Modelování a predikce hokejových zápasů*

Patrice Marek

Západočeská univerzita v Plzni

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

* Podpořeno z OPVK CZ.1.07/2.2.00/15.0377

OBSAH

- Předmět modelování
- Přehled modelů
- Vlastní model
- Data
- Experimenty
- Výsledky a budoucí práce

PŘEDMĚT MODELOVÁNÍ

PŘEHLED MODELŮ

- Dva základní přístupy:
  - Nepřímý: modelování počtu gólů.
  - Přímý: modelování výhry, prohry a remízy.
- Modely tvořeny často pro fotbal.
- Maher (1982) – dvě nezávislá Poissonova rozdělení.
- Novější modely používají dvourozměrné Poissonovo rozdělení.
  - Problém – umožňují pouze pozitivní korelaci.
- Karlis and Ntzoufras (2003):
  - Použití modelů pro vodní pólo.

MODELY

- Pracujeme s nepřímým přístupem, tj. modelování počtu gólů.
- Předpokládáme, že počet gólů má Poissonovo rozdělení (např. Maher (1982)).

$$X_H \sim Poisson(\lambda_{1H}), \qquad \lambda_{1H} = \mu \cdot \gamma \cdot \alpha_H \cdot \beta_A$$

$$Y_A \sim Poisson(\lambda_{2A}), \qquad \lambda_{2A} = \mu \cdot \alpha_A \cdot \beta_H$$

- Pro každý tým uvažujeme:
  - α_i … síla útoku ($\alpha_i > 0 \ \forall i, \quad \frac{1}{n}\sum_{i=1}^{n} \alpha_i = 1$) a
  - β_i … slabost obrany ($\beta_i > 0 \ \forall i, \frac{1}{n}\sum_{i=1}^{n} \beta_i = 1$).
- Globální parametry
  - μ … konstanta.
  - γ … efekt domácího prostředí,
  - ρ … korelace (bude použito později).

MODELY

- Dvourozměrné Poissonovo rozdělení

$$P(x,y) = e^{-(\lambda_H+\lambda_A+\lambda_3)} \cdot \sum_{i=0}^{\min(x,y)} \frac{\lambda_H^{x-i}\lambda_A^{y-i}\lambda_3^{i}}{(x-i)!\,(y-i)!\,i!}$$

  - $\mathrm{corr}(X,Y) = \lambda_3 = \rho$
  - $\lambda_3 > 0$, tj. nelze modelovat negativní korelaci.
  - V našich datech se vyskytuje negativní korelace mezi počtem gólů domácích a hostů ⇒ nový model.

MODELY

- Můžeme použít copuly ke spojení dvou Poissonovo rozdělení.
  - Umožní modelovat negativní korelaci.
  - Copuly jsou vhodné pro spojitá data.
  - Použití pro diskrétní data je problematické, ale možné (Genest and Nešlehová (2007)).
- Použili jsme Frankovu copulu (Dobson and Goddart (2011))

$$G(F_H(x), F_A(y)) = P(X \leq x, Y \leq y) = \\ = \frac{1}{\varphi} \ln\left(1 + \frac{\left(e^{\varphi F_H(x)} - 1\right) \cdot \left(e^{\varphi F_A(y)} - 1\right)}{e^{\varphi} - 1}\right)$$

  - $F_{\cdot}(.)$ značí jednorozměrnou distribuční funkci.
  - $\varphi = -\rho = -\mathrm{corr}(X, Y)$

MODELY

- Odehrané zápasy neobsahují stejnou informaci.
  - Toto bereme v modelu v úvahu.
  - Místo maximalizace logaritmicko-věrohodnostní funkce používáme pseudo-logaritmicko-věrohodnostní funkci (inspirace z Dixon a Coles (1997)).

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \mu, \rho) = \sum_{i=1}^{k} \kappa(t_i) \cdot \ln P(x_i, y_i)\,,$$

kde i vyjadřuje $i$tý zápas.

