, že test založený na koeficientech šikmosti a špičatosti (Jarque-Bera test) zamítl normalitu v menším počtu případů

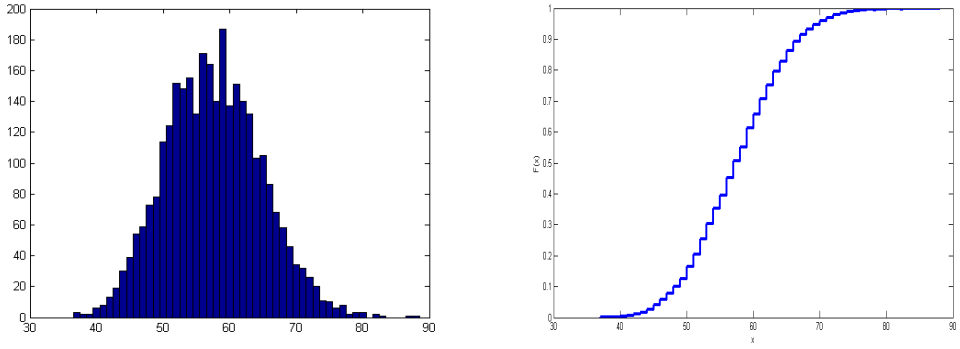
Obrázek 4.1: Histogram a empirická distribuční funkce-góly domácí



Obrázek 4.2: Histogram a empirická distribuční funkce-góly hosté



Obrázek 4.3: Histogram a empirická distribuční funkce-góly celkem



než test založený na distribuční funkci (Lilliefors test). P-hodnoty testů naznačují, že porušení normality není zas tak závažné, v mnoha případech by na hladině významnosti $\alpha = 1\%$ nebyla normalita zamítnuta.

Tabulka 4.1: Výsledky Jarque-Bera testu pro jednotlivé sezóny

sezóny	zápasy	góly domácí		góly hosté		góly celkem	
		p	hypotéza	p	hypotéza	p	hypotéza
2011-2012	306	$0,221$	H_0	$0,500$	H_0	$0,396$	H_0
2010-2011	305	$0,032$	H_1	$0,203$	H_1	$0,171$	H_0
2009-2010	306	$0,070$	H_0	$0,053$	H_1	$0,500$	H_0
2008-2009	306	$0,073$	H_0	$0,500$	H_1	$0,089$	H_0
2007-2008	306	$0,070$	H_0	$0,010$	H_1	$0,006$	H_1
2006-2007	306	$0,500$	H_0	$0,019$	H_1	$0,500$	H_0
2005-2006	306	$0,001$	H_1	$0,035$	H_1	$0,017$	H_1
2004-2005	306	$0,500$	H_0	$0,002$	H_1	$0,092$	H_0
2003-2004	306	$0,010$	H_1	$0,001$	H_1	$0,074$	H_0
2002-2003	306	$0,048$	H_1	$0,073$	H_1	$0,126$	H_0

Tabulka 4.2: Výsledky Lilliefors testu pro jednotlivé sezóny

sezóny	zápasy	góly domácí		góly hosté		góly celkem	
		p	hypotéza	p	hypotéza	p	hypotéza
2011-2012	306	$0,001$	H_1	$0,001$	H_1	$0,048$	H_1
2010-2011	305	$0,001$	H_1	$0,001$	H_1	$0,011$	H_1
2009-2010	306	$0,001$	H_1	$0,001$	H_1	$0,017$	H_1
2008-2009	306	$0,001$	H_1	$0,001$	H_1	$0,009$	H_1
2007-2008	306	$0,001$	H_1	$0,001$	H_1	$0,008$	H_1
2006-2007	306	$0,011$	H_1	$0,001$	H_1	$0,111$	H_0
2005-2006	306	$0,001$	H_1	$0,001$	H_1	$0,003$	H_1
2004-2005	306	$0,025$	H_1	$0,001$	H_1	$0,005$	H_1
2003-2004	306	$0,001$	H_1	$0,001$	H_1	$0,001$	H_1
2002-2003	306	$0,001$	H_1	$0,001$	H_1	$0,001$	H_1

Tabulka 4.3: Výsledky Jarque-Bera testu pro jednotlivá kola

kola	zápasy	góly domácí		góly hosté		góly celkem	
		p	hypotéza	p	hypotéza	p	hypotéza
1	90	$0,224$	H_0	$0,301$	H_0	$0,326$	H_0
2	90	$0,133$	H_0	$0,106$	H_0	$0,500$	H_0
3	90	$0,412$	H_0	$0,500$	H_0	$0,207$	H_0
4	90	$0,151$	H_0	$0,011$	H_1	$0,057$	H_0
5	90	$0,235$	H_0	$0,500$	H_0	$0,453$	H_0
6	90	$0,250$	H_0	$0,027$	H_1	$0,500$	H_0
7	90	$0,500$	H_0	$0,209$	H_0	$0,459$	H_0
8	90	$0,500$	H_0	$0,094$	H_0	$0,120$	H_0
9	90	$0,500$	H_0	$0,500$	H_0	$0,186$	H_0
10	90	$0,057$	H_0	$0,032$	H_1	$0,500$	H_0
11	90	$0,500$	H_0	$0,278$	H_0	$0,500$	H_0
12	90	$0,358$	H_0	$0,248$	H_0	$0,500$	H_0
13	90	$0,135$	H_0	$0,001$	H_1	$0,053$	H_0
14	90	$0,172$	H_0	$0,236$	H_0	$0,059$	H_0
15	90	$0,103$	H_0	$0,221$	H_0	$0,403$	H_0
16	90	$0,001$	H_1	$0,075$	H_0	$0,001$	H_1
17	90	$0,444$	H_0	$0,370$	H_0	$0,034$	H_1
18	89	$0,100$	H_0	$0,384$	H_0	$0,259$	H_0
19	90	$0,500$	H_0	$0,500$	H_0	$0,424$	H_0
20	90	$0,205$	H_0	$0,500$	H_0	$0,500$	H_0
21	90	$0,126$	H_0	$0,046$	H_1	$0,415$	H_0
22	90	$0,056$	H_0	$0,002$	H_1	$0,158$	H_0
23	90	$0,369$	H_0	$0,500$	H_0	$0,208$	H_0
24	90	$0,500$	H_0	$0,500$	H_0	$0,500$	H_0
25	90	$0,182$	H_0	$0,017$	H_1	$0,500$	H_0
26	90	$0,500$	H_0	$0,500$	H_0	$0,500$	H_0
27	90	$0,298$	H_0	$0,295$	H_0	$0,500$	H_0
28	90	$0,001$	H_1	$0,500$	H_0	$0,132$	H_0
29	90	$0,126$	H_0	$0,149$	H_0	$0,500$	H_0
30	90	$0,500$	H_0	$0,222$	H_0	$0,500$	H_0
31	90	$0,466$	H_0	$0,010$	H_1	$0,134$	H_0
32	90	$0,500$	H_0	$0,053$	H_0	$0,382$	H_0
33	90	$0,225$	H_0	$0,094$	H_0	$0,500$	H_0
34	89	$0,290$	H_0	$0,500$	H_0	$0,477$	H_0

Tabulka 4.4: Výsledky Lilliefors testu pro jednotlivá kola

kola	zápasy	góly domácí		góly hosté		góly celkem	
		p	hypotéza	p	hypotéza	p	hypotéza
1	90	$0,102$	H_0	$0,052$	H_0	$0,500$	H_0
2	90	$0,001$	H_1	$0,005$	H_1	$0,083$	H_0
3	90	$0,455$	H_0	$0,172$	H_0	$0,257$	H_0
4	90	$0,295$	H_0	$0,001$	H_1	$0,009$	H_1
5	90	$0,026$	H_1	$0,294$	H_0	$0,101$	H_0
6	90	$0,062$	H_0	$0,045$	H_1	$0,165$	H_0
7	90	$0,500$	H_0	$0,003$	H_1	$0,419$	H_0
8	90	$0,184$	H_0	$0,002$	H_1	$0,007$	H_1
9	90	$0,500$	H_0	$0,168$	H_0	$0,086$	H_0
10	90	$0,019$	H_1	$0,012$	H_1	$0,500$	H_0
11	90	$0,036$	H_1	$0,011$	H_1	$0,350$	H_0
12	90	$0,094$	H_0	$0,032$	H_1	$0,500$	H_0
13	90	$0,039$	H_1	$0,034$	H_1	$0,459$	H_0
14	90	$0,006$	H_1	$0,062$	H_0	$0,212$	H_0
15	90	$0,047$	H_1	$0,076$	H_0	$0,148$	H_0
16	90	$0,046$	H_1	$0,001$	H_1	$0,010$	H_1
17	90	$0,042$	H_1	$0,010$	H_1	$0,285$	H_0
18	89	$0,044$	H_1	$0,064$	H_0	$0,116$	H_0
19	90	$0,040$	H_1	$0,056$	H_0	$0,024$	H_1
20	90	$0,013$	H_1	$0,045$	H_1	$0,175$	H_0
21	90	$0,058$	H_0	$0,013$	H_1	$0,440$	H_0
22	90	$0,009$	H_1	$0,001$	H_1	$0,156$	H_0
23	90	$0,067$	H_0	$0,366$	H_0	$0,272$	H_0
24	90	$0,027$	H_1	$0,263$	H_0	$0,185$	H_0
25	90	$0,026$	H_1	$0,001$	H_1	$0,500$	H_0
26	90	$0,323$	H_0	$0,404$	H_0	$0,105$	H_0
27	90	$0,078$	H_0	$0,308$	H_0	$0,132$	H_0
28	90	$0,001$	H_1	$0,302$	H_0	$0,078$	H_0
29	90	$0,020$	H_1	$0,001$	H_1	$0,500$	H_0
30	90	$0,024$	H_1	$0,096$	H_0	$0,465$	H_0
31	90	$0,199$	H_0	$0,028$	H_1	$0,235$	H_0
32	90	$0,006$	H_1	$0,021$	H_1	$0,259$	H_0
33	90	$0,309$	H_0	$0,021$	H_1	$0,221$	H_0
34	89	$0,149$	H_0	$0,012$	H_1	$0,475$	H_0

4.2 Lineární závislost počtu gólů a počtu kol

4.2.1 Motivace k vytvoření hypotézy

Házená je ve světě chápána jako jeden z nejrychlejších a nejdynamičtějších sportů. V házenkářských zápasech padá hodně gólů a často se mění skóre.

