UNIVERZITA PALACKÉHO V OLOMOUCI

PŘÍRODOVĚDECKÁ FAKULTA

Katedra matematické analýzy a aplikací matematiky

Diplomová práce

Optimalizace výpočtu účinnosti rotačních hydraulických strojů pomocí metod predikce v časových řadách

Vedoucí diplomové práce:
RNDr. Tomáš Fürst, Ph. D.
Rok odevzdání: 2017

Vypracoval:
Bc. Daniela Malá
AME, II. ročník

Prohlašuji, že diplomovou práci jsem vypracovala samostatně pod vedením Tomáše Fürsta. Všechny použité zdroje jsem uvedla v seznamu. Souhlasím, aby práce byla uložena na Univerzitě Palackého v knihovně Přírodovědecké fakulty a ve studijní agendě a dále zpřístupněna ke studijním účelům.

V Olomouci dne: ……. Daniela Malá

Poděkování

Děkuji Tomáši Fürstovi za odborné vedení při tvorbě mé diplomové práce, za ochotu při konzultacích a za poskytnutí velmi cenných rad a poznatků, které mi dopomohly k jejímu úspěšnému dokončení. Rovněž děkuji všem mým drahým blízkým osobám za trpělivost a podporu.

Bibliografická identifikace:

Jméno a příjmení	Daniela Malá
Název práce	Optimalizace výpočtu účinnosti rotačních hydraulických strojů pomocí metod predikce v časových řadách
Typ práce	Diplomová
Katedra	Matematické analýzy a aplikací matematiky
Vedoucí práce	Tomáš Fürst
Rok obhajoby práce	2017
Abstrakt	Práce obsahuje analýzu možnosti předpovědí v časových řadách simulovaných účinností čerpací pumy. Konkrétně jde o to, zda je už na začátku CFD výpočtu možné odhadnout, zda se výpočet ustálí a na jaké hodnotě se tak stane. Použitý matematický aparát sestává z regresních modelů (lineární a logické regrese) a shlukové analýzy.
Klíčová slova	Časové řady, Shlukovací metody, K-means, Vážená lineární regrese, logistická regrese.
Abstrakt v angličtině	The Thesis analyzes possible means of predicting the time series of simulated efficiency of water pumps. The goal is to show if it is possible to predict the equilibrium late-time values of the time series from the knowledge of the beginning of the series. The mathematical tools include regression models (linear and logistic), and cluster analysis.
Klíčová slova v angličtině	Time series, clustering method, K-means, weighted linear regression, logistic regression.
Počet stran	50
Jazyk	Český

Obsah:

Úvod

Na této práci jsem spolupracovala s Centrem hydraulického výzkumu, Sigma Lutín, ve spolupráci s panem Tomášem Krátkým. Sigma Lutín, byla založena v roce 1868 a specializuje se na výrobu čerpacích pump, které slouží k čerpání vody. Cílem Centra hydraulického výzkumu je převážně co nejrychlejší vývin pump, které budou mít co nejvyšší účinnost přeměny mechanické energie na kinetickou energii kapaliny a zároveň budou tyto pumpy co nejlevnější. U jednotlivých pump se liší požadavky jak na průtok, tak na dopravní výšku. Jak průtok, tak dopravní výška vyplývá z přírodních i stavebních podmínek, které nelze změnit.

Účinnost závisí na proudění kapaliny v čerpadle. Toto chování popisují nelineární parciální diferenciální rovnice, které jsou skrze geometrii oblasti a okrajové podmínky závislé na tvaru čerpadla. Víme, že neexistuje univerzálně dokonalý tvar čerpadla, vždy závisí na požadovaných parametrech. Není známé obecné analytické řešení těchto rovnic (tj. jak získat analytické chování proudění pro zadaný tvar), taktéž není známo jednoduché řešení inverzního problému (tj. jak získat správný tvar čerpadla pro zadaný charakter proudění). Hydraulický návrh pumpy probíhá iteračním způsobem tak, že se vytvoří návrh, na kterém se ověří výsledky požadovaného průtoku vody a požadované dopravní výšky. Následně proběhne korekce. Další způsob k vytvoření hydraulického návrhu pumpy je využití tvarové optimalizace.

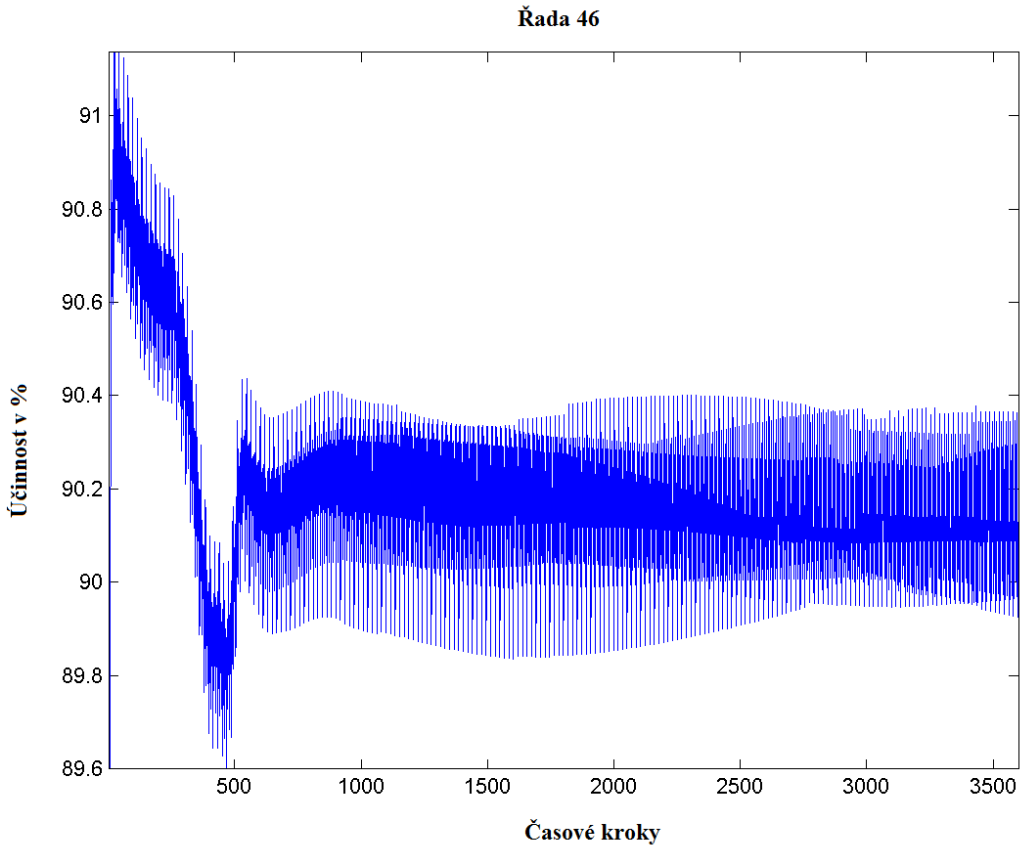
K vytvoření návrhu hydraulického návrhu pumpy lze pro zjednodušení využít 2D modely proudění umožňující přibližný nástřel správného tvaru pumpy. Korekci lze samozřejmě provádět na základě statistiky již vyřešených tvarů. K vytvoření konečného modelu je potřeba získat přesné parametry hydraulického návrhu pumpy, čehož lze dosáhnout dvěma způsoby:

1. Experimentem, který je dosti finančně náročný, protože typicky stojí v řádu milionů korun.
2. CFD výpočtem.

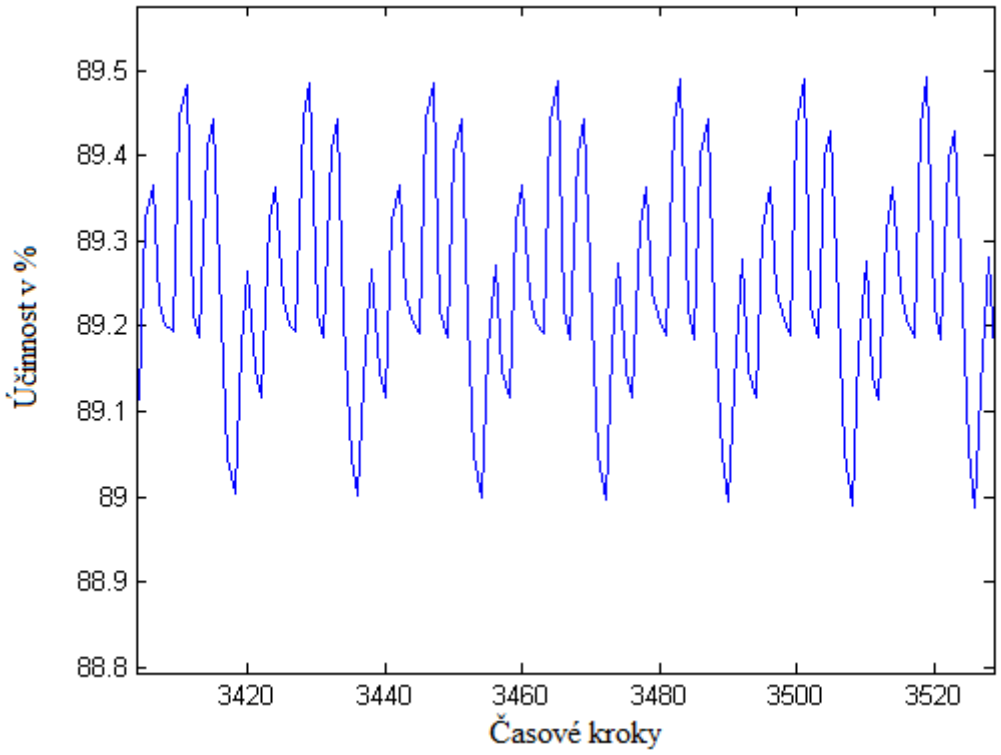
CFD (Computational Fluid Dynamics) výpočet představuje numerické modelování proudění vody v pumpě. Toto numerické modelování je založeno na numerické aproximaci rovnic proudění a snaží se zachytit rychlostní pole pomocí po částech polynomiálních funkcí. CFD využívá metody konečných prvků respektive objemů, což ve výsledku vede na řešení velkých soustav algebraických rovnic, z čehož vyplývá velká výpočetní náročnost. Na běžných počítačích zabírá řešení dny až týdny a to pro každou drobnou změnu v návrhu (geometrii) čerpací pumpy. Řešení se vždy počítá pro více průtoků čerpací pumpy, protože čerpací pumpa pracuje v určitém rozmezí, čímž výpočet dále narůstá. Výsledky těchto řešení jsou však v dobré shodě s experimenty a cena se pohybuje zhruba o řád níže, tedy v řádech statisíců korun. Z tohoto důvodu dnes Centrum hydraulického

výzkumu používá CFD výpočty a experimentem pouze ověřuje finální návrh tvaru čerpacích pump.