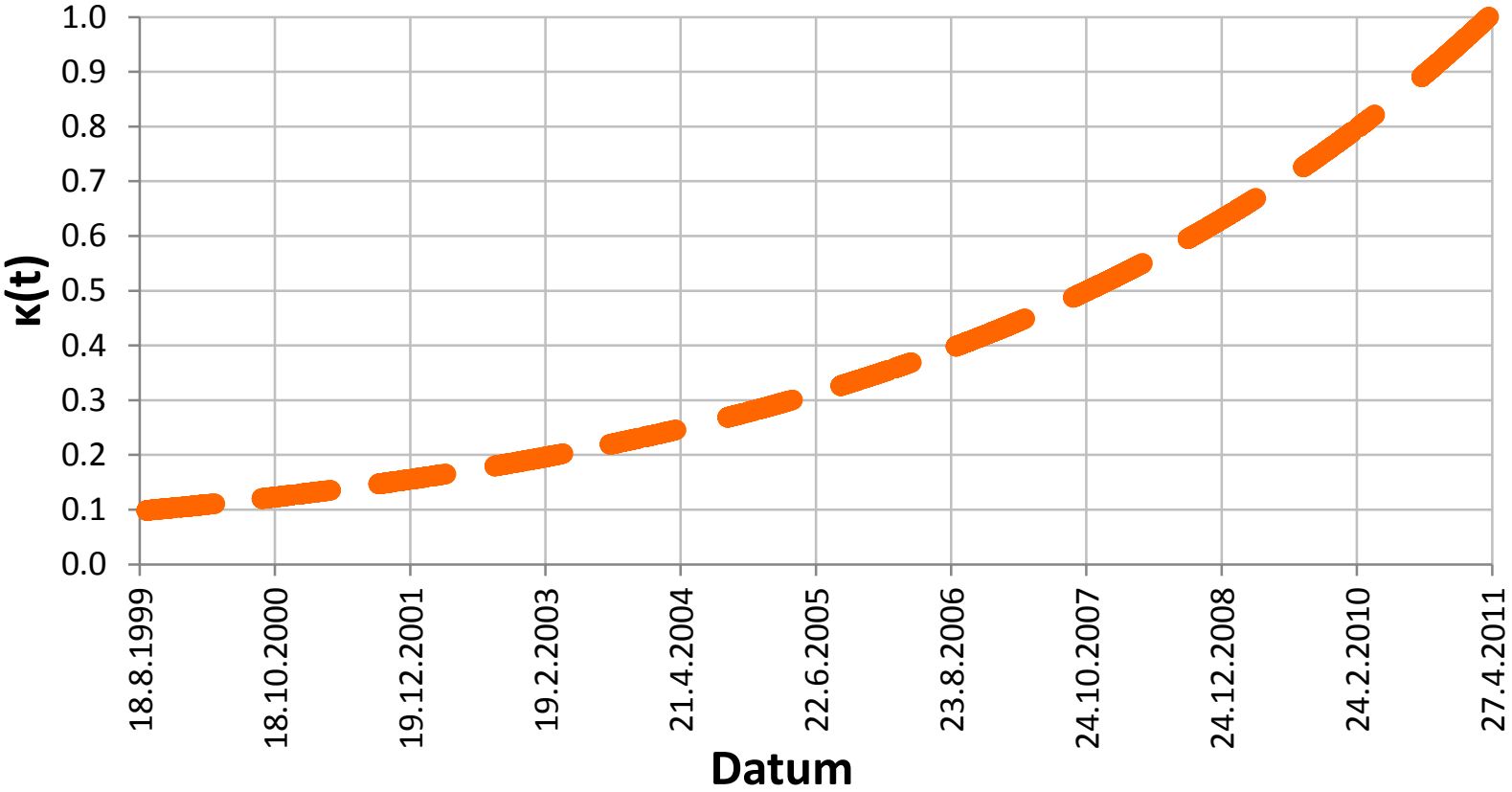
  - Váhy pro každý zápas jsou dány následujícím předpisem

$$\kappa(t_i) = e^{-\frac{\omega(T-t_i)}{365.25}}, \qquad \omega > 0,$$

kde T je aktuální datum, t_i je datum, kdy byl ten který zápas odehrán a rozdíl je měřen v dnech.

MODELY

- Váhy $\kappa(t) = e^{-\frac{\omega(T-t)}{365.25}}$, kde $\omega = 0.2$.