Byla sestavena hypotéza, zda je pravidlem, že v jednotlivých ročnících roste počet gólů s růstem počtu kol. Tato hypotéza je založena na předpokladu, že na začátku soutěžního ročníku jsou týmy připraveny velmi kvalitně na sezónu, mají dostatek hráčů, téměř žádní hráči nejsou zraněni a všechny týmy mají velkou motivaci vyhrát utkání. Ovšem s přibývajícím počtem kol roste počet zraněných hráčů a zápasy se stávají méně vyrovnanými. Také se postupně začíná rýsovat finální pořadí týmů, a proto jak se lidově říká, v některých zápasech na konci sezóny již „nejde téměř o nic“.

4.2.2 Výběr testu pro testování hypotézy

Statistické ověření pravdivosti této hypotézy bylo formulováno na základě předpokladu lineární závislosti počtu gólů v jednotlivých kolech a počtu odehraných kol.

Je uvažován následující lineární regresní model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1, 2, 3, ..., n)$$

kde

- $-\infty < \beta_0,\ \beta_1 < +\infty$ jsou neznámé parametry modelu;
- x_i vyjadřuje kolo, ve kterém byl zápas sehrán;
- y_i je počet všech gólů v zápase v daném kole;
- ϵ_i jsou neznáme náhodné odchylky.

4.2.3 Předpoklady modelu a jejich ověření

Jak je uvedeno v Kapitole 2.2 v regresních modelech se pracuje s několika předpoklady, které byly v rámci regresní diagnostiky ověřeny [8]. Regresní diagnostika byla provedena pro studentizovaná rezidua.

(P1) střední hodnota odchylek je nulová, $E(\epsilon_i) = 0$ pro všechna $i = 1, 2, ..., n$;

Nulovost střední hodnoty byla otestována pomocí jednovýběrového testu střední hodnoty při neznámém parametru σ^2 (t-test hypotézy $H_0 : \mu = 0$ s alternativní hypotézou $H_1 : \mu \neq 0$) [7, str. 55]

(P2) rozptyl odchylek je shodný, $D(\epsilon_i) = \sigma^2$ pro všechna $i = 1, 2, ..., n$;

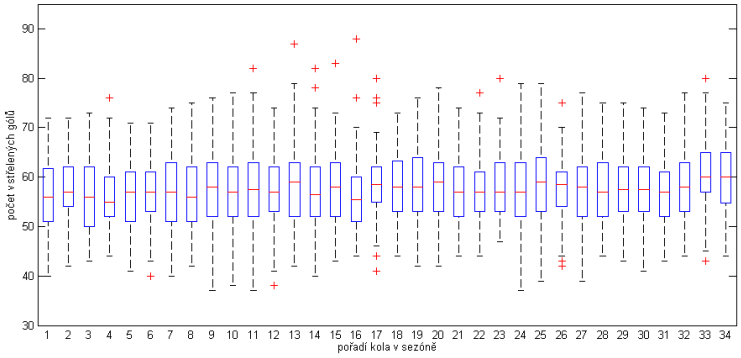
Testy heteroskedasticity zaměřené na dodržení předpokladů konstantnosti rozptylu σ^2 (grafická analýza rozptylů reziduí, Goldfeld-Quandtův test).

Zachycené boxploty pro jednotlivá kola na Obrázku 4.4 nesignalizují žádná silně heteroskedastická data, v žádném z kol ve sledovaném období není variabilita významně větší. Homoskedasticita byla dále ověřena Goldfeld-Quandtovo testem. [7, str. 109]

Tabulka 4.5: Výsledky testu o nulovosti středních hodnot odchylek

testová statistika	kritický obor	p-hodnota	přijatá hypotéza
$1.854 \cdot 10^{-4}$	$(-\infty, -1,961) \cup (1,961, \infty)$	$0,999$	H_0

Obrázek 4.4: Boxplot graf pro jednotlivá kola



Tabulka 4.6: Výsledky testu o shodnosti rozptylů odchylek

rozptyl v 1.-10. kole	$49,827$
rozptyl v 25.-34. kole	$50,658$
testovací statistika F	$0,984$
obor kritických hodnot	$(0,863; 1,121)$
p-hodnota testu	$0,804$

Výsledky testu v Tabulce 4.6 potvrzují shodnost rozptylů v počtu vstřelených gólů na začátku sezóny (prvních 10 kol) a na konci sezóny (posledních 10 kol).

(P3) veličiny ϵ_i jsou nezávislé;

Tento předpoklad nebyl statisticky testován, ale nezávislost jednotlivých měření je předpokládána z charakteru dat.

(P4) složky ϵ_i mají normální rozdělení $\epsilon_i \sim N(0, \sigma^2)$.

Ověření testu normality odchylek bylo ověřeno pomocí Jarque-Bera testu a Lillieforsova testu [2, str. 57] a v rámci standardně používaných postupů regresní diagnostiky byly testy provedeny se studentizovanými rezidui. Výsledky obou testů jsou shrnuty v Tabulce 4.7.

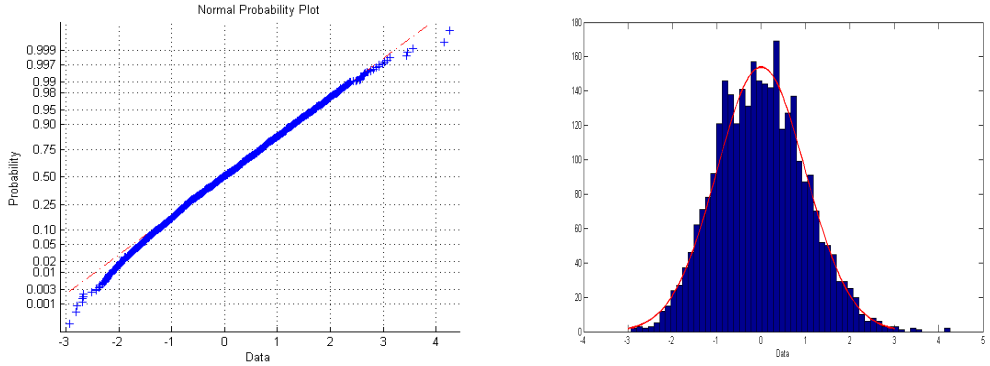
Z Tabulky 4.7 je vidět, že výsledky obou testů zamítají normalitu odchylek dat. Grafická prezentace dat, kde na Obrázku 4.5 je vykreslený histogram relativních četností odchylek

Tabulka 4.7: Výsledky testů normality dat

	testová statistika	kritický obor	p-hodnota	přijatá hypotéza
JB test	$20,962$	$(5,971;\infty)$	$1 \cdot 10^{-3}$	H_1
Lillenfors test	$0,023$	$(0,016;\infty)$	$1 \cdot 10^{-3}$	H_1

dat a p-p graf, však naznačuje, že odchylky dat mají rozdělení velmi blízké rozdělení normálnímu. Vzhledem k velkému rozsahu dat je možno předpokládat asymptotickou normalitu odhadů metodou nejmenších čtverců a dále pracovat i bez splnění tohoto explicitního předpokladu [2, str. 53].

Obrázek 4.5: Histogram relativních četností odchylek a p-p graf



4.2.4 Formulace hypotézy o kladnosti směrnicového parametru

Pro ověření kladné lineární závislosti mezi kolem v sezóně a počtem gólů se vychází z následujících hypotéz:

$$H_0 : \beta_1 = 0 \text{ a } H_1 : \beta_1 > 0 \ .$$

Při platnosti předpokladů uvedených výše byla použita testovací statistika: $t_1 = \dfrac{b_1}{SE(b_1)}$, kterou porovnáváme s kritickým oborem $(t_{1-\alpha}(n-2); +\infty)$.

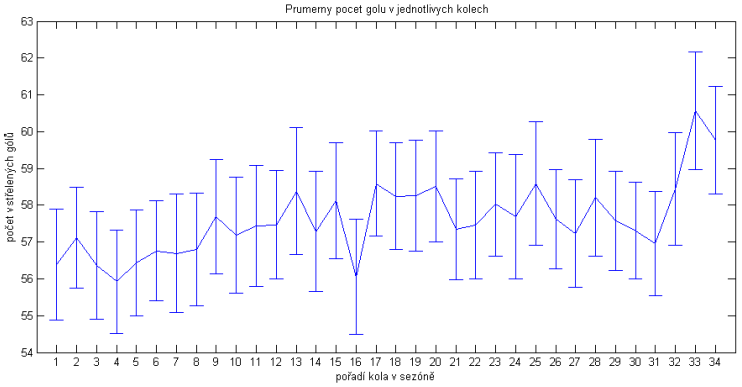
4.2.5 Výsledky testu

Test hypotézy byl proveden v matematickém prostředí MATLAB pomocí funkce *regstats*. Následně byla provedena regresní diagnostika reziduí $e_i = y_i - \widehat{y}$, která slouží k ověření předpokladů (P1)-(P4) a dále byla zaměřena na následující problémy:

- test významnosti odhadnutých koeficientů (t-test) β_0, β_1;
- test významnosti regrese založený na koeficientu determinace R^2 (F-test);
- test detekce odlehlých pozorování (analýza studentizovaných reziduí, Cookova míra vlivu i-tého bodu na výslednou regresi).

Výše uvedený testovací postup byl použit pro reálná data české a německé ligy a zde jsou uvedeny podrobné výsledky zpracování pro německou mužskou nejvyšší soutěž v házené (tzv. bundesliga) pro celé sledované období, tj. sezony od roku 2002 do roku 2011. Dále byla provedena podrobná analýza pro jednotlivé soutěžní sezóny. Následující graf na Obrázku 4.6 zachycuje průměrný počet vstřelených gólů (průměr je počítán ze všech odehraných zápasů v i-tém kole) a příslušné kvartilové rozpětí, který byl motivací pro formulování výše uvedené hypotézy.

Obrázek 4.6: Průměrný počet gólů v jednotlivých kolech a příslušné kvartilové rozpětí



Metodou nejmenších čtverců byly odhadnuty parametry lineárního regresního modelu a byl získán výsledný model ve tvaru:

$$\widehat{y_i} = 56,485 + 0,064 x_i$$

kde $\widehat{y_i}$ jsou odhadnuté průměrné počty gólů v jednotlivých kolech a $x_i = 1, 2, \ldots, 34$ jsou jednotlivá kola.

Výsledky ověření kladné lineární závislosti jsou uvedeny v Tabulce 4.8

Tabulka 4.8: Výsledky testu ověření kladné lineární závislosti

b_1	$0,064$
$SE(b_1)$	$0,013$
t_1	$4,824$
obor kritických hodnot W	$(1,645; \infty)$
p-hodnota	$3,265 \cdot 10^{-6}$

Protože hodnota testové statistiky t_1 leží v oboru kritických hodnot, je na hladině významnosti $\alpha = 5\%$ zamítnuta hypotéza H_0. Proto lze usuzovat, že původně formulovaná hypotéza o tom, že s přibývajícím kolem se zvyšuje počet vstřelených gólů, je platná.