Ve spolupráci s panem Tomášem Krátkým mi Sigma Lutín poskytla data jedné sady tvarové optimalizace. Data obsahují 73 CFD simulací, z nichž každá je evaluována v 3600 časových okamžicích. Z vypočtených rychlostních a tlakových polí je evaluována celková účinnost pumpy v daném čase. Tyto dlouhé (3600 časových kroků) simulace CFD výpočtu považujme za časové řady průběhu vypočtené celkové účinnosti pumpy. Časový krok je z důvodu numerické přesnosti zvolený zhruba tak, aby se oběžné kolo během jednoho časového kroku pootočilo v rozmezí 2 až 3 úhlových stupňů. Úhel 2 až 3 stupňů vychází z empirických zkušeností s numerickými výpočty čerpacích pump. U této volby se jedná o rozumný kompromis mezi přesností (čím kratší časový krok, tím více časových kroků na jednu otáčku, a tím přesnější výpočet) a rychlostí výpočtu. Na celou otočku oběžného kola (360°) připadá tedy asi 126 časových kroků. V našem případě se kolo vždy otočí 28,5 krát. Průběh účinnosti je teoreticky periodický s předem známou periodou. Děj je periodický z fyzikální podstaty, kolo rotuje kolem své osy. Perioda je předem známá a uživatelsky volená. V mém případě je perioda 18 časových kroků. Těchto 73 různých návrhů tvaru se týká jedné navrhované čerpací pumpy. CFD simulace se provádí v softwaru REF. Na následujícím **Obrázku 1. 1.** vidíme ukázku výstupu jednoho CFD výpočtu. Na **Obrázku 1. 2.** je ukázka přibližné periodicity těchto časových řad.



Obrázek 1. 1.: Ukázka výstupu CFD výpočtu



Obrázek 1. 2.: Ukázka periody v časových řadách

Účinnost je měřena od 0-100% já se omezím na rozmezí 89-95%, protože v tomto rozmezí se mé časové řady pohybují. Z **Obrázku 1. 1.** je patrné, že ačkoliv má být děj periodický, CFD simulaci chvíli trvá, než se výpočet ustálí. Z praktického hlediska je tedy problém, kdy má člověk výpočet ukončit a děj považovat za již ustálený. Vzhledem k tomu, že je výpočet časově velice náročný, bylo by velmi výhodné mít nějaký matematický nástroj, který pomůže s odhadem času potřebného k ustálení výpočtu.

Cílem mé práce je proto pokusit se použít počáteční úseky, těchto 73 dlouhých časových řad a prozkoumat zda se dají použít jako prediktory koncových úseků časových řad. Pokud by bylo možné s vysokou pravděpodobností predikovat již na začátku výpočtu, na jaké hodnotě se výpočet asi ustálí, výrazně by to zkrátilo čas potřebný k celému optimalizačnímu procesu. Typicky se počítá okolo 100 variant a cílem je nalézt tu nejlepší, tedy takovou, která dosahuje nejvyšší účinnosti. Účinnost vítězné varianty se poté ověří přesnějším výpočtem.

Přiložená diplomová práce se skládá ze dvou kapitol. V první kapitole popíši vyhlazení časových řad. Následně se pokusím predikovat průměr posledních 10 hodnot vyhlazených časových řad pomocí dvou metod: shlukové metody k-means a metody vážené lineární regrese. V poslední části první kapitoly porovnám výsledky obou metod. Ve druhé kapitole se pokusím odhadnout pravděpodobnost toho, že časová řada bude kolísavá, tedy že se výpočet neustálí kolem rovnovážné hodnoty. K této předpovědi využiji metody logistické regrese.

K analýze výsledků jsem využívala programů **Matlab R2013, R** verze **i386 3.1.2** a funkcí v programu **Microsoft Office Excel 2007**

1 Model pro optimalizaci výpočtu účinnosti rotačních hydraulických strojů pomocí metod predikce v časových řadách

V následující kapitole budu predikovat průměr posledních deseti period výpočtu CFD simulace ze znalosti průběhu začátku této časové řady. V následující kapitole nejprve popíši vyhlazení dat. Vyzkouším dvě metody predikce, jednou pomocí shlukové metody k-means a jednou pomocí metody vážené lineární regrese. Nakonec této kapitoly porovnám výsledky obou způsobů.

1.1 Vyhlazování časových řad

Z **Obrázku 1. 1.** je patrné, že původní časové řady (3600 kroků) obsahují vysokofrekvenční složku s velkou amplitudou, která má periodu 18 kroků. To odpovídá otočení oběžného kola o jednu lopatku. Tato perioda v časové řadě pouze ruší, protože výsledná účinnost, která nás zajímá, je typicky brána jako průměr přes poslední dvě periody, tedy posledních 36 časových kroků. Proto nejprve u všech časových řad provedeme down-sampling, tedy definujeme novou časovou řadu délky 200, jejíž každý člen je průměr 18 členů původní časové řady. Označuji časovou řadu v původním rozlišení jako:

$$a = a_1, \dots, a_{3600.}$$

Vyhlazenou časovou řadu označuji jako časovou řadu průměrů následovně:

$$b = b_1, \dots, b_{200.}$$

kde

$$b_1 = \frac{1}{18}(a_1 + \cdots + a_{18})$$

$$b_2 = \frac{1}{18}(a_{19} + \cdots + a_{37})$$

$$\vdots$$

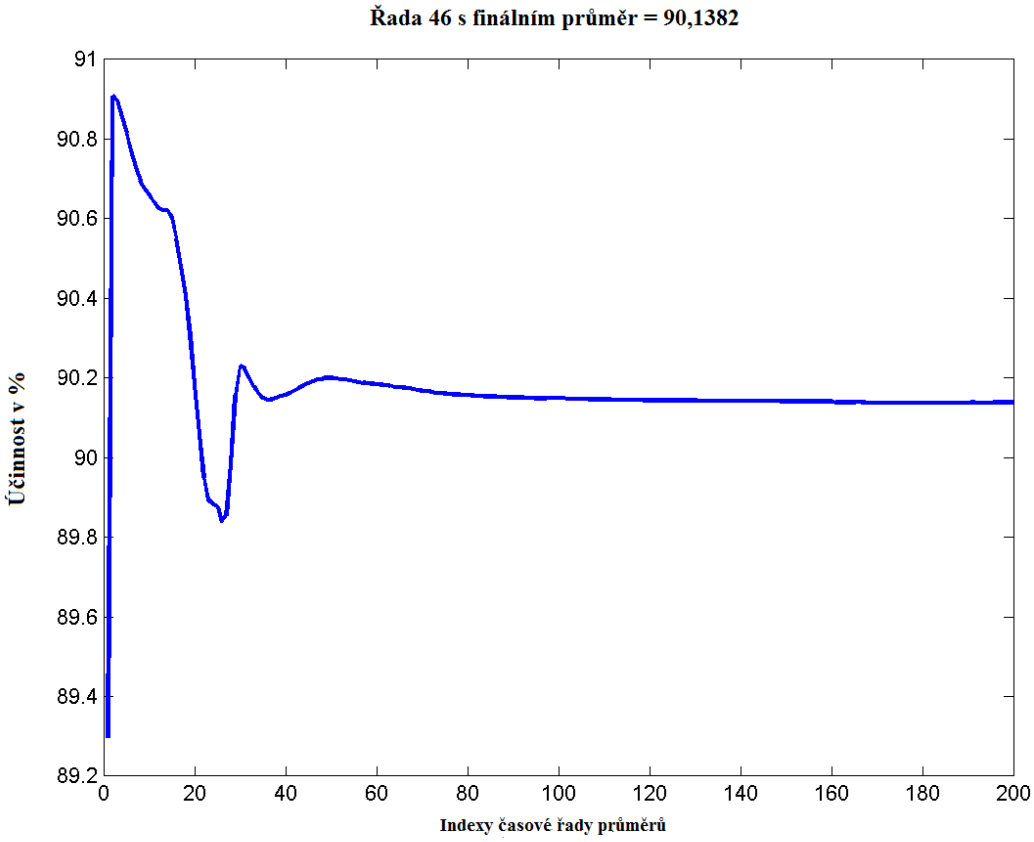
$$b_{200} = \frac{1}{18}(a_{3582} + \cdots + a_{3600})$$

Takto vyrovnané časové řady obsahují průměrné hodnoty přes (teoretické) periody původní časové řady. Cílem této kapitoly je predikovat průměr posledních deseti členů řady b pomocí průběhu jejího začátku. Chceme tedy predikovat číslo:

$$b_{190} = \frac{1}{10}(b_{190} + \cdots + b_{200})$$

pomocí čísel $b_1, \ldots, b_{80}$. Průměr posledních deseti period je robustnější ukazatel účinnosti, než pouhá konečná hodnota b_{200}. Prvních 80 datových bodů nazýváme „počáteční úsek“ časové řady b a 81 až dvoustý datový bod nazýváme „koncový úsek“ časových řad průměrů. Hodnota 80 je zvolena proto, že takto dlouhé jsou většinou výpočty, které se při rutinní optimalizaci tvaru provádějí. Odpovídá to 1440 časovým krokům v původní časové řadě.

Na **Obrázku 1. 1.** je vykreslena časová řada v původním rozlišení a a na **Obrázku 1. 3.** je tato časová řada vyhlazená neboli časová řada průměrů b.