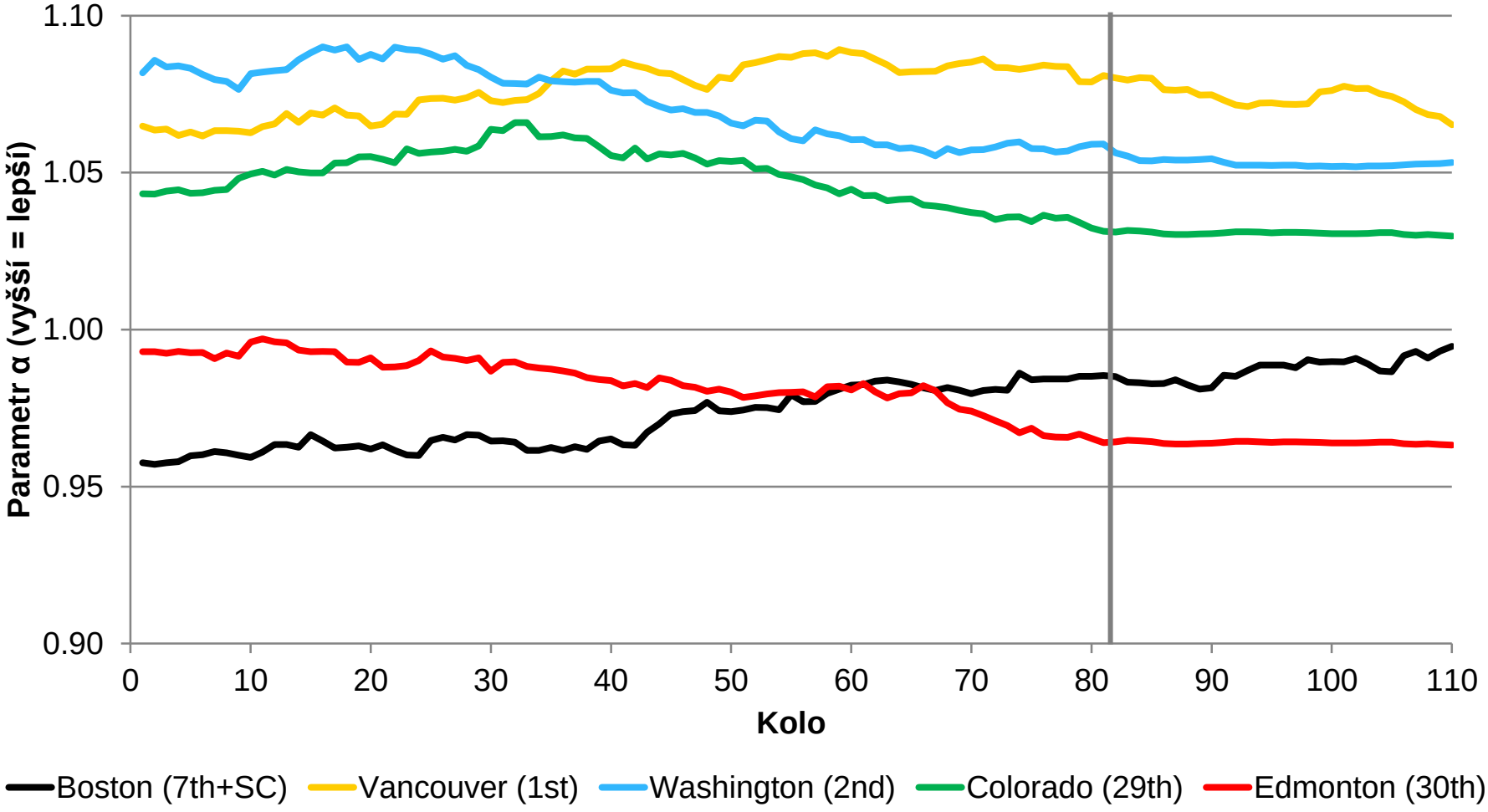
DATA

- NHL – National Hockey League (USA a Kanada):
  - data od sezóny 1999/2000 do sezóny 2010/2011,
  - 14 398 zápasů,
  - Model testován pro sezónu 2010/2011 (82 + 28 kol).
- Extraliga – Nejvyšší hokejová liga v ČR:
  - data od sezóny 1999/2000 do sezóny 2010/2011,
  - 4 998 zápasů,
  - Model testován pro sezónu 2010/2011 (52 + 25 kol).
- Zajímavý (problematický) fakt: negativní korelace mezi počtem gólů domácích a hostů.