4.2.6 Výsledky regresní analýzy a diagnostiky:

4.2.6.1 Test na významnosti odhadnutých koeficientů (t-test) β_0, β_1

Výsledky testování významnosti odhadnutých koeficientů jsou shrnuty v Tabulce 4.9.

Tabulka 4.9: Výsledky testu o významnosti odhadnutých koeficientů

	b	$SE(b)$	d_b	h_b	t	p
β_0	$56,485$	$0,265$	$55,965$	$57,005$	$212,981$	0
β_1	$0,064$	$0,013$	$0,038$	$0,090$	$4,824$	$1,477 \cdot 10^{-6}$

Statistické vyhodnocení odhadnutých parametrů ukazuje, že oba koeficienty β_0 i β_1 lze považovat za statisticky významně odlišné od 0. To ukazují jak p-hodnoty příslušných statistik, které jsou menší než $\alpha = 5\%$ tak i intervaly spolehlivosti na hladině významnosti $\alpha = 5\%$ pro koeficienty, $\beta_0 \in (55,965; 57,005)$ a $\beta_1 \in (0,013; 0,090)$.

4.2.6.2 Test na významnost regrese založený na koeficientu determinace R^2 (F-test)

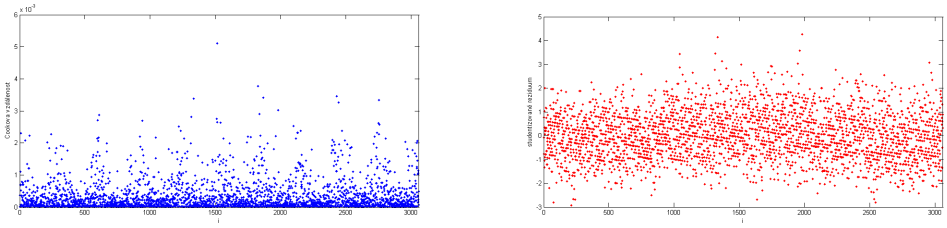
Výsledky testování významnosti regrese jsou shrnuty v Tabulce 4.10.

Přestože koeficient determinace R^2 i jeho upravená varianta R^2_{adj} je velmi malá a model proto nelze použít pro předpovídání a vysvětlení počtu gólů, které padnou v jednotlivých zápasech, p-hodnota F-testu ukazuje, že model jako celek lze považovat za významný.

Tabulka 4.10: Výsledky testu významnosti regrese

R^2	$0,008$
R^2_{adj}	$0,007$
F	$23,270$
p	$1,477 \cdot 10^{-6}$

Obrázek 4.7: Studentizovaná rezidua a Cookova míra



4.2.6.3 Test na detekci odlehlých pozorování (analýza studentizovaných reziduí, Cookova míra vlivu i-tého bodu na výslednou regresy)

Z Obrázku 4.7 je vidět, že hodnoty studentizovaných reziduí i Cookovy vzdálenosti jsou pro některé zápasy vysoké, přesto se neprokázalo, že se jedná o statisticky významná odlehlá pozorování. V Tabulce 4.11 je uvedeno 5 zápasů s nejvyšší hodnotou Cookovy vzdálenosti.

Tabulka 4.11: Zápasy s největší Cookovo vzdáleností

Domácí	Hosté	sezóna	kolo	gólyD	gólyH	Cook
TV Großwallstadt	VfL Gummersbach	2007/2008	33	42	42	$0,005$
SG Flensburg-Handewitt	THW Kiel	2006/2007	33	41	36	$0,004$
SG Wallau-Massenheim	VfL Gummersbach	2004/2005	32	32	45	$0,003$
THW Kiel	FA Göppingen	2005/2006	4	43	33	$0,003$
MT Melsungen	SG Flensburg-Handewitt	2007/2008	13	40	47	$0,003$

Z Tabulky 4.11 je vidět, že zápasy s největší Cookovou vzdáleností se odehrály nejvíce na konci soutěžního ročníku.

Dále byla provedena analýza platnosti hypotézy o kladnosti směrnicového koeficientu také pro jednotlivé soutěžní týmy a jednotlivé soutěžní ročníky. Výsledky jsou uvedeny v Tabulkách 4.12 a 4.13. Z těchto tabulek je vidět, že uvedenou hypotézu H_1 lze přijmout pro 9 týmů a 3 soutěžní ročníky, které jsou v tabulce zvýrazněny červenou barvou.

Tabulka 4.12: Výsledky testů pro jednotlivé týmy

tým	zápasy	β_0	β_1	$SE(b_0)$	$SE(b_1)$	p (b_0)	p-hodnota (b_1)	R^2	R^2_{adj}	p
Bergischer HC	34	60,824	-0,185	2,143	0,107	$2,921 \cdot 10^{-24}$	0,093	0,086	0,057	0,093
Concordia Delitzsch	34	52,829	0,237	2,138	0,107	$2,063 \cdot 10^{-22}$	0,034	0,134	0,106	0,034
DHC Rheinland	34	52,481	0,117	2,226	0,111	$8,588 \cdot 10^{-22}$	0,299	0,034	0,003	0,299
Eintracht Hildesheim	68	57,353	0,052	1,818	0,091	$1,580 \cdot 10^{-41}$	0,567	0,005	-0,010	0,567
FA Göppingen	340	55,488	0,040	0,760	0,038	$5,816 \cdot 10^{-209}$	0,287	0,003	0,000	0,287
Füchse Berlin	170	55,216	0,092	1,059	0,053	$1,181 \cdot 10^{-105}$	0,083	0,018	0,012	0,083
GWD Minden	272	54,390	0,095	0,876	0,044	$7,701 \cdot 10^{-162}$	0,030	0,017	0,014	0,030
HBW Balingen-Weilstetten	204	55,392	0,029	0,939	0,047	$2,861 \cdot 10^{-129}$	0,538	0,002	-0,003	0,538
HSG Ahlen-Hamm	34	56,727	0,054	2,093	0,104	$1,215 \cdot 10^{-23}$	0,607	0,008	-0,023	0,607
HSG D/M Wetzlar	272	55,542	0,007	0,806	0,040	$2,611 \cdot 10^{-173}$	0,869	0,000	-0,004	0,869
HSG Düsseldorf	136	56,485	-0,001	1,170	0,058	$1,257 \cdot 10^{-86}$	0,983	0,000	-0,007	0,983
HSG Nordhorn	238	57,387	0,111	0,882	0,044	$1,047 \cdot 10^{-152}$	0,012	0,026	0,022	0,012
HSG Wetzlar	67	54,488	-0,064	1,696	0,086	$1,295 \cdot 10^{-41}$	0,460	0,008	-0,007	0,460
HSV Hamburg	340	55,849	0,100	0,777	0,039	$8,380 \cdot 10^{-207}$	0,010	0,019	0,016	0,010
MT Melsungen	237	59,438	0,049	1,033	0,052	$1,644 \cdot 10^{-140}$	0,345	0,004	0,000	0,345
Rhein-Neckar-Löwen	170	58,611	0,085	0,978	0,049	$2,674 \cdot 10^{-115}$	0,084	0,018	0,012	0,084
SC Magdeburg	340	57,916	0,043	0,794	0,040	$6,989 \cdot 10^{-209}$	0,276	0,004	0,001	0,276
SG Flensburg-Handewitt	340	58,719	0,061	0,794	0,040	$7,554 \cdot 10^{-211}$	0,124	0,007	0,004	0,124
SG Kronau/Östringen	102	55,916	0,046	1,549	0,077	$3,618 \cdot 10^{-59}$	0,551	0,004	-0,006	0,551
SG Wallau-Massenheim	102	58,494	0,069	1,621	0,081	$3,694 \cdot 10^{-59}$	0,393	0,007	-0,003	0,393
SG Willstätt/Schutterwald	34	52,882	0,311	2,421	0,121	$8,570 \cdot 10^{-21}$	0,015	0,172	0,146	0,015
Stralsunder HV	68	56,176	-0,079	1,963	0,098	$6,393 \cdot 10^{-39}$	0,422	0,010	-0,005	0,422
SV Post Schwerin	34	52,091	0,380	1,932	0,096	$1,419 \cdot 10^{-23}$	0,000	0,327	0,306	0,000
TBV Lemgo	340	57,798	0,073	0,744	0,037	$1,552 \cdot 10^{-217}$	0,051	0,011	0,008	0,051
ThSV Eisenach	68	54,166	0,017	1,715	0,085	$1,444 \cdot 10^{-41}$	0,839	0,001	-0,015	0,839
THW Kiel	340	58,653	0,097	0,848	0,042	$1,491 \cdot 10^{-201}$	0,022	0,015	0,013	0,022
TSG Ludwigshafen-Friesenheim	34	55,968	0,198	2,428	0,121	$1,685 \cdot 10^{-21}$	0,111	0,078	0,049	0,111
TSV Dormagen	68	55,005	0,090	1,438	0,072	$8,924 \cdot 10^{-47}$	0,216	0,023	0,008	0,216
TSV Hannover-Burgdorf	102	57,624	-0,024	1,284	0,064	$4,617 \cdot 10^{-68}$	0,710	0,001	-0,009	0,710
TuS N-Lübbecke	272	56,442	0,073	0,862	0,043	$9,457 \cdot 10^{-24}$	0,089	0,011	0,007	0,089
TUSEM Essen	170	55,015	0,095	0,972	0,048	$2,439 \cdot 10^{-111}$	0,051	0,023	0,017	0,051
TV 05/07 Hüttenberg	34	54,636	0,056	2,199	0,110	$1,728 \cdot 10^{-22}$	0,612	0,008	-0,023	0,612
TV Großwallstadt	340	54,240	0,034	0,726	0,036	$3,357 \cdot 10^{-212}$	0,350	0,003	0,000	0,350
VfL Gummersbach	340	57,361	0,094	0,778	0,039	$2,575 \cdot 10^{-210}$	0,016	0,017	0,014	0,016
VfL Pfullingen	136	52,786	0,172	1,324	0,066	$3,294 \cdot 10^{-76}$	0,010	0,048	0,041	0,010
Wilhelmshavener HV	204	56,033	-0,031	1,050	0,052	$4,772 \cdot 10^{-121}$	0,558	0,002	-0,003	0,558