Obrázek 1. 3.: Časová řada průměrů *b*

Vzhledem k tomu, že máme jen 73 vzorků časových řad, není možné jako prediktory finálního průměru použít přímo všech 80 počátečních bodů. To je příliš mnoho prediktorů na příliš málo vzorků. Proto se pokusíme chování časových řad na jejich začátku zachytit charakteristikami, kterých bude výrazně méně. Po visuální inspekci celé datové sady navrhuji na „počátečním úseku“ časových řad průměrů, neboli na jejich prvních 80-ti datových bodech, definovat těchto šest charakteristik.

Rozdíl prvních dvou datových bodů ($Char_1$):

$$Char_1 = b_2 - b_1$$

Rozdíl prvních dvou datových bodů počítám proto, že časové řady v prvních dvou bodech mají dramatické chování, někdy hodnota prudce stoupne a někdy prudce klesne. Na **Obrázku 1. 2.** je znázorněn rozdíl mezi prvními dvěma datovými body časové řady průměrů číslo 46 a je roven 28,5.

Minimální hodnota počátečního úseku časových řad průměrů ($Char_2$):

$$Char_2 = \min(b_i)\,, kde\ i = 1, \dots, 80$$

Maximální hodnota z počátečního úseku časových řad průměrů ($Char_3$):

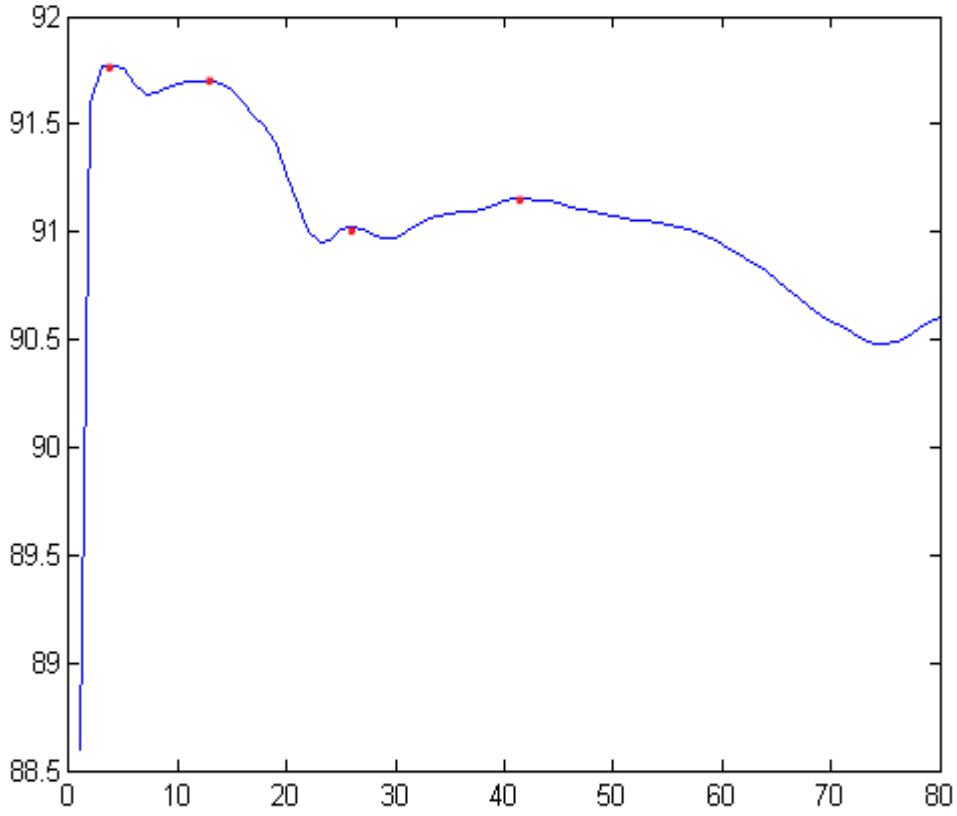
$$Char_3 = \max(b_i), kde\ i = 1, \dots, 80$$

Průměrná hodnota z počátečního úseku časových řad průměrů ($Char_4$):

$$Char_4 = \frac{1}{80} \sum_{i=1}^{80} b_i$$

Charakteristiky minim, maxim a průměrů z počátečního úseku časových řad počítám, protože těmito charakteristikami nejjednodušeji zachytím variabilitu v datech.

Počet vrcholů na počátečním úseku časových řad průměrů ($Char_5$): Tato charakteristika udává počet lokálních maxim počátečního úseku. Lokální maximum je definováno jakožto bod, v němž je hodnota větší než u jeho dvou sousedů. Lokální maxima vrací funkce *findpeaks* v Matlabu. Na **Obrázku 1. 4.** je zachycena časová řada se 4 vrcholy s hodnotami účinnosti z leva: 91,77%, 91,7%, 91,02% a 91,15%.



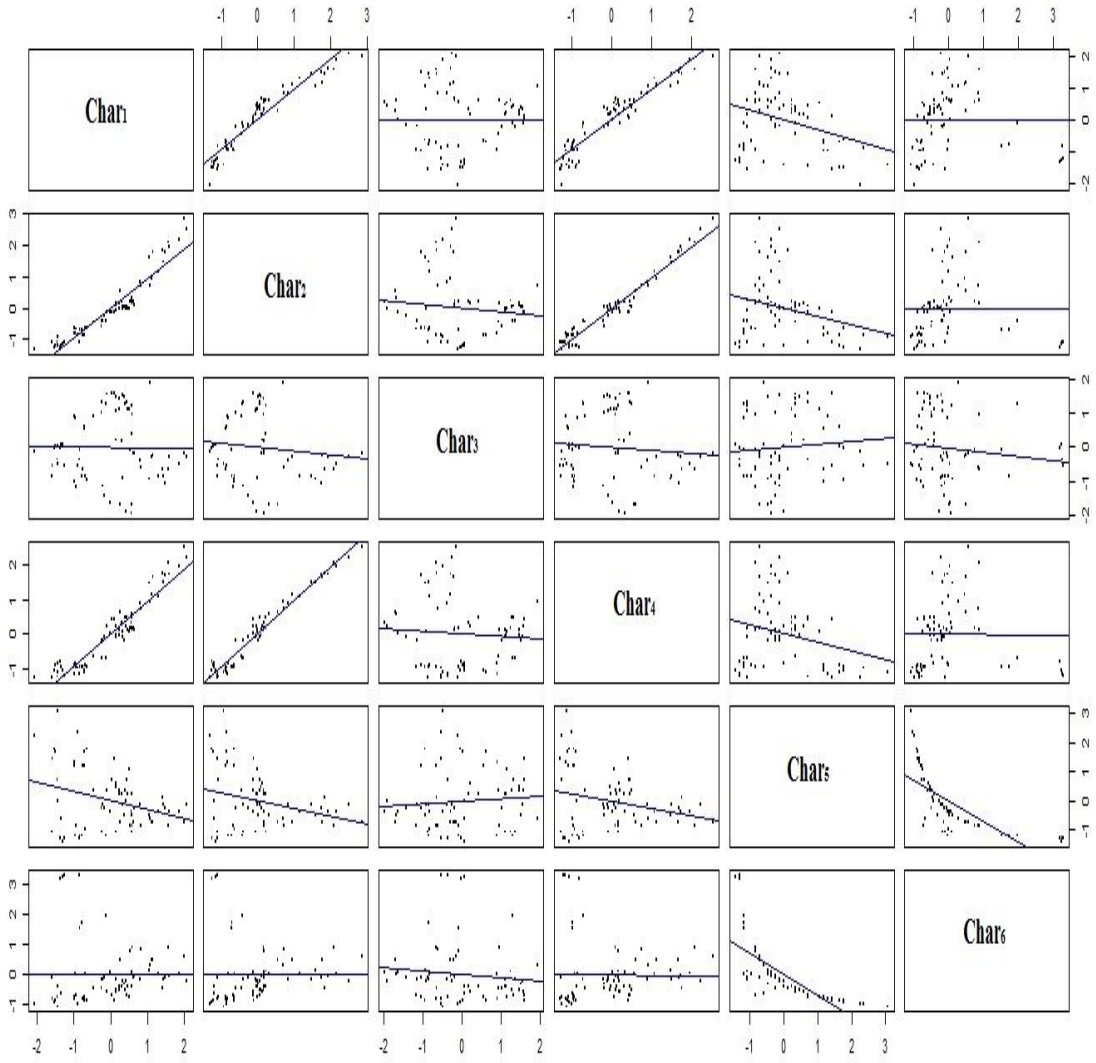
Obrázek 1. 4.: Vrcholy časové řady

Průměrná vzdálenost mezi vrcholy na počátečním úseku časových řad průměrů ($Char_6$): Tato charakteristika se počítá pomocí funkce *diff* v Matlabu, která vrátí vzdálenosti mezi vrcholy nalezenými v předchozí charakteristice.

Z charakteristik počátečního úseku časových řad ($Char_1$-$Char_6$), chci predikovat průměr posledních deseti hodnot časové řady průměrů, který označíme y a který považujeme za vysvětlovanou proměnnou.

$$y = \frac{1}{10}(b_{191} + b_{192} + \cdots + b_{200})$$

Nejprve je dobré se podívat, jak spolu jednotlivé charakteristiky souvisejí. Na **Obrázku 1. 5.** je grafická reprezentace vzájemného vztahu dvojic.



Obrázek 1. 5.: Vzájemné vztahy mezi charakteristikami

Z obrázku vyplívá, že pokud se změní $Char_1$ změní se i charakteristiky $Char_2$ a $Char_4$, protože mají mezi sebou téměř lineární vztah. Což znamená, že jejich korelační koeficient je téměř roven 1. Stejný vztah mají mezi sebou $Char_2$ a $Char_4$. Mezi $Char_5$ a $Char_6$ je hyperbolický vztah, což znamená, že jakákoliv změna nezávisle proměnné, dost ovlivní závisle proměnnou.