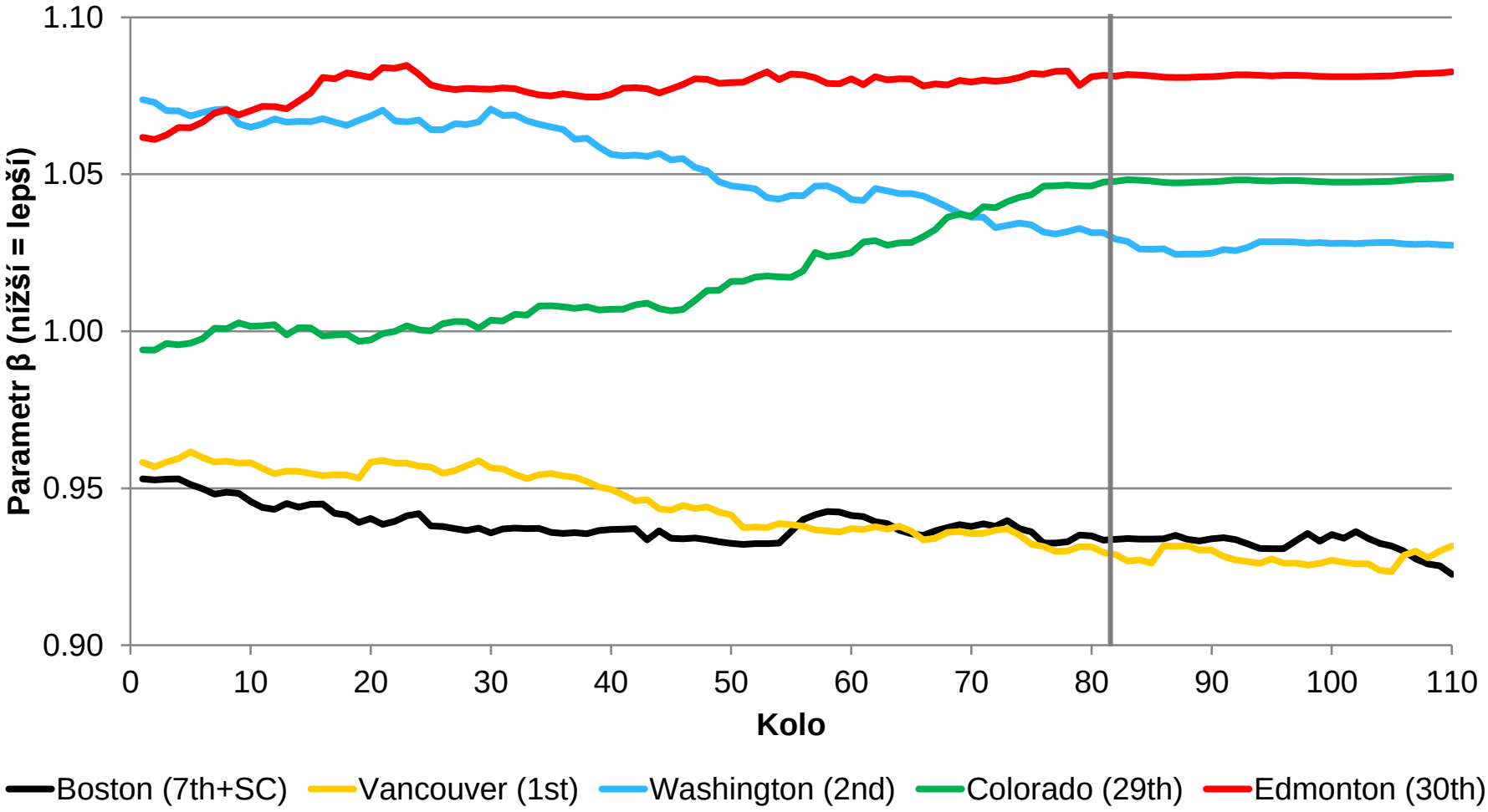
EXPERIMENTY

- Představíme parametry některých týmů z NHL.
- Dva nejlepší týmy z NHL (82 kol)
  - Vancouver Canucks (117 bodů) a
  - Washington Capitals (107 bodů).
- Dva nejhorší týmy z NHL (82 kol)
  - Colorado Avalanche (68 bodů) a
  - Edmonton Oilers (62 bodů).
- Tým Boston Bruins získal Stanley Cup (playoff) ve finále proti Vancouver Canucks.