Tabulka 4.13: Výsledky testů pro jednotlivé soutěžní ročníky

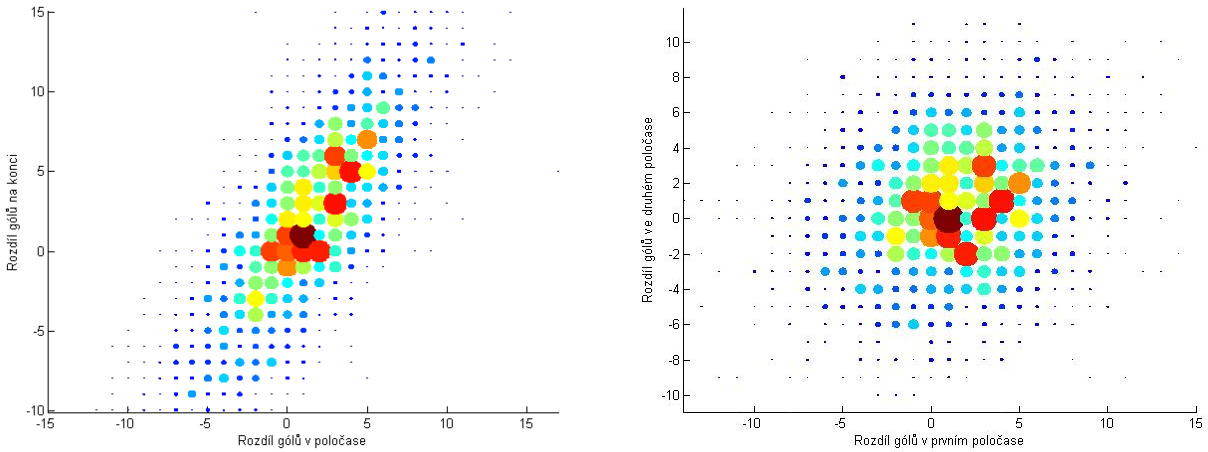
sezóna	zápasy	β_0	β_1	$SE(b_0)$	$SE(b_1)$	p-hodnota (b_0)	p-hodnota (b_1)	R^2	R^2_{adj}	p-hodnota
2011-2012	306	56,955	-0,019	0,779	0,039	$5,211 \cdot 10^{-195}$	0,619	0,001	-0,002	0,619
2010-2011	305	55,879	0,053	0,780	0,039	$4,969 \cdot 10^{-192}$	0,175	0,006	0,003	0,175
2009-2010	306	55,738	0,051	0,705	0,035	$9,834 \cdot 10^{-205}$	0,149	0,007	0,004	0,149
2008-2009	306	59,388	-0,004	0,808	0,040	$1,211 \cdot 10^{-195}$	0,923	0,000	-0,003	0,923
2007-2008	306	56,837	0,124	0,852	0,042	$1,415 \cdot 10^{-183}$	0,004	0,027	0,024	0,004
2006-2007	306	57,843	0,086	0,884	0,044	$4,060 \cdot 10^{-181}$	0,053	0,012	0,009	0,053
2005-2006	306	59,443	-0,054	0,860	0,043	$6,694 \cdot 10^{-188}$	0,214	0,005	0,002	0,214
2004-2005	306	54,578	0,185	0,835	0,042	$5,219 \cdot 10^{-181}$	0,000	0,061	0,058	0,000
2003-2004	306	54,913	0,048	0,860	0,043	$3,302 \cdot 10^{-178}$	0,260	0,004	0,001	0,260
2002-2003	306	53,251	0,169	0,849	0,042	$5,207 \cdot 10^{-176}$	0,000	0,050	0,047	0,000

4.3 Závislost konečného výsledku zápasu na poločasovém skóre

4.3.1 Motivace k vytvoření hypotézy

Inspirací k vytvoření této hypotézy bylo zamyšlení nad tím, zda výsledky v prvním a v druhém poločase spolu významně souvisí. Ve sportovní terminologii se často objevuje situace, kdy po přestávce mezi jednotlivými částmi zápasů vstupuje do zápasu „jiný tým“. Tato změna se projevuje například tím, že mužstvo do další části zápasu vstoupí větším zápalem do hry a cílem otočit nepříznivý vývoj zápasu. Lze předpokládat, že zvláště u zápasů, které jsou po prvním poločase vyrovnané, budou týmy v druhém poločase hrát aktivněji. Po tomto zamyšlení byly vytvořeny grafy uvedené na Obrázku 4.8, zachycující vztah mezi rozdíly počtu gólů v prvním poločase a na konci zápasu/ve druhém poločase. Velikost bublin zachycuje počet zápasů s konkrétní kombinací rozdílu počtu gólů v poločase a na konci/ve druhém poločase zápasu. Pro větší přehlednost jsou stejně velké bubliny označeny stejnou barvou.

Obrázek 4.8: Vztah rozdílů gólů v prvním poločase a na konci zápasu/ve druhém poločase



Předpokládá se silná vazba mezi rozdílem gólů v poločase a rozdílem gólů na konci zápasu, které byla potvrzena také vypočtenými korelačními koeficienty, které jsou uvedeny v Tabulce 4.14. Analýzou dat bylo zjištěno, že pokud je rozdíl v poločase dva a více gólů, pak je korelace mezi rozdílem gólů v poločase a na konci zápasu velmi vysoká a naopak pokud je rozdíl v poločase malý (rozdíl je maximálně jeden gól) je korelace statisticky neprokázána. Lze tedy předpokládat, že pokud je rozdíl v poločase malý nebo je remízový stav, konečný výsledek zápasu je stále otevřený.

Přestože jsou sledované veličiny implicitně závislé, byla závislost podrobně otestována a analyzována.

Byla sestavena hypotéza, zda konečný výsledek zápasu závisí na rozdílu gólů v poločase. Základním předpokladem této hypotézy je fakt, že tým, který v poločase vyhrává je vítězem i na konci zápasu. Z Tabulky 4.14 jsou patrné vysoké hodnoty korelačních koeficientů pro všechny zápasy, proto testování bylo zaměřeno pouze na otevřené zápasy. Jako otevřené zápasy byly

považovány zápasy, ve kterých rozdíl v poločase činil maximálně jeden gól, jejichž hodnoty korelačních koeficientů jsou uvedeny v Tabulce 4.15, ze které jsou patrné jejich nižší hodnoty.

Také byla sestavena hypotéza, zda výsledek druhého poločasu závisí na rozdílu počtu gólů v poločase. Cílem této hypotézy bylo analyzovat skutečnost, zda minimální výhra v poločase o jeden gól motivuje více družstvo, které v poločase vyhrává nebo prohrává.

Tabulka 4.14: Hodnoty korelačních koeficientů pro jednotlivé sezóny

sezona	zápasy	Pearson	Kendall	Spearman
2011-2012	306	$0,778$	$0,601$	$0,764$
2010-2011	305	$0,797$	$0,626$	$0,790$
2009-2010	306	$0,783$	$0,608$	$0,771$
2008-2009	306	$0,822$	$0,639$	$0,801$
2007-2008	306	$0,812$	$0,638$	$0,800$
2006-2007	306	$0,777$	$0,597$	$0,754$
2005-2006	306	$0,750$	$0,554$	$0,713$
2004-2005	306	$0,764$	$0,594$	$0,752$
2003-2004	306	$0,777$	$0,596$	$0,756$
2002-2003	306	$0,735$	$0,557$	$0,718$
celkem	3059	$0,780$	$0,600$	$0,762$

Tabulka 4.15: Hodnoty korelačních koeficientů pro jednotlivé sezóny, v zápasech, kdy je v poločase rozdíl maximálně jeden gól

sezona	zápasy	Pearson	Kendall	Spearman
2011-2012	88	$0,208$	$0,168$	$0,214$
2010-2011	85	$0,166$	$0,181$	$0,231$
2009-2010	94	$0,169$	$0,142$	$0,179$
2008-2009	75	$0,369$	$0,267$	$0,336$
2007-2008	78	$0,034$	$0,059$	$0,067$
2006-2007	81	$0,153$	$0,122$	$0,150$
2005-2006	93	$0,220$	$0,181$	$0,226$
2004-2005	76	$0,205$	$0,165$	$0,204$
2003-2004	76	$0,061$	$0,059$	$0,078$
2002-2003	94	$0,022$	$0,040$	$0,052$
celkem	841	$0,149$	$0,129$	$2,650 \cdot 10^{-6}$

4.3.2 Výběr testů pro testování hypotézy

Pro testování hypotézy o závislosti konečného výsledku byly vytvořeny kontingenční tabulky zachycující počty zápasů, ve kterých jsou zápasy členěny ze dvou hledisek: podle rozdílu počtu gólů v poločase a podle stavu zápasu (výhra, prohra, remíza). Dále bylo pomocí χ^2 - testu nezávislosti ověřováno zda konečný výsledek zápasu, případně výsledek druhého poločasu, závisí nebo nezávisí na stavu zápasu v poločase.

4.3.3 Formulace hypotézy o závislosti konečného výsledku zápasu na rozdílů počtu gólů v poločase a předpoklady modelu

Předpokládejme, že konečný výsledek zápasu je diskrétní náhodná veličina X, která nabývá hodnot Prohra, Výhra, Remíza a Y je diskrétní náhodná veličina vyjadřující rozdíl gólů v poločase zápasu, která nabývá hodnot 1, 0, -1. Testujeme hypotézu H_0 : X *a* Y jsou nezávislé proti alternativní hypotéze H_1 : X *a* Y nejsou nezávislé.

Kritickou hodnotu tohoto testu porovnáváme s kvantily χ^2 - rozdělení. Tento postup je však založen na limitním chování kritéria. Ke shodě s limitním rozdělení se vyžaduje, aby všechny teoretické četnosti byly větší než 5. Tento předpoklad byl v našem případě splněn, proto nebylo třeba slučovat žádné řádky ani sloupce dohromady.

4.3.4 Výsledky testů

Nejprve byl proveden test závislosti konečného výsledku a rozdílu počtu gólů v poločase. V Tabulkách 4.16 a 4.17 je použito následující značení: V (na konci) znamená výhra týmu na konci zápasu, R (na konci) znamená remízový stav na konci zápasu a P (na konci) znamená prohra týmu na konci zápasu.

Tabulka 4.16: Tabulka naměřených četností

	V (na konci)	R (na konci)	P (na konci)	celkem
-1	100	34	111	245
0	159	34	102	295
1	185	35	81	301
celkem	444	103	294	841

Za pomoci Tabulek 4.16 a 4.17 byla určena hodnota testového kritéria $\chi^2 = 24,854$. Jelikož je tato hodnota větší než kritická hodnota $\chi^2_4(\alpha = 0,05) = 9,490$, byla zamítnuta hypotéza H_0, že výsledek na konci zápasu a minimální rozdíly počtu gólů v poločase jsou nezávislé veličiny.