1.2 Predikce průměru posledních deseti hodnot časových řad průměrů pomocí metody k-means

V této části uvažuji všech 73 časových řad průměrů. Z každé časové řady průměrů extrahuji mnou definovaných šest charakteristik ($Char_1$-$Char_6$). Následně na takto vytvořených, 6-ti rozměrných vektorech najdu shluky pomocí metody k-means. Predikce bude probíhat takto: když přijde nová časová řada průměrů, dle počátečního úseku ji zařadím do správného shluku. Následně budu konec této časové řady předpovídat typickým koncem časových řad v tomto shluku. Předpovídat typickým koncem znamená, že v rámci každého shluku se podívám na posledních deset hodnot časové řady průměrů a udělám jejich medián. Tento medián je hodnota, kterou budu predikovat.

1.2.1 Metoda k-means

Metoda k-means si klade za cíl rozdělit n datových bodu do k-shluků. V mém případě odpovídá n-tý datový bod vektoru charakteristik ($Char_1$,…,$Char_6$), vypočítaných z počátečního úseku n-té časové řady průměrů. Přeznačme pro jednoduchost zápisu nyní datovou sadu ($\mathbf{x_1}$,…,$\mathbf{x_n}$), kde každé pozorování x_i je d rozměrný reálný vektor. V našem případě je tedy $d = 6$ a každé x_i i odpovídá vektoru šesti charakteristik jedné časové řady Metoda k-means rozděluje těchto n pozorování do $k \leq n$ shluků. Jejím výsledkem jsou množiny S_1,…, S_k, kde S_i obsahuje datové body náležející ke shluku i. Rozdělení probíhá tak, aby byl vnitřní skupinový součet čtverců (WWS) minimální, neboli hledáme

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

kde $\boldsymbol{\mu}_i$ je střední hodnota bodů ve shluku S_i a S je systém všech možných dělení datové sady na k -shluků. Nejběžnější používaný algoritmus vypadá následovně:

Algoritmus:

Na počátku vybereme počáteční polohu středů shluků $(\boldsymbol{\mu}_1^0, \ldots, \boldsymbol{\mu}_k^0)$. Například mohu vybrat k-náhodně zvolených datových bodů z původní datové sady. Následně pokračuje algoritmus opakováním následujících dvou kroků.

1. Klasifikace: přiřazuje pozorování do odpovídajícího shluku. Tedy přiřadí pozorování x_i k tomu shluku, jehož střed je danému pozorování nejblíže.

2. Přepočítávání: vypočítají se nové polohy středů shluků, jakožto střední hodnoty těch datových bodů, které byli klasifikovány do daného shluku.

Algoritmus skončí, jakmile v kroku 1 žádný datový bod nezmění svou příslušnost. Neexistuje žádná záruka, že bude tímto algoritmem nalezeno globální optimum. Výsledek závisí na volbě počátečních podmínek.

Do programu **R** je metoda k-means implementována pomocí funkce *kmeans*. Do funkce *kmeans* se zadávají d rozměrné vektory a předem zvolený počet shluků k. Následně algoritmus rozdělí pozorování do shluků tak, aby se minimalizoval vnitřní skupinový součet čtverců.

Já udělám na charakteristikách ($Char_1$-$Char_6$) z počátečního úseku časových řad průměrů (73 časových řad průměrů) shlukovou analýzu s využitím metody k-means pro předem zadaný počet shluků 2, 3, 4, 5, 6, 7, 8 a 9. Optimální počet shluků určuji dvěma způsoby:

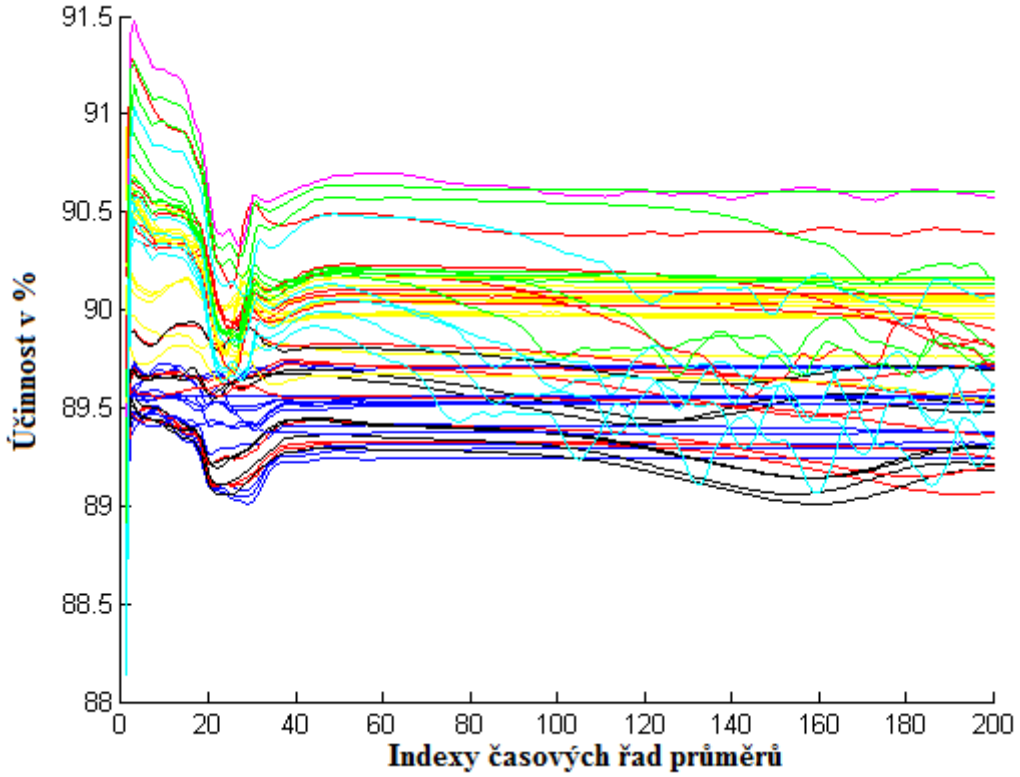
(1) Pomocí vykreslení závislostí vnitřního skupinového součtu čtverců na počtu shluků.

(2) Nalezením počtu shluků, kde celková chyba predikce je co nejmenší a zároveň shluků není příliš mnoho. Celková chyba predikce je definovaná takto:

$$J = \sum_{i=1}^{73} váha\ |predikce - skutečnost|.$$

Podrobnější popis této metody následuje níže.

Pro ilustraci shlukové metody jsme nejdříve do **Obrázku 1. 6.** vykreslila všech 73 časových řad průměrů v rozmezí účinnosti 87% - 93%. Pomocí funkce *k-means* jsem provedla shlukovou analýzu pro předem zadaný počet devíti shluků, na charakteristikách vypočítaných z počátečního úseku časových řad průměrů. Následně jsem pomocí devíti barev označila na **Obrázku 1. 6.** jednotlivé shluky. Z obrázku je patrné, že je jistá šance, že tento postup predikce bude fungovat, protože koncové úseky řad stejného shluku se skutečně chovají poměrně podobně.



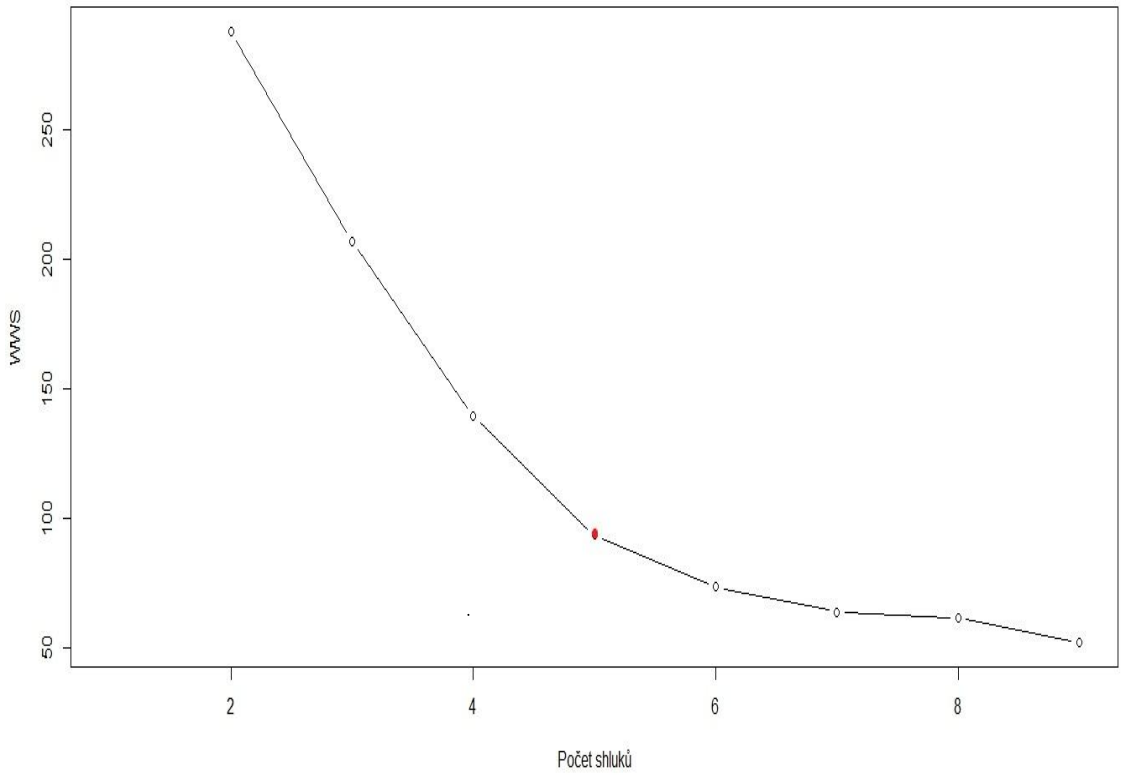
Obrázek 1. 6.: Výsledky shlukové analýzy podle charakteristik z počátečního úseku časových řad průměrů

1.2.2 Varianta 1 výpočtu optimálního počtu shluků:

Na charakteristiky ($Char_1$ – $Char_6$), vypočítaných z počátečního úseku časových řad průměrů, aplikuji metodu k-means pro předem zadaný počet shluků 2, 3, 4, 5, 6, 7, 8 a 9. Optimální počet shluků budu hledat pomocí vykreslení vnitřního skupinového součtu čtverců:

$$WWS = \min_{S} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

Tato varianta vychází pouze z datových bodů a nebere v potaz predikce. Vnitřní skupinový součet čtverců vykreslím pro počty shluků: 2, 3, 4, 5, 6, 7, 8 a 9 do **Obrázku 1. 7.**. Optimální počet shluků se nachází tam kde graf WSS přestává prudce klesat v tzv. *„zlomu"*. Dle obrázku by se dalo říct, že optimální počet shluků je 5, ačkoliv zlom není nijak výrazný.