EXPERIMENTY – ÚTOK



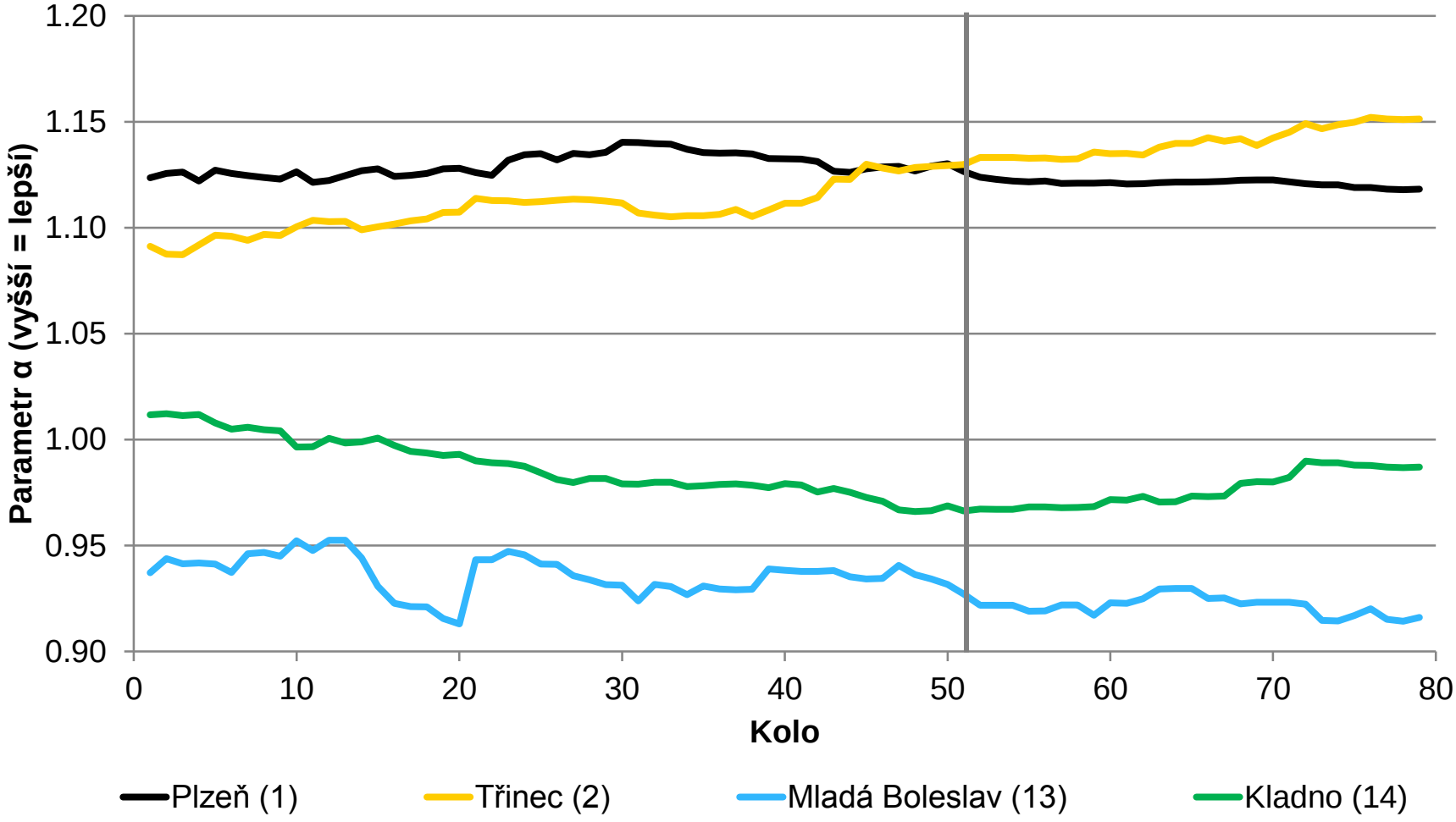
EXPERIMENTY – OBRANA



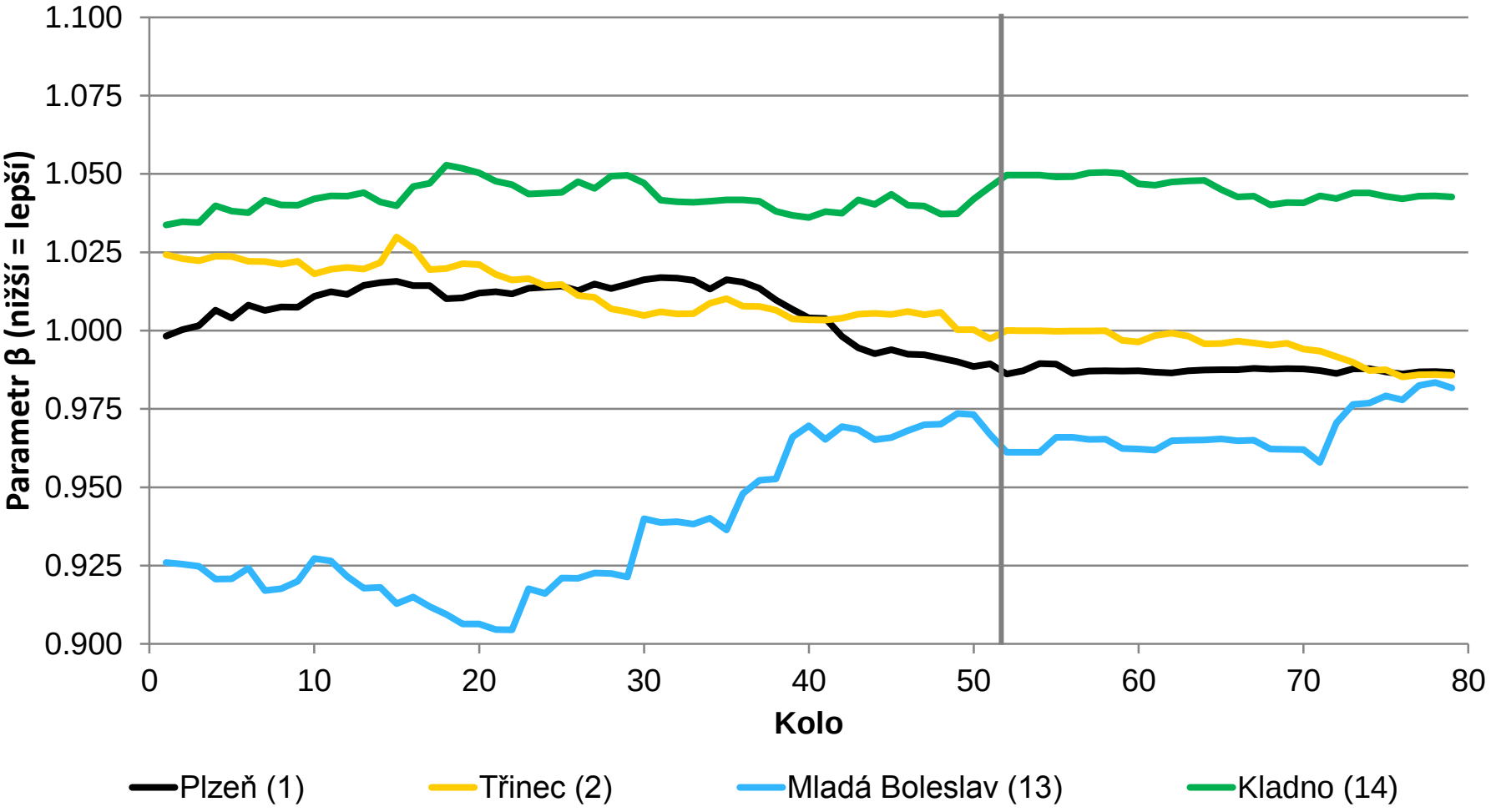
EXPERIMENTY

- Představíme parametry některých týmů z Extraligy.
- Dva nejlepší týmy z Extraligy (52 kol)
  - Plzeň (96 bodů) a
  - Třinec (96 bodů).
- Dva nejhorší týmy z Extraligy (52 kol)
  - Mladá Boleslav (55 bodů) a
  - Kladno (41 bodů).
- Třinec následně zvítězil v playoff.
- Parametry jsou ovlivněny i přítomností týmů z historie, tj. těch které se v Extralize objevili ale již sestoupili.