V druhé fázi byl postup zopakován pro testování závislosti výsledku druhého poločasu a rozdílu počtu gólů v poločase. V Tabulkách 4.18 a 4.19 je použito následující značení: V (2.poločas) znamená výhra týmu ve druhém poločase zápasu, R (2.poločas) znamená remízový stav ve druhém poločase zápasu a P (2.poločas) znamená prohra týmu ve druhém poločase zápasu.

V Tabulkách 4.18 a 4.19 jsou opět zachyceny naměřené a očekávané četnosti pro jednotlivé studované stavy. Porovnání hodnoty testového kritéria $\chi^2 = 5,233$ a kritické hodnoty testu

Tabulka 4.17: Tabulka očekávaných četností pro vyrovnané zápasy

	V (na konci)	R (na konci)	P (na konci)	celkem
-1	129,346	30,006	85,648	245
0	155,743	36,130	103,127	295
1	158,911	36,864	105,225	301
celkem	444	103	294	

Tabulka 4.18: Tabulka naměřených četností

	V (2. poločas)	R (2. poločas)	P (2. poločas)	celkem
-1	134	22	89	245
0	159	34	102	295
1	143	42	116	301
celkem	436	98	307	841

Tabulka 4.19: Tabulka očekávaných četností pro vyrovnané zápasy

	V (2. poločas)	R (2. poločas)	P (2. poločas)	celkem
-1	127,015	28,549	89,435	245
0	152,937	34,376	107,687	295
1	156,048	35,075	109,878	301
celkem	436	98	307	

$\chi^2_4(\alpha = 0,05) = 9,490$ byla přijata hypotéza H_0, že výsledek druhého poločasu je nezávislý na rozdílu počtu gólů po prvním poločase. To posiluje domněnku, že zápasy, které v poločase končí malým rozdílem počtu gólů, jsou stále otevřené a ve sportovní terminologii lze říci, že se začíná hrát znovu od začátku.

4.4 Srovnání dynamičnosti hry v české a německé lize

4.4.1 Motivace k vytvoření hypotézy

V Německu patří házená k nejpopulárnějším kolektivním sportům, velký zájem o ni projevují nejen diváci, ale také sponzoři. Německá bundesliga je řazena mezi nejlepší házenkářské ligy světa a v zápasech je vidět velká taktická vyzrálost celého týmu, což je největší odlišnost od české extraligy. Jednou z možností, kde se tato odlišnost může projevit, je větší nasazení týmů a tím i častější změna stavu zápasu mezi prvním poločasem a konečným výsledkem zápasu. Pro tuto hypotézu byly využity opět zápasy, které mají v poločase otevřené skóre (rozdíl maximálně jeden gól), ve kterých by se tato skutečnost mohla nejvíce projevit.

4.4.1.1 Výběr testu pro testování hypotézy

Zápasy byly rozděleny do devíti skupin podle poločasového výsledku a výsledku na konci zápasu. Skupin V-V znamená výhra týmu v poločase nejvýše o jeden gól a výhra týmu na konci zápasu, skupina P-V znamená prohru týmu nejvýše o jeden gól v poločase a výhru týmu na konci zápasu, atd. Pro testování srovnání dynamičnosti německé a české ligy byl vybrán χ^2 - test homogenity, který porovnával četnosti zápasů v jednotlivých skupinách v české a německé lize.

4.4.2 Formulace hypotézy

Jako nulová hypotéza je uvažována hypotéza H_0 : *rozdělení do jednotlivých skupin je homogenní v české a německé lize* oproti alternativní hypotéze H_1 : *rozdělení do jednotlivých skupin není homogenní v české a německé lize.*

4.4.3 Výsledky testu

V Tabulkách 4.20 a 4.21 jsou uvedeny naměřené a očekávané četnosti pro jednotlivé skupiny zápasů. Z těchto tabulek byla určena hodnota testového kritéria $\chi^2 = 8,478$ a tato hodnota byla porovnána s kritickou hodnotou testu $\chi^2_8(\alpha = 5\%) = 15,500$. Protože $\chi^2 = 8,478 < 15,500$, nebyla zamítnuta hypotéza H_0, že rozložení do jednotlivých skupin v německé a české lize je shodné.

Tabulka 4.20: Hodnoty naměřených četností pro jednotlivé skupiny

	V-V	V-R	V-P	R-V	R-R	R-P	P-V	P-R	P-P	celkem
Německo	185	35	81	159	34	102	100	34	111	841
ČR	90	8	24	62	11	38	42	18	53	346
celkem	275	43	105	221	45	140	142	52	164	1187

Tabulka 4.21: Hodnoty očekávaných četností pro jednotlivé skupiny

	V-V	V-R	V-P	R-V	R-R	R-P	P-V	P-R	P-P	celkem
Německo	194,840	30,466	74,393	156,580	31,883	99,191	100,608	36,842	116,195	841
ČR	80,160	12,534	30,607	64,420	13,117	40,809	41,392	15,158	47,805	346
celkem	275	43	105	221	45	140	142	52	164	

Při dalším testování byly zápasy rozděleny podle toho, zda na konci zápasu byl výsledek stejný nebo rozdílný oproti poločasového stavu (např. zda se týmu, který vyhrával v poločase, podařilo tuto výhru udržet i na konci zápasu). Zápasy s remízovým stavem v poločase byly pro tuto analýzu vynechány. Naměřené a očekávané četnosti jednotlivých skupin zápasů jsou uvedeny v Tabulkách 4.22 a 4.23.

Z těchto tabulek byla určena hodnota testového kritéria $\chi^2 = 2,942$ a tato hodnota byla porovnána s kritickou hodnotou testu $\chi^2_1(\alpha = 5\%) = 3,840$. Protože $\chi^2 = 2,942 < 3,840$, nemůže

Tabulka 4.22: Hodnoty naměřených četností pro jednotlivé skupiny

	zachování poločasového stavu	změna poločasového stavu	celkem
Německo	296	250	546
ČR	143	92	235
celkem	439	342	781

Tabulka 4.23: Hodnoty očekávaných četností pro jednotlivé skupiny

	zachování poločasového stavu	změna poločasového stavu	celkem
Německo	306,907	239,093	546
ČR	132,093	102,907	235
celkem	439	342	

být zamítnuta hypotéza H_0, že rozložení do jednotlivých skupin dle zachování poločasového stavu na konci zápasu je v české a německé lize shodné.

Na základě výsledků testů nemůže být potvrzena domněnka, že pravděpodobnosti jednotlivých skupin zápasů členěných podle výsledku v prvním poločase a ve druhém poločase/na konci zápasu je v české a německé soutěži odlišný.

Kapitola 5

Rankingy na reálných datech

Pomocí tří modelů, které jsou uvedeny v teoretické části této práce, byly vypočteny rankingy jednotlivých týmů v každé sezóně a také za celé sledované období pro německou soutěž.

Výraznou nevýhodou těchto modelů je to, že nezohledňují časový faktor jednotlivých zápasů, tedy výsledek zápasu z předcházejícího týdne má stejnou váhu jako například výsledek zápasu odehraného před půl rokem. Modely proto zachycují spíše dlouhodobou sílu týmu, nikoliv aktuální. Lze tedy očekávat, že predikce výsledků založená na hodnotách rankingů počítaných pro celé sledované období bude méně spolehlivá než predikce vycházející z rankingů sezonních. Sezonní rankingy by měly více odpovídat aktuální síle týmů.

Pro ilustraci výsledků modelu budou využity rankingy pro tři nejúspěšnější týmy podle průměrného pořadí ve sledovaném období a které byly nasazeny ve všech soutěžních ročnících. Konkrétně se bude jednat o týmy THW Kiel, SG Flensburg-Handewitt a HSV Hamburg.

Výsledné hodnoty rankingů byly využity pro předpověď výsledku zápasů a porovnání predikovaných výsledků s reálnými výsledky.

5.1 Colley Matrix Model

5.1.1 Výhody a nevýhody modelu

Mezi základní výhody tohoto modelu patří jeho jednoduchost a snadná interpretovatelnost získaných rankingů. Hlavním předpokladem tohoto modelu je náhodnost rozlosování týmů do jednotlivých ročníků soutěže. V případě, že v německé lize se využívá stejný systém rozlosování týmu do soutěžního ročníku jako v české lize, nemůže být tento předpoklad splněn, a proto výsledky mohou být ovlivněny. Hlavně mohou být ovlivněny rankingy, které jsou určeny v počátečních kolech soutěže. Další nevýhodou tohoto modelu je fakt, že při vytváření rankingů model nepřiřazuje větší váhu zápasům, ve kterých má tým silnějšího soupeře a naopak nižší váhu zápasům, které tým sehraje se slabším soupeřem. Tento model pracuje pouze s výsledkem zápasu prohra, výhra a remíze, proto nezohledňuje kolika gólový byl rozdíl branek na konci zápasu, což je způsobeno jednoduchostí tohoto modelu. Veškeré výpočty rankingů jsou uloženy v CD příloze této bakalářské práce.

5.1.2 Intepretace výsledků

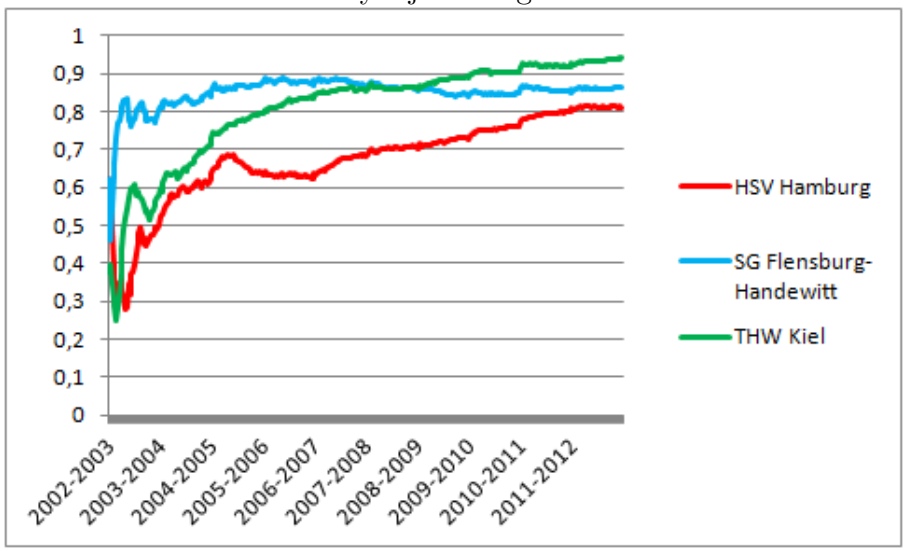
Tabulka 5.1 zachycuje pořadí tří nejúspěšnějších týmu německé soutěže v každé sezóně za celé sledované období. Za nejúspěšnější lze považovat tým THW Kiel, který na počátku sledovaného období, tedy v sezóně 2002-2003, obsadil 6. příčku a od následující sezóny do konce sledovaného období byl nejhůře druhý. Naproti tomu tým SG Flensburg-Handewitt na začátku sledovaného období byl úspěšnější a naopak v průběhu se postupně zhoršoval. U tým HSV Hamburg lze vysledovat spíše opačnou tendenci průběhu pořadí ve sledovaném období.