Obrázek 1. 7.: Optimální počet shluků využitím Varianty 1

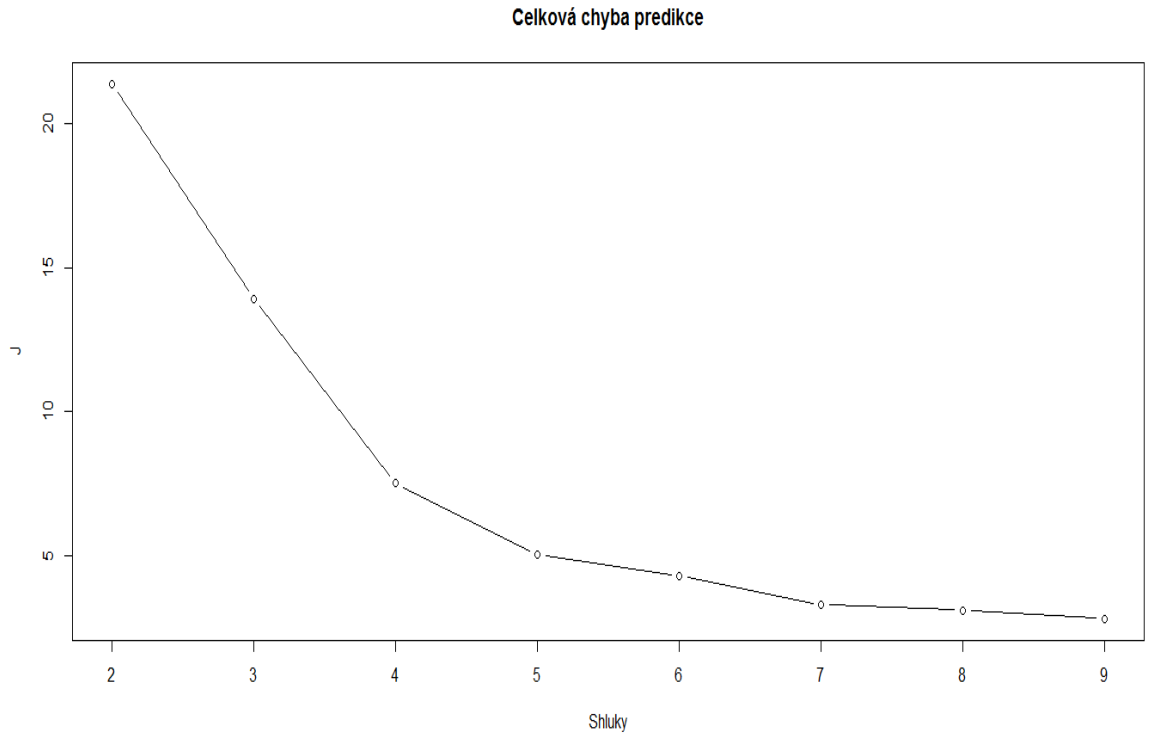
1.2.3 Varianta 2 výpočtu optimálního počtu shluků:

Tato varianta bere v potaz predikci a snaží se počet shluků optimalizovat, tak aby predikce byla co nejlepší. Pro daný počet shluků 2, 3, 4, 5, 6, 7, 8, 9, aplikuji metodu k-means a dle výsledků rozdělím časové řady do shluků. V každém shluku udělám predikci. Predikci udělám mediánem z charakteristiky y, průměru posledních 10 hodnot časových řad, které patří do příslušného shluku.

Pro každou časovou řadu máme tedy skutečnou finální hodnotu a predikovanou hodnotu. Následně hledám rozumný počet shluků, který je takový, kde je celková chyba predikce co nejmenší, ale shluků není příliš mnoho. Celková chyba predikce:

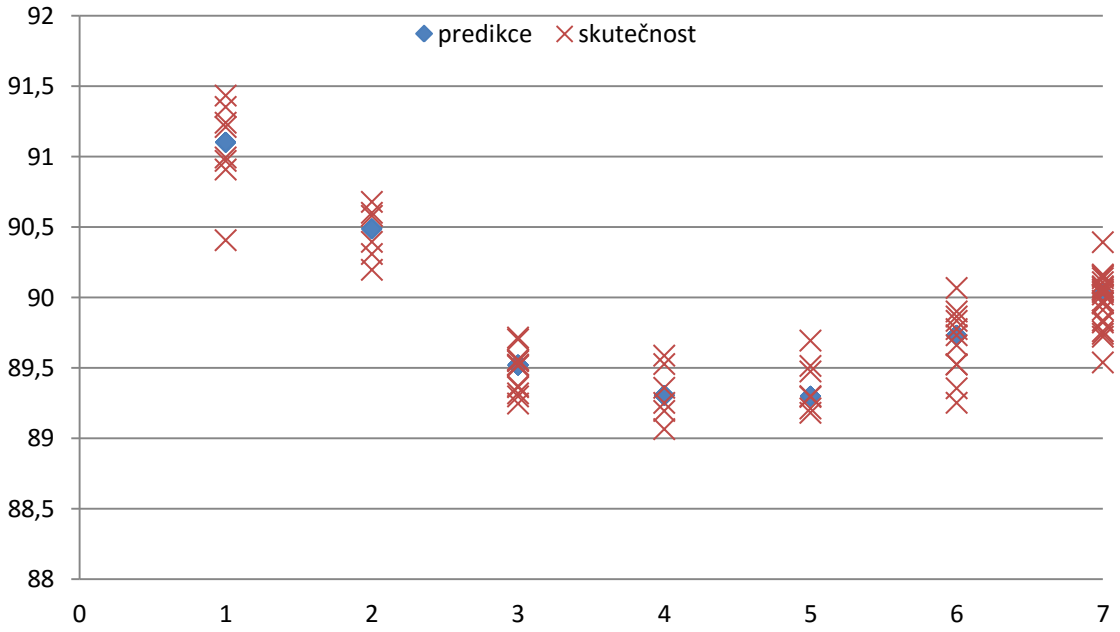
$$J = \sum_{i=1}^{73} (predikce - skutečnost)^2.$$

Na **Obrázku 1. 8.** je vykreslena celková chyba předpovědi (J).



Obrázek 1. 8.: Celková chyba předpovědi

Z **Obrázku 1. 8.** vyplývá, že rozumný počet shluků dle míry chybovosti (J) je pět až sedm shluků. Na **Obrázku 1. 9.** jsou výsledky predikce (průměru posledních 10 hodnot časových řad průměrů) proti skutečným hodnotám pro rozumný počet shluků sedm.



Obrázek 1. 9.: Výsledky predikce oproti skutečným hodnotám pro rozumný počet shluků 7

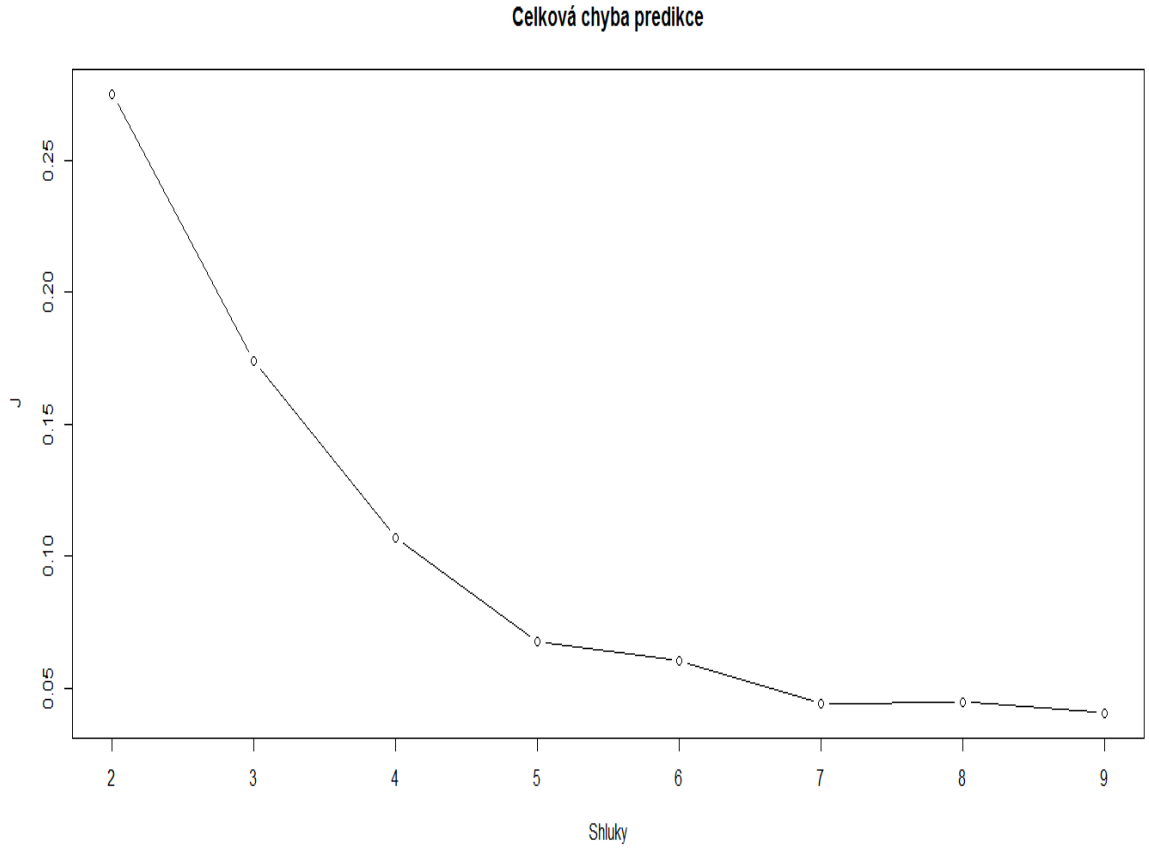
Na **Obrázku 1. 6.** jsou vidět dva typy časových řad průměrů:

1. Ustálená časová řada průměrů
2. Kolísavá časová řada průměrů

U časových řad průměrů, které se ustálí, je predikovaná hodnota celkem jistá. U časových řad průměrů, které jsou na svém konci kolísavé, tam predikovaná hodnota tak jistá není. Z toho důvodu postup výpočtu predikce provedu znovu. Tentokrát nebudu počítat celkovou chybu predikce, ale vypočítám váženou celkovou chybu predikce. Váha je tím větší, čím bude časová řada průměrů „hezčí“. Časová řada průměrů je tím hezčí, čím má nižší variabilitu svého konce. Variabilitu konce časových řad průměrů počítám jako rozptyl posledních 10 hodnot časových řad průměrů. Váhu počítám jako převrácenou hodnotu těchto rozptylů. Váženou celkovou chybu predikce definuji (J_2):

$$J_2 = \sum_{i=1}^{73} váha * (predikce - skutečnost)^2.$$

Na **Obrázku 1. 10.** jsou výsledky vážené celkové chyby predikce (J_2).