EXPERIMENTY – ÚTOK



EXPERIMENTY – OBRANA



EXPERIMENTY – PREDIKČNÍ SCHOPNOST

- Použili jsme χ^2 test dobré shody pro srovnání za celou sezónu.
  - Pro NHL jsme získali p-hodnotu < 0.001.
  - Pro Extraligu jsme získali p-hodnotu $= 0.18$.
- Tabulky obsahují poměr $\frac{o_{ij}}{n_{ij}}$ a zvýraznění je založeno na $\frac{(n_{ij}-o_{ij})^2}{o_{ij}} = A$

  n_{ij} … skutečné počty výsledků a o_{ij} … teoretické počty výsledků.

Extraliga		Hosté					
		0	1	2	3	4	5+
Domácí	0	---	2.41	0.72	0.77	0.93	0.53
	1	0.88	0.90	2.14	0.97	0.90	0.88
	2	1.36	1.37	0.67	1.43	0.70	2.22
	3	0.76	0.94	0.99	0.64	1.65	0.86
	4	0.70	0.87	1.15	1.48	1.70	1.05
	5+	1.02	1.13	1.02	0.88	1.06	1.16

NHL		Hosté					
		0	1	2	3	4	5+
Domácí	0	0.73	1.15	1.38	0.54	0.84	0.88
	1	0.87	0.67	1.26	1.09	0.73	1.23
	2	1.25	1.72	0.73	1.12	0.93	0.94
	3	0.88	1.23	1.22	0.63	1.09	1.17
	4	0.92	0.99	1.19	1.54	0.84	1.91
	5+	0.87	1.18	0.88	0.94	1.18	1.27

Zvýraznění: $A > 10$, $A \in [5, 10)$, $A \in [1, 5)$

EXPERIMENTY – PREDIKČNÍ SCHOPNOST

- Srovnáme výsledky za celou sezónu

	Extraliga			NHL		
	Výhra	Remíza	Prohra	Výhra	Remíza	Prohra
Realita (R)	0.5214	0.2032	0.2754	0.4085	0.2422	0.3493
Sázková kancelář (B)	0.4695	0.2163	0.3142	0.4307	0.2288	0.3404
Náš model (M)	0.5267	0.1746	0.2987	0.4555	0.1829	0.3616

Absolutní rozdíl						
$abs(R-B)$	0.0520	0.0131	0.0388	0.0222	0.0134	0.0089
$abs(R-M)$	0.0053	0.0286	0.0233	0.0470	0.0593	0.0123

Relativní vychýlení						
$abs\left(\frac{R}{B}-1\right)$	0.0996	0.0645	0.1411	0.0544	0.0552	0.0253
$abs(\frac{R}{M}-1)$	0.1219	0.1928	0.0494	0.0576	0.2008	0.0621

VÝSLEDKY A BUDOUCÍ PRÁCE

- Ukázali jsme, že modely používané pro fotbal mohou být použity i pro hokej.
- Spojili jsme model založený na Frankově copule a použití vah pro historické zápasy (dosud nebylo provedeno pro fotbal).
- V obou případech lze pozorovat podhodnocování pravděpodobnosti remíz ⇒ musí být vyřešeno.
  - Možné řešení spočívá v použití směsi dvou rozdělení:

$$P_D(x,y) = \begin{cases} (1-\pi)\cdot P(x,y) & pro \quad x \neq y \\ (1-\pi)\cdot P(x,y) + \pi \cdot D(x,\theta) & pro \quad x = y \end{cases}$$

kde $D(x,\theta)$ je diskrétní rozdělení (např. Karlis and Ntzoufras (2003)).

ZDROJE

- Dixon, M., Coles, S. (1997). *Modelling Association Football Scores and Inefficiencies in Football Betting Market*. Applied Statistics, 46, 265–280.
- Dobson, S., Goddard, J. (2011). *The Economics of Football, 2nd ed.*, Cambridge University Press, Cambridge.
- Genest, C., Nešlehová, J. (2007). *A primer on copulas for count data*, Astin Bulletin, 37(2), pp. 475–515.
- Maher, M. J. (1982). *Modelling association football scores*. Statistica Neerlandica, 36: 109–118.
- Karlis D, Ntzoufras I (2003). *Analysis of Sports Data by Using Bivariate Poisson Models*. Journal of the Royal Statistical Society D (The Statistician), 52, 381-393.

DĚKUJI ZA VAŠI POZORNOST!