Tabulka 5.1: Tabulka pořadí vybraných týmů v celém sledovaném období

	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	10-11	11-12
THW Kiel	6	2	1	1	1	1	1	1	2	1
SG Flensburg-Handewitt	2	1	2	2	3	2	6	3	6	2
HSV Hamburg	8	5	6	10	2	3	2	2	1	4

V následujícím Obrázku 5.1 je zachycen časový průběh vývoje rankingů za celé sledované období, což vyjadřuje výše zmíněné dlouhodobé tendence síly jednotlivých týmů.

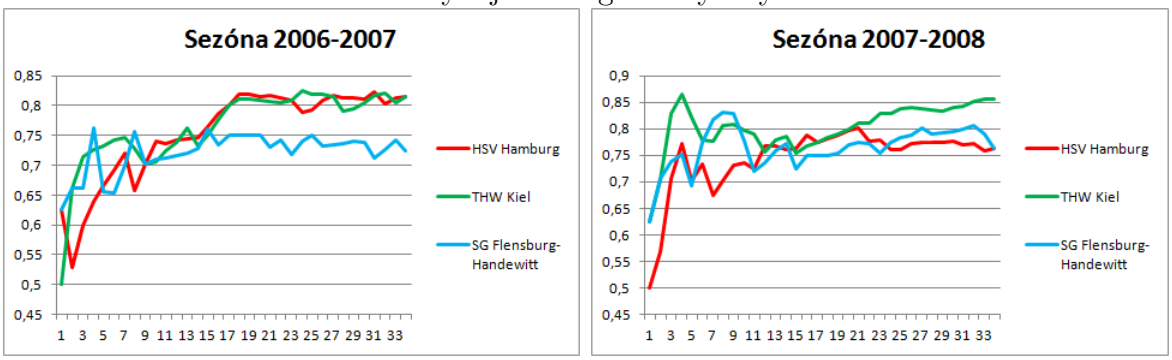
Obrázek 5.1: Vývoj rankingů za celé období



Pro další analýzu vývoje rankingů byly vybrány sezóny 2006-2007 a 2007-2008. Tyto sezóny byly vybrány, protože vybrané týmy obsadily první tři pozice. V Obrázku 5.2 je zobrazen vývoj rankingů vybraných týmu ve vybraných sezónách.

O sezóně 2006-2007 lze říci, že v první třetině sezóny byl průběh mezi třemi vybranými týmy napínavý a vyrovnaný. Naopak od 15. kola došlo k poklesu výkonnosti týmu SG Flensburg-Handewitt a bylo zřejmé, že už se nezapojí do bojů o první místo soutěže. Boj o první příčku probíhal tedy mezi týmy HSV Hamburg a THW Kiel a byl nerozhodnutý až do konce soutěže, kde v posledním kole dosáhly oba týmy shodného rankingového ohodnocení. O konečném pořadí mezi těmito dvěma týmy mohl rozhodnout například jen jeden prohraný zápas týmu HSV Hamburg.

Obrázek 5.2: Vývoj rankingů ve vyraných sezónách



V sezóně 2007-2008 došlo k rozhodujícímu okamžiku v boji o první příčku přibližně v 22. kole, kdy byly dominance týmu THW Kiel zřejmá a dále probíhal boj pouze o druhou příčku v soutěži. Tento boj svedly týmy HSV Hamburg a SG Flensburg-Handewitt, které po posledním kole opět dosáhly stejného rankingového ohodnocení. Skutečnost, že tým SG Flensburg-Handewitt skončil v konečném pořadí na druhé příčce, byla zřejmě způsobena lepšímy výsledky zápasů tohoto týmu v posledních kolech sezóny.

5.2 Keener Ranking Model

5.2.1 Výhody a nevýhody modelu

Tento model se od Colleyho modelu liší zejména tím, že je založen na počtu vstřelených a obdržených gólů gólů. Navíc nelineární složka modelu vyjádřena funkcí $h(x)$ zdůrazňuje pozici dominantního týmu, který v případě jasného vítězství hraje s lehkostí a tím pádem je rozdíl počtu gólů na konci zápasu daleko markantnější. Nevýhodou tohoto model je jeho počáteční variabilita, která se projevuje v počátečních kolech soutěže. Další nevýhodou je, že velké množství vstřelených branek v házené ovlivňuje hodnotu rankingů dle tohoto modelu a výsledný ranking nejsilnějších týmů je velmi podobný.

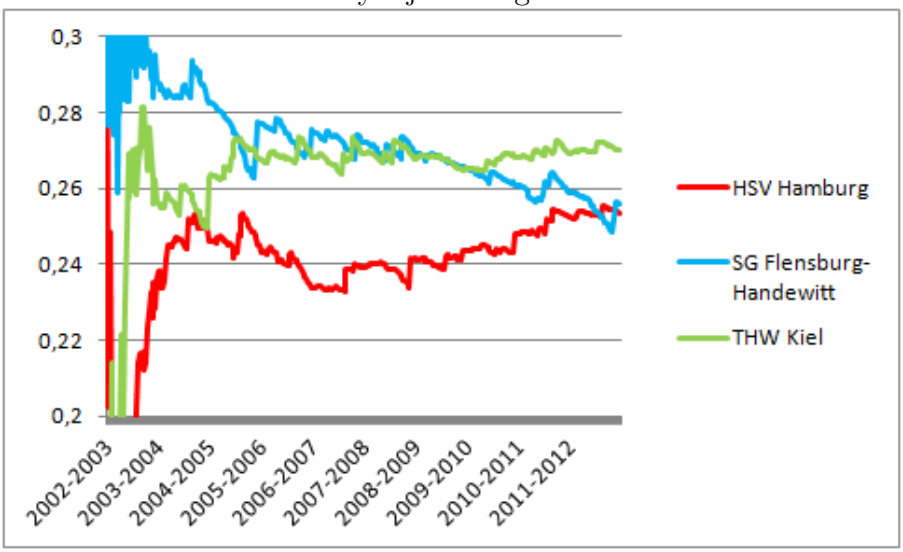
5.2.2 Interpretace výsledků

V Obrázku 5.3 je zachycen časový průběh vývoje rankingů za celé sledované období, což vyjadřuje výše zmíněné dlouhodobé tendence síly jednotlivých týmů.

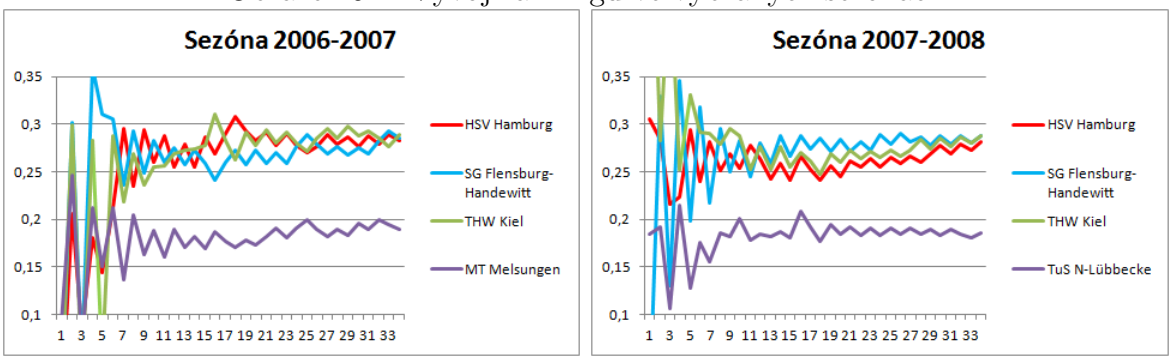
Pro další analýzu vývoje rankingů byly opět vybrány sezóny 2006-2007 a 2007-2008. Pro ilustraci byl vybrán jeden ze slabších týmu, konkrétně MT Melsungen. V Obrázku 5.4 je zobrazen vývoj rankingů vybraných týmu ve vybraných sezónách.

Na Obrázku 5.4 je vidět, že tento model není schopen rozlišit sílu u týmů, které střílí podobně vysoký počet gólů. Naopak sílu slabšího týmu je schopen zachytit velmi dobře.