Obrázek 1. 10.: Celková chyba predikce J_2

Z **Obrázku 1. 10.** vyplývá, že rozumný počet se shluků se nezměnil, a můžeme opět vzít pět, šest nebo sedm shluků. Obrázek celkové chyby predikce se moc nezměnil z důvodu ne příliš velké významnosti vah. Rozptyl na **Obrázku 1. 6.** u některých řad vypadá, že je významný. Ve skutečnosti se však jedná o rozptyl v řádu desetin procenta účinnosti a z toho důvodu vychází váhy velmi malé.

Numerické porovnání celkových chyb predikce pro rozumné počty shluků:

- Pro pět shluků, zvolených jak variantou jedna tak variantou 2, je celková vážená chyba předpovědi dle vztahu

$$J_2 = \sum_{i=1}^{73} váha * (predikce - skutečnost)^2 = 0{,}067$$

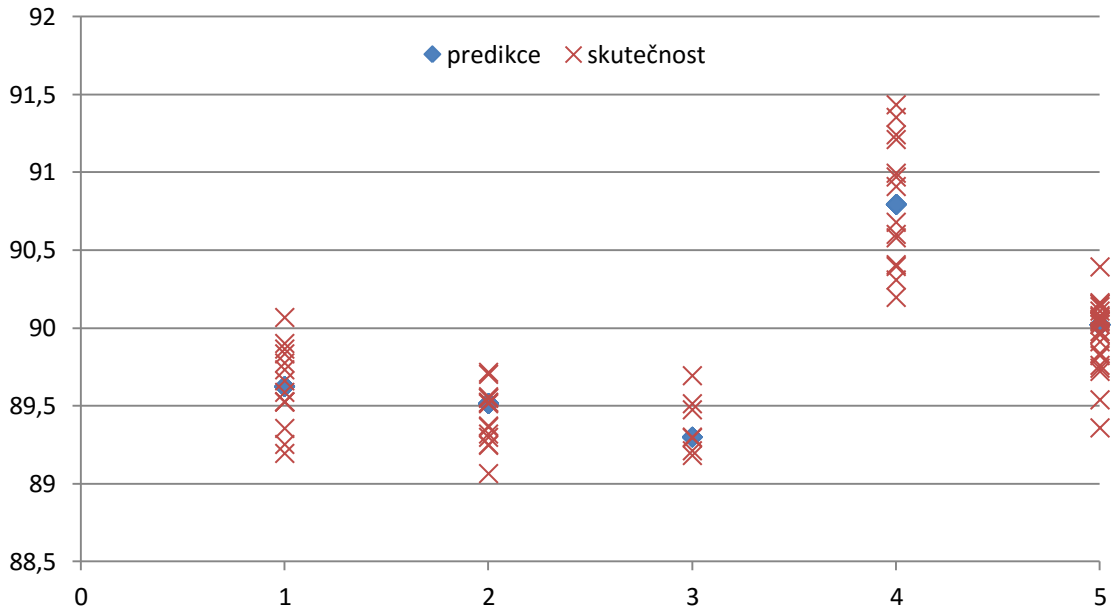
- Pro šest shluků, zvolených variantou dvě, je celková vážená chyba předpovědi dle vztahu:

$$J_2 = \sum_{i=1}^{73} váha * (predikce - skutečnost)^2 = 0{,}062$$

- Pro sedm shluků, zvolených variantou dvě, je celková vážená chyba předpovědi dle vztahu

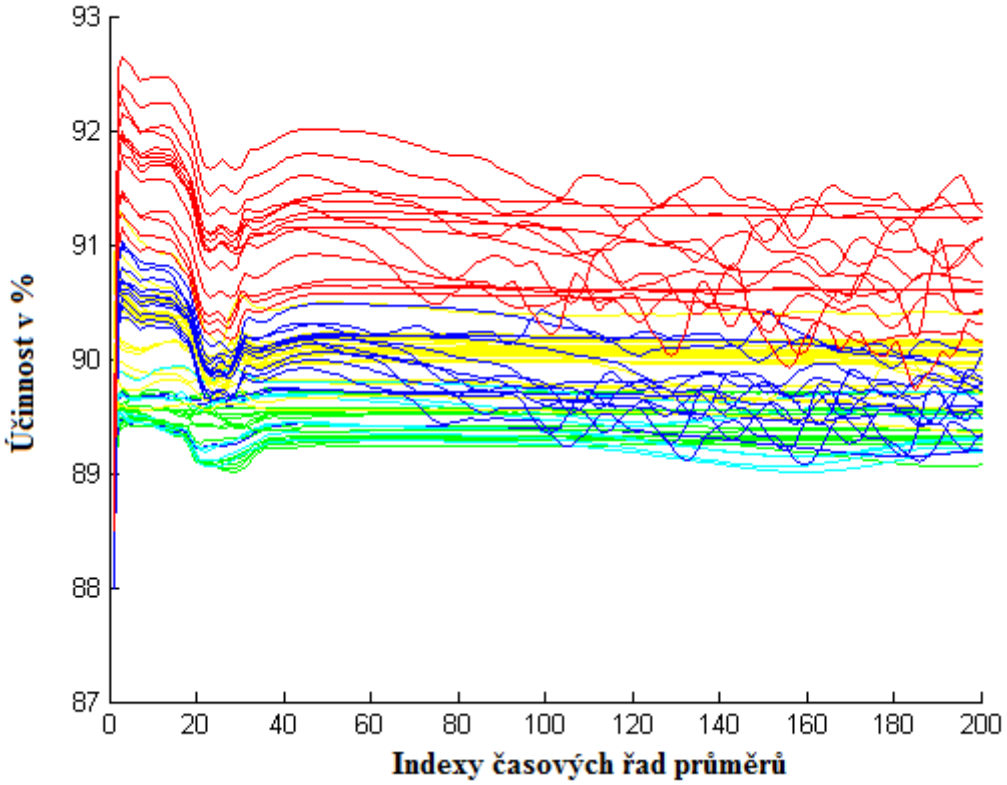
$$J_2 = \sum_{i=1}^{73} váha * (predikce - skutečnost)^2 = 0{,}044$$

Dle kvantitativního porovnání vychází o něco lépe rozumný počet shluků pět, který jsem stanovila jak metodou číslo 1, tak i metodou číslo 2. Zde je počet shluků rozumný a zároveň shluků není příliš mnoho. Na **Obrázku 1. 11.** jsou vykresleny výsledky predikce oproti skutečným hodnotám průměru posledních 10 hodnot časových řad průměrů, pro rozumný počet shluků pět.



Obrázek 1. 11.: Výsledky predikce oproti skutečným hodnotám pro optimální počet shluků 5

Na **Obrázku 1. 12.** jsou vykresleny výsledky shlukování pro rozumný počet shluků pět. Použila jsem pět barev pro vykreslení jednotlivých shluků.



Obrázek 1. 12.: Výsledky shlukování pro optimální počet shluků 5

Z **Obrázku 1. 12.,** vidíme, že kolísavé časové řady se zařadili pouze do dvou shluků z pěti možných. Z toho vyplývá, že časové řady, které se neustálí, mají pouze dva typy průběhů na svém počátečním úseku.

1.3 Predikce průměru posledních deseti hodnot časových řad průměrů metodou vážené lineární regrese

V této části budu predikovat průměr posledních 10 hodnot časových řad průměrů, využitím vážené lineární regrese, aplikované na charakteristiky počátečního úseku časových řad průměrů.

Stanovuji model:

$$y = \beta_0 + \beta_1 Char_1 + \beta_2 Char_2 + \beta_3 Char_3 + \beta_4 Char_4 + \beta_5 Char_5 + \beta_6 Char_6$$

kde $Char_1$,…, $Char_6$ jsou prediktory:

- y… průměr posledních 10 hodnot časových řad průměrů
- $Char_1$… minimum z počátečního úseku časových řad průměrů
- $Char_2$… maximum z počátečního úseku časových řad průměrů
- $Char_3$… průměr z počátečního úseku časových řad průměrů
- $Char_4$… počet vrcholů z počátečního úseku časových řad průměrů
- $Char_5$…průměrná vzdálenost mezi vrcholy z počátečního úseku časových řad průměrů
- $Char_6$… sklon počátečního úseku časových řad průměrů

Využívám k odhadu parametrů $\beta_0 - \beta_6$ váženou metodu nejmenších čtverců (VMNČ). Vážená metoda nejmenších čtverců požaduje, aby součet čtverců (druhých mocnin) naměřených hodnot $Char_{7i}$ a funkčních hodnot $f(Char_{1i}, \dots, Char_{6i})$ s váhou w_i pro stejné hodnoty $(Char_{1i}, \dots, Char_{6i})$ byl co nejmenší. Uvažuji $(Char_{1i}, \dots, Char_{6i})$, $i = 1, \dots, 73$ a váhy w_i, kde i je z intervalu $\langle 0,1 \rangle$. Váhy vyjadřují variabilitu v datech a jsou vypočítány jako převrácená hodnota rozptylu posledních deseti hodnot časových řad průměrů. Funkce $f(Char_{1i}, \dots, Char_{6i})$ je funkce s neznámými parametry $(\beta_0; \beta_1; \dots; \beta_6)$, které je třeba odhadnout z dat.

$$VMNČ = \sum_{i=1}^{73} w_i \, (Char_{7i} - (\beta_0 + \beta_1 Char_1 + \beta_2 Char_2 + \beta_3 Char_3 + \beta_4 Char_4 + \beta_5 Char_5 + \beta_6 Char_6))^2$$

V programu **R** je funkce vážené lineární regrese implementována funkcí *lmw*. Jako argumenty se do této funkce zadává model ve tvaru $Y \sim X_1 + \cdots + X_n$, dále se zadává vektor vah. Model počítá koeficienty metodou vážených nejmenších čtverců.

Následně jsem na data charakteristik $Char_1, \ldots, Char_7$, z počátečního úseku časových řad průměrů, aplikovala funkci *lmw* v programu **R**. Výsledky odhadu koeficientů metodou vážené lineární regrese jsou v **Tabulce 1. 1.**

Tabulka 1. 1.: Výsledky odhadu koeficientů metodou vážené mnohonásobné lineární regrese

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr (>|t|)**	
$\boldsymbol{\beta_0}$	0.9875325	2.7692411	0.357	0.7225	
$\boldsymbol{\beta_1}$	-0.0008372	0.0008258	-1.014	0.3144	
$\boldsymbol{\beta_2}$	1.0109869	0.0360288	28.060	<2e-16	***
$\boldsymbol{\beta_3}$	-0.0054558	0.0091614	-0.596	0.5535	
$\boldsymbol{\beta_4}$	-0.0166750	0.0161126	-1.035	0.3045	
$\boldsymbol{\beta_5}$	0.0012484	0.0010329	1.209	0.2312	
$\boldsymbol{\beta_6}$	-0.0031402	0.0017492	-1.795	0.0772	.

Značení významnosti regresoru: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tato metoda nám oproti metodě k-means z odhadů parametrů dokáže říci na, které proměnné a jak moc předpověď závisí. Z **Tabulky 1. 1.** vyplývá, že významné jsou pouze ukazatele $Char_2$ (minimum) a $Char_6$ (průměrná vzdálenost mezi vrcholy), z toho důvodu všechny ostatní charakteristiky z modelu vyjmu. Vytvořím redukovaný model, který vypadá následovně:

$$y = \beta_0 + \beta_2 Char_2 + \beta_6 Char_6$$

Na data charakteristik $Char_2, Char_6$ *a* y znovu aplikuji funkci lmw v programu R. Výsledky odhadu koeficientů metodou vážené lineární regrese pro redukovaný model jsou v **Tabulce 1. 2.**

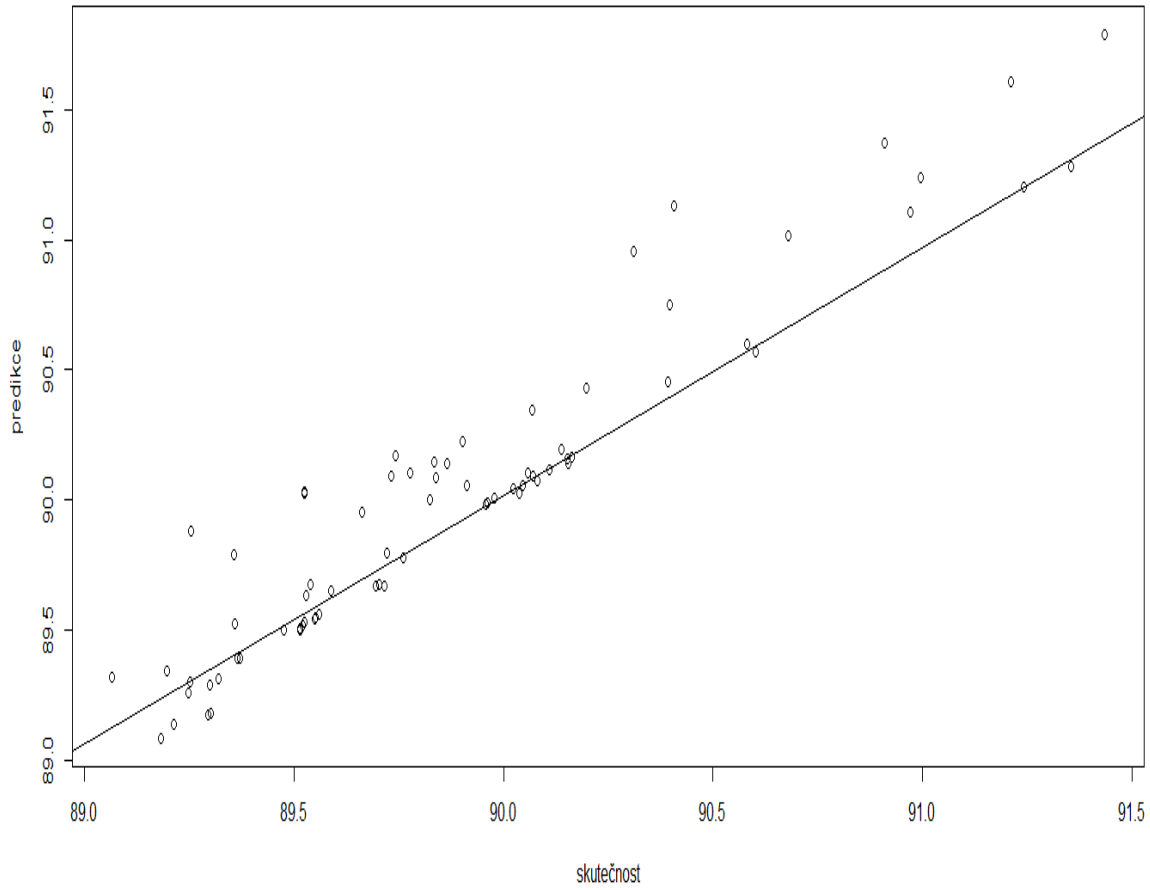
Tabulka 1. 2.: Výsledky odhadu koeficientů metodou vážených nejmenších čtverců pro redukovaný model

Coefficients:	**Estimate**	**Std. Error**	**t value**	**Pr (>|t|)**	
$\boldsymbol{\beta_0}$	4.157141	0.845264	4.918	5.57e-06	***
$\boldsymbol{\beta_2}$	0.953978	0.009493	100.491	<2,00E-16	***
$\boldsymbol{\beta_6}$	-0.005112	0.001250	-4.090	0.000114	***

Značení významnosti regresorů: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Z **Tabulky 1. 2.** vyplývá, že všechny parametry redukovaného modelu jsou významné. Tento redukovaný model použiji k predikci průměru posledních 10 hodnot časových řad průměrů. Predikci provedu dosazením hodnot časových řad průměrů do redukovaného modelu. V programu **R** získám predikované hodnoty příkazem *predict*, kam

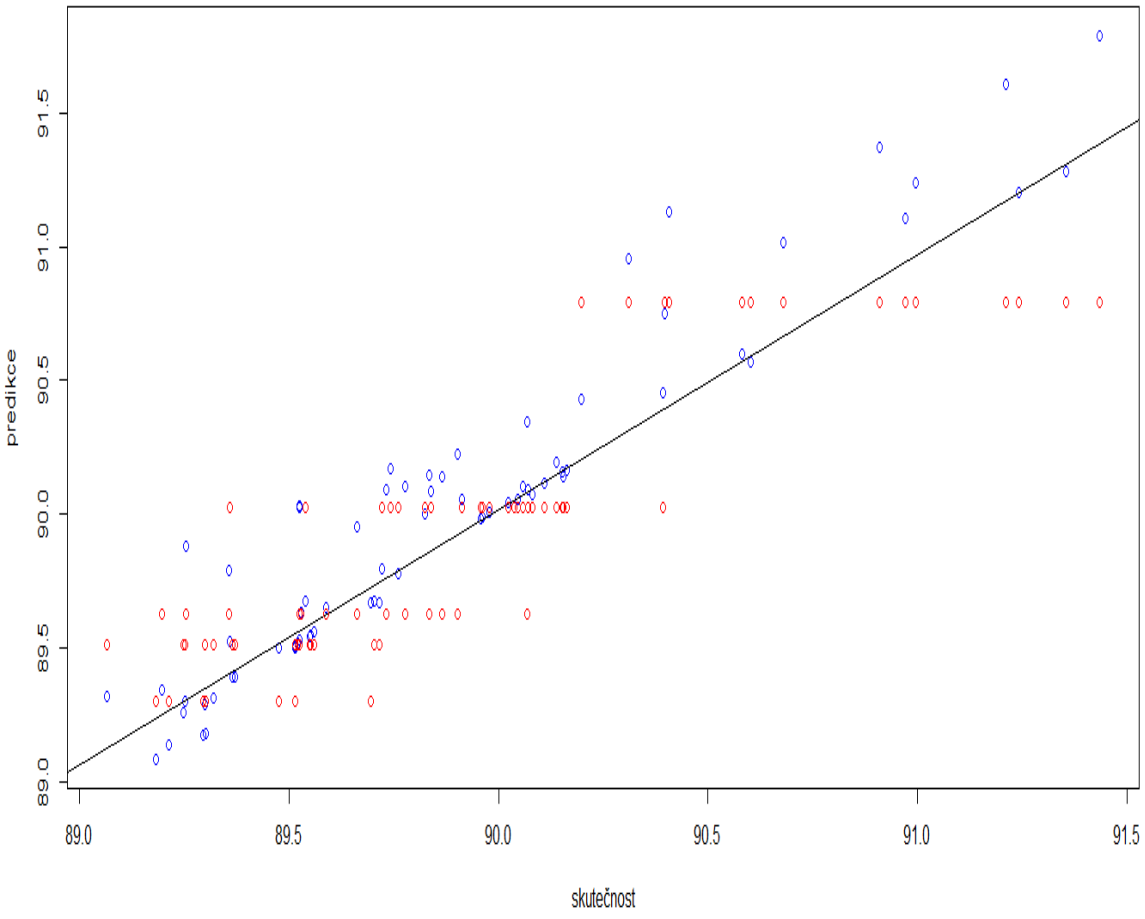
se jako argument zadává stanovený model. Výsledky predikce jsou vidět na **Obrázku 1. 13.**



Obrázek 1. 13.: Výsledky predikce průměru posledních 10 hodnot časových řad průměrů metodou vážené lineární regrese

1.4 Srovnání predikce metodou k-means a metodou vážené lineární regrese

V této části porovnám výsledky predikce průměru posledních 10 hodnot časových řad průměrů, které jsem počítala dvěma metodami a to: metodou k-means a metodou vážené lineární regrese. U metody k-means volím rozumný počet shluků 5. Na **Obrázku 1. 14.** je vyobrazeno porovnání výsledků výše zmíněných metod.