Obrázek 5.3: Vývoj rankingů za celé období



Obrázek 5.4: Vývoj rankingů ve vybraných sezónách



5.3 Offense-Defense Model

5.3.1 Výhody a nevýhody modelu

Při výpočtu obranných a útočných rankingů vycházíme z předpokladu, že vyšší útočný ranking odpovídá vyšší útočné síle týmu a tedy vyššímu počtu vstřelených gólů, naopak nižší obranný ranking vypovídá o větší obranné síle týmu a tedy nižšímu počtu obdržených gólů. Nevýhodou tohoto modelu je fakt, že optimální ranking je získáván iteračním procesem. Konvergence procesu výpočtu rankingů je garantována Sinkhorn-Knoppovo větou [3, str. 33]. Předpoklady této věty jsou pro matici skóre S_{ij} takové, že všechny řádkové i sloupcové součty jsou kladné. Tento předpoklad nemusí být splněn v prvních kolech sezóny, kdy některý z týmů nemá sehraný ani jeden zápas. V tomto modelu je tento nedostatek eliminován přičtením malé kladné chyby ϵ. Další nevýhodou modelu je kromě volby ϵ pro výpočet rankingu modelu také volba zastavující podmínky iteračního procesu. Problematice citlivosti modelu na volbu ϵ a volbu zastavovací podmínky je věnován článek [3, str. 35]. Pro potřeby praktických výpočtu bylo voleno

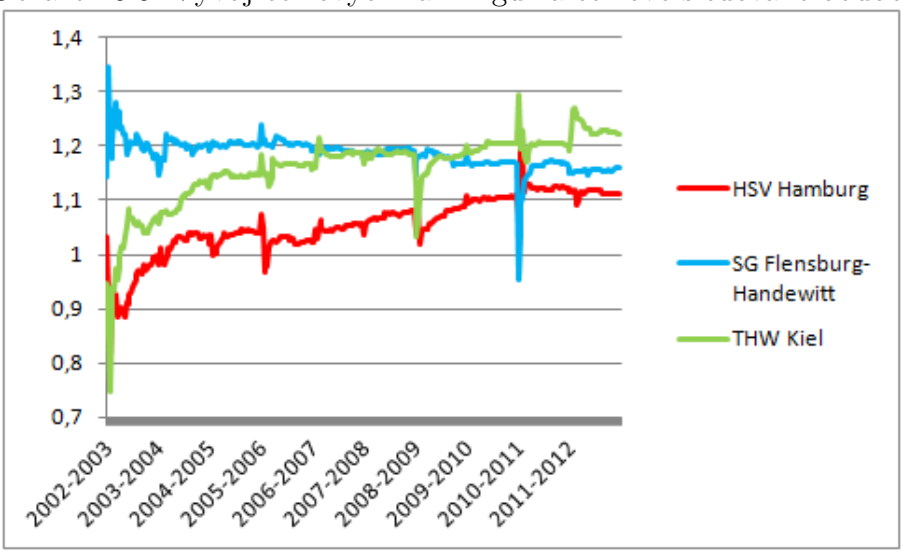
$\epsilon = 0,00001$ a zastavovací podmínky $||\boldsymbol{d}^{(k)} - \boldsymbol{d}^{(k-1)}|| < 0,01$.

Výhodou tohoto modelu je zejména to, že pro určení celkového rankingu týmu musí být určeny také jednotlivé rankingy obrany a útoku, které je možno využít pro podrobnější analýzu síly týmu.

5.3.2 Interpretace výsledků

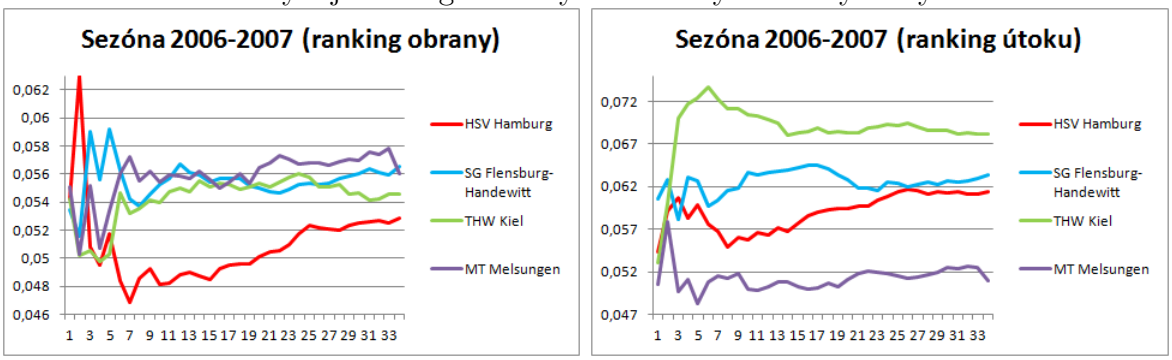
Na Obrázku 5.5, který zachycuje vývoj rankingu za celé sledované období, je zřejmá citlivost modelu v začátcích jednotlivých sezón, proto je tento model využitelnější pro výpočet rankingu pouze v jednotlivých sezónách, kdy nedochází ke změně týmů, které hrají danou sezónu.

Obrázek 5.5: Vývoj celkových rankingů za celkové sledované období

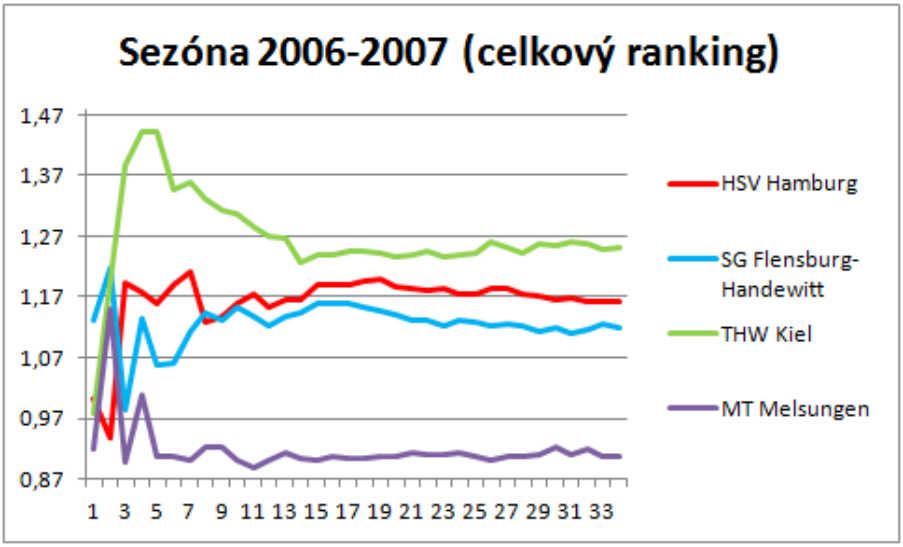


Pro ilustraci vývoje rankingů byla vybrána sezóna 2006-2007. Dále byl vybrán také tým MT Melsungem jako zástupce slabších týmů. V Obrázku 5.6 je zobrazen vývoj obranných a útočných rankingů vybraných týmu ve vybraných sezónách. Zachycené rankingy útoku a obrany ukazují, že například tým HSV Hamburg má lepší obranu oproti týmu THW Kiel, ale jeho útočná síla je menší než u THW Kiel, proto má celkový ranking (na vykreslený na Obrázku 5.7) nižší a byl také ve vybrané sezóně méně úspěšný.

Obrázek 5.6: Vývoj rankingů obrany a útoku týmů ve vybraných sezónách



Obrázek 5.7: Vývoj celkových rankingů ve vybraných sezónách



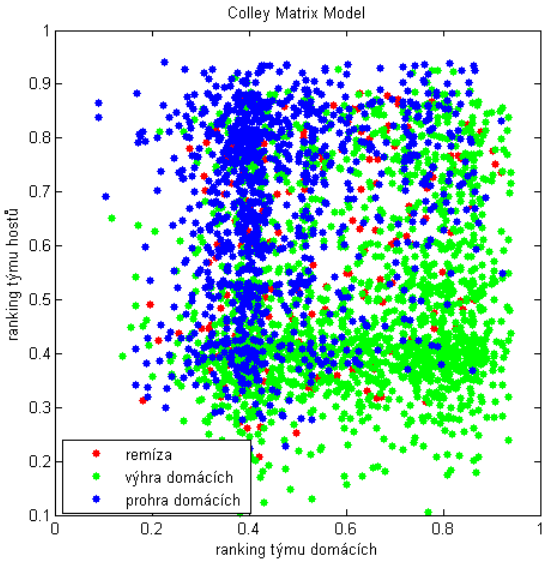
5.4 Porovnání hodnot rankingů s reálnými výsledky zápasů

Výsledné hodnoty rankingů byly využity pro předpověď výsledku zápasů a porovnání predikovaných výsledků s reálnými výsledky. Byly porovnávány výsledky z modelů:

1. Colley Matrix Model pro celé období (CMMC)
2. Colley Matrix Model pro jednotlivé sezóny (CMMJ)
3. Keener Ranking Model pro celé období (KRMC)
4. Keener Ranking Model pro jednotlivé sezóny (KRMJ)
5. Offense-Defense Model pro celé období (ODMC)
6. Offense-Defense Model pro jednotlivé sezóny (ODMJ)

Na následujícím Obrázku 5.8 jsou pro ilustraci zachyceny rankingy týmu hostů a týmu domácích určené dle CMMC. Barevně jsou znázorněny jednotlivé konečné stavy zápasů, modrá odpovídá výhře hostů, zelená výhře domácích a červená remízovému stavu.

Obrázek 5.8: Hodnoty rankingů týmu domácích a týmu hostů získané z Colley Matrix Model



Zápasy, které skončily remízou (červeně znázorněné), jsou rozmístěny po celém prostoru, což naznačuje, že predikovatelnost remízy na základě rankingů a z nich odvozených pravděpodobností je problematická. Proto nebyly remízové zápasy v této práci predikovány. Jedinou výjimkou byly situace, kdy rankingy obou týmů byly shodné, což byl velmi řídký jev.

Nejprve bylo vycházeno z hraniční hodnoty pravděpodobnosti výhry domácích $\pi_{ij} = 0.5$, což znamená, že pokud pravděpodobnost výhry domácích byla větší než 0.5 byla predikována výhra domácích, pokud pravděpodobnost výhry domácích byla menší než 0.5 byla predikována výhra hostů a pokud byla pravděpodobnost rovna přesně hodnotě 0.5, byla predikována remíza. Úspěšnost predikce byla měřena počtem, resp. procentem, shody predikovaného a reálného výsledku. Celková úspěšnost predikce na základě Colley Matrix Modelu pro celé období byla $68,73\%$, pro jednotlivé sezony $68,14\%$; na základě Keener Ranking Modelu pro celé období byla $67,52\%$, pro jednotlivé sezony $61,73\%$ a na základě Offense-Defense Modelu pro celé období byla $67,06\%$, pro jednotlivé sezony $68,53\%$. Výsledky ukazují, že úspěšnosti predikce podle jednotlivých modelů jsou si velmi blízké. Neočekávaně se ukázalo, že predikce získaná na základě nejjednoduššího modelu (Colley Matrix Model) byla nejúspěšnější.

V následujících Tabulkách 5.2, 5.3, 5.4, 5.5, 5.6 a 5.7 jsou zachyceny podrobné výsledky pro všechny modely. Je zde vždy zachycen počet zápasů, které skončily daných stavem a procentuální vyjádření predikovaných výsledků pro sledované konečné stavy zápasů. Procentuální úspěšnost předpovědi výhry domácích se pohybovala v rozmezí od $55,11\%$ do $69,20\%$, výhry hostů v rozmezí od $80,04\%$ do $86,89\%$. Předpovědi remíz byly dle očekávání neúspěšné.

Tabulka 5.2: Úspěšnost předpovědi CMMC

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,85	43,16	55,98
výhra domácích	1789	58,46	0,39	69,20	30,41
výhra hostů	1037	33,89	0,00	16,78	83,22

Tabulka 5.3: Úspěšnost předpovědi CMMJ

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	1,28	41,88	56,84
výhra domácích	1789	58,46	1,29	67,47	31,25
výhra hostů	1037	33,89	0,58	15,04	84,38

Tabulka 5.4: Úspěšnost předpovědi KRMC

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,85	44,44	54,70
výhra domácích	1789	58,46	0,39	67,41	32,20
výhra hostů	1037	33,89	0,00	17,26	82,74

Tabulka 5.5: Úspěšnost předpovědi KRMJ

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,85	34,62	64,53
výhra domácích	1789	58,46	1,12	55,11	43,77
výhra hostů	1037	33,89	0,48	12,63	86,89

Tabulka 5.6: Úspěšnost předpovědi ODMC

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	2,56	41,03	56,41
výhra domácích	1789	58,46	3,24	67,97	28,79
výhra hostů	1037	33,89	3,47	16,49	80,04

Tabulka 5.7: Úspěšnost předpovědi ODMJ

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	1,28	42,74	55,98
výhra domácích	1789	58,46	1,23	68,64	30,13
výhra hostů	1037	33,89	0,87	15,62	83,51

V další fázi byla brána v úvahu predikce všech modelů a výsledek zápasu byl předpovězen na základě nejčastěji se objevující predikce konečného stavu. Tedy pokud například čtyři ze šesti modelů předpověděly výhru domácích, byla předpovězena výhra domácích. Pokud modely předpověděly výsledek nerozhodně (tedy tři předpověděly výhru domácích a tři výhru hostů), byla předpovězena výhra domácích, neboť lze očekávat výhodu domácího prostředí. Při takto získaných předpovědích byla předpověď správná v $70,13\%$ zápasů. Podrobnější informace jsou shrnuty v následující Tabulce 5.9. Z této tabulky je vidět, že využití informací ze všech modelů

Tabulka 5.8: Úspěšnost předpovědi na základě všech modelů

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,85	47,44	51,71
výhra domácích	1789	58,46	1,06	71,72	27,22
výhra hostů	1037	33,89	0,29	16,68	83,03

umožnilo zlepšit předpověď výhry domácích na $71,72\%$ a výhry hostů na $83,03\%$.

Na závěr byl zkoumán vliv hraniční hodnoty 0.5 na úspěšnost predikce výsledku zápasu. Lze předpokládat, že i v házené je podstatná výhoda domácího prostředí. V takovémto případě by optimální hraniční hodnota byla menší než 0.5, tedy byla predikována výhra i týmu, který má pravděpodobnost výhry v domácím prostředí o něco menší než 0.5 , ale větší než zvolená optimální hodnota. Simulačně byly otestovány hodnoty blízké 0.5, konkrétně hodnoty od $0,4$ do $0,6$. na základě výsledků byly určeny optimální hodnoty hraničních bodů pro jednotlivé modely. Shrnutí je uvedeno v Tabulce 5.9.

Tabulka 5.9: Úspěšnost předpovědi na základě všech modelů

model	optimální hraniční hodnota	% úspěšně předpovězených zápasů
CMMC	$0,432$	$71,80$
CMMJ	$0,435$	$71,57$
KRMC	$0,459$	$69,25$
KRMJ	$0,458$	$66,96$
ODMC	$0,480$	$70,29$
ODMJ	$0,481$	$72,03$

Z Tabulky je zřejmé, že ve všech modelech se projevila výhoda domácího prostředí, což znamená, že optimální hraniční hodnota byla menší než 0.5. Snížení hraniční hodnoty vedlo ve všech šesti modelech ke zvýšení úspěšnosti predikce výsledků. Jako nejúspěšnější byl považován Offense-Defense Model, který ranking odhadoval na základě výsledků jednotlivých sezón. I když rankingy určované pro jednotlivé sezóny by měly vyjadřovat aktuální formu týmu a predikce výsledků dle nich by měly být přesnější, potvrdil se tento předpoklad jen u Offense-Defense Modelu.

Pro ilustraci jsou uvedeny podrobné výsledky nejúspěšnějšího modelu se sníženou hranicí rozhodování o výsledku zápasu v Tabulce 5.10.

Tabulka 5.10: Úspěšnost předpovědi na základě ODMJ se sníženou hranicí

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,43	69,66	29,91
výhra domácích	1789	58,46	0,84	82,28	16,88
výhra hostů	1037	33,89	0,87	28,64	70,49

V konečné fázi byl použit postup, kdy výsledný predikovaný stav byl odvozen ze všech modelů. U modelů byly voleny optimální hraniční hodnoty uvedené výše a výsledek byl opět predikován podle nejčastěji se objevujícího predikovaného stavu. Tímto postupem se zvýšila úspěšnost predikce na $72,29\%$. Podrobné výsledky jsou uvedeny v Tabulce 5.11.

Tabulka 5.11: Úspěšnost předpovědi na základě všech optimalizovaných modelů

reálné výsledky			% predikovaných výsledků		
výsledek	počet	%	remíza	výhra domácích	výhra hostů
remíza	234	7,65	0,00	75,64	24,36
výhra domácích	1789	58,46	0,17	88,26	11,57
výhra hostů	1037	33,89	0,48	38,48	61,04

Závěr

Tato práce byla zaměřena na analýzu výsledků zápasů z mužské házené. Jako datové zdroje byly použity výsledky zápasů v německé a české nejvyšší soutěži z let 2002-2012. Celkově bylo k dispozice 3060 výsledků zápasů z německé soutěže a 1614 výsledků zápasů z české nejvyšší soutěže. Nejprve byla data upravena. Zejména bylo třeba sjednotit názvy týmů v české soutěži, protože se v průběhu sledovaného období často měnily. Následně byla provedeno základní statistické zpracování dat.

Teoretické postupy a modely byly shrnuty v Kapitolách 2 a 3. V těchto kapitolách lze najít základní informace o použitých metodách , formulace metod ve formě, ve které byly dále využívány v praktické části této práce.

V Kapitole 4 je vždy nejprve uvedena motivace k vytvoření té konkrétní hypotézy, následně je zdůvodně výběr testu pro testování hypotézy, její matematické formulace a interpretovány získané výsledky testu. Výsledky testování normality dat ukazují, že porušení normality není při vysokých počtech gólů, které v házené padají, závažným problémem, protože v řadě případů nebyla normalita dat zamítnuta. Další hypotéza byla zaměřena na identifikaci lineární závislosti počtu gólů na počtu odehraných kol. Z výsledků testů lze usuzovat, že hypotéza o tom, že s přibývajících počtem kol se zvyšuje také počet vstřelených gólů je platná. Další studovaná hypotéza se zaměřila na závislost konečného výsledku zápasu na poločasovém skóre. Podrobná analýza ukázala, že u zápasů, kde rozdíl v poločase je nejvýše jeden gól, jsou konečné výsledky nezávislé na rozdílu počtu gólů v prvním poločase. To posílilo domněnku, že zápasy, které v poločase končí malým rozdílem v počtu gólů jsou otevřené až do konce zápasu. Jako poslední byla srovnávána dynamičnost hry v české a německé lize. Dynamičnost byla porovnávána podle toho, jak často v české a v německé lize dochází ke změně stavu zápasu mezi prvním poločasem a konečných výsledkem zápasu. Výsledky testů ukázaly, že domněnka o tom, že v německé lize dochází k častější změně stavu mezi výsledkem v poločase a na konci zápasu, se nepotvrdila.

V poslední Kapitole 5 bylo provedeno hodnocení síly jednotlivých týmů z německé soutěže na základě třech rankingových modelů. Jednalo se o Colley Matrix Model, Keener Ranking Model a Offense-Defense Model. Byly odhadnuty rankingy pro jednotlivé týmy a z těchto rankingů byl odhadnut výsledek konkrétního zápasu. Tyto predikované výsledky byly porovnány s reálnými výsledky zápasů. Pro jednotlivé modely se úspěšnost predikce pohybovala v rozmezí od $61,73\%$ do $72,29\%$.Jako nejúspěšnější byl při pevně zvolené rozhodovací hranici zhodnocen Colley Matrix Model s úspěšností odhadu $68,73\%$. Pokud byly brána v úvahu výhoda domácího prostředí, tedy optimalizována rozhodovací hranice, za nejúspěšnější byl považován Offense-Defense Model s úspěšností $72,03\%$. Pokud byly využity pro predikci výsledků zápasů výstupy ze všech modelů, byla úspěšnost predikce $72,29\%$.

Literatura

[1] Jiří Anděl. *Základy matematické statistiky.* Matfyzpress, 2007.

[2] Tomáš Cipra. *Ekonometrie.* Ekopress, 2008.

[3] Anjela Yuryevna Govan. *Ranking Theory with Application to Popular Sports.* Doctor Thesis of North Carolina State University, 2008.

[4] P. James Keener. *The Perron-Frobenius Theorem and the Ranking of Football Teams.* Siam Review, 1993.

[5] Lukáš Kotlorz. *Testy normality.* Bakalářská práce MFF UK, 2012.

[6] Daniel Novák. *The Czech Handball Server.* <http://www.hazena.pb.cz/> [cit. 29. 10. 2012].

[7] Jiří Reif. *Metody matematické statistiky.* Západočeská univerzita, 2004.

[8] Karel Zvára. *Regrese.* Matfyzpress, 2009.

Obsah přiloženého CD

1. **hlavnitext.pdf**-soubor obsahující text bakalářské práce
2. **1-Německá handbalová soutěž.xlsx**-soubor obsahující základní statistické zpracování dat pro německou soutěž
3. **2-Česká handbalová soutěž.xlsx**-soubor obsahující základní statistické zpracování dat pro českou soutěž
4. **dataNem.mat**-soubor obsahující data pro práci v Matlabu
5. **hypoteza01.m**-soubor obsahující program pro testy normality
6. **hypoteza02.m**-soubor obsahující program pro testy lineární závislosti
7. **hypoteza03.m**-soubor obsahující program pro testy závislosti konečného výsledku na poločasovém skóre
8. **ColleyMatrixModel.m** a **ColleyMatrixModelRocni.m**-soubory obsahující program pro výpočty rankingů dle Colley Matrix Modelu
9. **KeenerModel.m** a **KeenerModelRocni.m**-soubory obsahující program pro výpočty rankingů dle Keener Ranking Modelu
10. **OffenseDefenseModel.m** a **OffenseDefenseModeR.m**-soubory obsahující program pro výpočty rankingů dle Offense-Defense Modelu
11. **Ranking.mat**-soubor obsahující data pro práci s rankingy
12. **predikce.m**-soubor obsahující program pro predikci výsledků zápasů