Obrázek 1. 14.: Porovnání výsledků predikce posledních 10 hodnot časových řad průměrů metodou k-means a metodou vážené lineární regrese

Z **Obrázku 1. 14.** vyplývá, že lepší výsledky predikce průměru posledních 10 hodnot časových řad průměrů vychází z metody vážené lineární regrese.

Reziduální součet čtverců (RSČ), neboli váženou chybu regrese definuji:

$$RSČ = \sum_{i=1}^{73} váha_i * (predikce - skutečnost)^2.$$

Reziduální součet čtverců ($RSČ_1$), neboli vážená chyba, pro výsledky predikce průměru posledních 10 hodnot časových řad průměrů metodou k-means je rovna:

$$RSČ_1 = 0{,}068$$

Reziduální součet čtverců ($RSČ_2$), neboli vážená chyba, pro výsledky predikce průměru posledních 10 hodnot časových řad průměrů metodou vážené lineární regrese je rovna:

$$RSČ_2 = 0{,}059$$

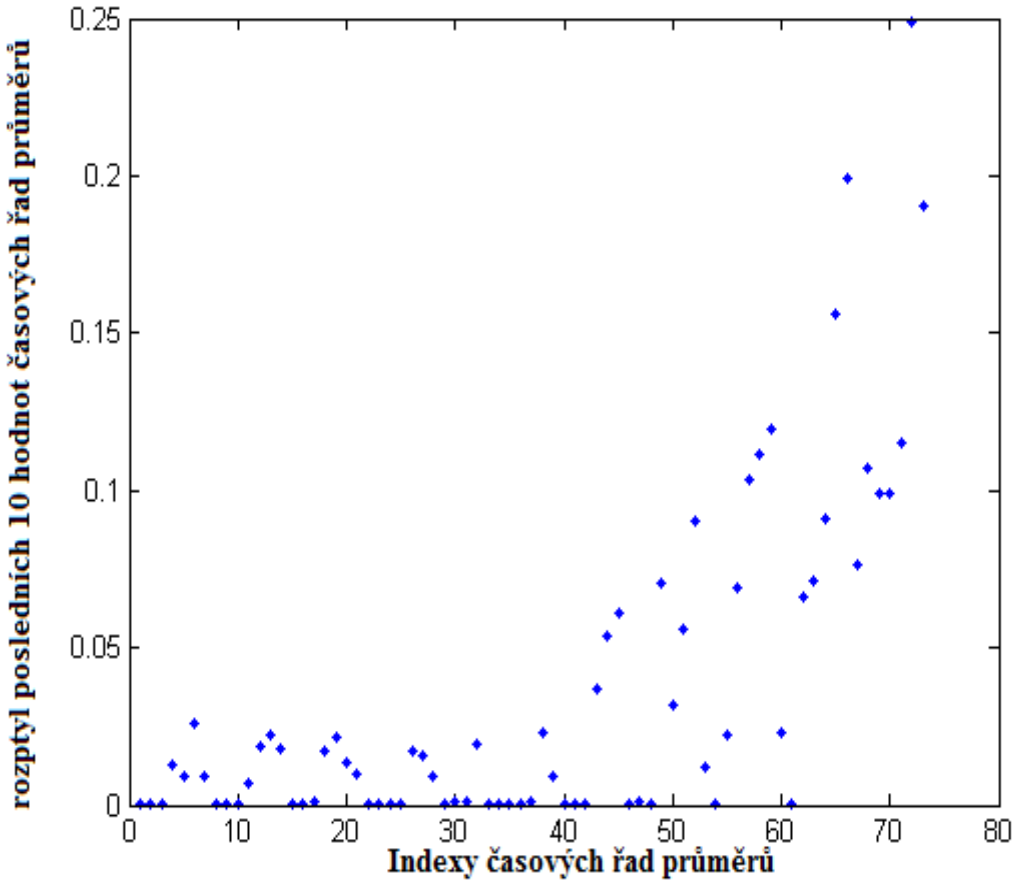
Srovnání reziduálního součtu čtverců obou metod:

$$RSČ_2 < RSČ_1$$

Podle výsledků hodnot reziduálních součtů čtverců jsou lepší výsledky predikce průměru posledních 10 hodnot časových řad průměrů získané metodou vážené lineární regrese.

2 Model předpovědi pravděpodobnosti ustálení časových řad průměrů za využití logistické regrese

V této kapitole budu predikovat, zda časová řada průměrů skončí ustálená nebo se neustálí a na svém konci bude kolísavá. K této predikci využiji model logistické regrese. Za klasifikátor, který určuje, zda řada skončí ustálená či kolísavá, použiji rozptyl posledních deseti hodnot časových řad průměrů. Jestliže je rozptyl těchto posledních deseti hodnot časových řad průměrů < 0,03 tak je binární hodnota 0 (neustálí se). Jestliže je rozptyl posledních deseti hodnot časových řad průměrů > 0,03, tak je binární hodnota rovna 1 (ustálí se). Stanovení hranice 0,03 vychází z rozložení variability posledních dvou hodnot časových řad průměrů. Tato variabilita je vykreslena na následujícím **Obrázku 2. 1.**



Obrázek 2. 1.: Rozptyl posledních deseti hodnot časových řad průměrů

2.1 Logistický regresní model

Logistická regrese se používá v případě, kdy závisle proměnná Y je dichotomická náhodná veličina. Dichotomická náhodná veličina nabývá pouze dvou hodnot 0 (nepravda) nebo 1 (pravda). V rámci této kapitoly se budeme zabývat vztahem, který umožňuje modelovat pravděpodobnost pro závisle proměnnou Y pomocí nezávislých náhodných veličin X_1, …, Xn.

Nechť náhodná veličina $Y \sim Alt(\vartheta)$, neboli závisle proměnná Y má alternativní rozdělení s parametrem ϑ, který nabývá hodnot $0 < \vartheta < 1$, což znamená:

$$P(Y = y) = \begin{cases} \vartheta, & y = 1 \\ 1 - \vartheta, & y = 0 \\ 0, & jinak \end{cases}$$

Tuto pravděpodobnost zapisujeme následovně:

$$P(Y = y) = \vartheta^y (1 - \vartheta)^{1-y},$$

pro $y = (0,1)$. Přímo lze vypočítat střední hodnotu i rozptyl veličiny Y následovně:

$$EY = \vartheta \qquad var(Y) = \vartheta\,(1 - \vartheta).$$

Budeme modelovat pravděpodobnost toho, že závisle proměnná Y nabude hodnoty 1 (pravda). Jelikož uvažujeme případ kdy závisle proměnná Y je dichotomická, platí tedy:

$$P(Y = 1) = 1 - P(Y = 0)$$

Tato pravděpodobnost nabývá hodnot z intervalu (0,1). V tuto chvíli nemůžeme vytvořit model:

$$P(Y = 1) = \beta_0 + \beta_1 X_1 + \ ... + \beta_m X_m,$$

protože bychom získali odhady koeficientů $\beta_0, \beta_1, \dots, \beta_m$ takové, které by pro určité realizace $x_1, \dots, x_m$ veličin $X_1, \dots, X_m$ nabývaly hodnot mimo interval (0,1). A to z důvodu, že lineární prediktor může nabývat všech hodnot z R pokud není konstantní. Z tohoto důvodu zavádíme pojem šance (tzv. odds) jako podíl

$$odds\left(P(Y = 1)\right) = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)}.$$

Což vyjadřuje kolikrát je vyšší pravděpodobnost, že závisle proměnná Y nabude hodnoty 1, než pravděpodobnost, že nabude hodnoty 0. Šance leží v intervalu $(0, \infty)$. Následně je potřeba transformovat interval $(0, \infty)$ na interval $(-\infty, \infty)$. Z tohoto důvodu zavádíme logitovou funkci, která využívá přirozeného logaritmu.

$$logit\left(P(Y = 1)\right) = \ln\left(odds\left(P(Y = 1)\right)\right) = \ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right)$$

Tuto pravděpodobnost modelujeme obdobně jako je tomu u lineární regrese a to následovně:

$$ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m,$$

vyjádříme pravděpodobnost:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m)}}.$$

Označíme-li $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$ a $\boldsymbol{X} = (1, X_1, \ldots, X_m)'$, můžeme pravděpodobnost přepsat jako:

$$P(Y = 1) = \frac{1}{1 + e^{-X'\beta}}.$$

Pro různé realizace x náhodného vektoru X nabývá tato pravděpodobnost různých hodnot, z tohoto důvodu je to podmíněná pravděpodobnost. Všimněme si

$$P(Y = 1) = \vartheta = EY,$$

což znamená, že modelem predikujeme střední hodnotu náhodné veličiny Y v závislosti na realizacích x. Máme-li soubor o rozsahu n. Následně je potřeba získat odhady koeficientů modelu $\beta = (\beta_0, \beta_1, \ldots, \beta_m)$, které získáme pomocí metody maximální věrohodnosti. K získání koeficientů modelu β musíme definovat logistickou funkci jako:

$$f(x', \beta_0, \beta) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \cdots + \beta_m)}}$$

kde parametry $\beta_0, \beta \in R$. Oborem hodnot logistické funkce je interval $(0,1)$. Logistická funkce umožňuje popsat logistický regresní model:

$$P(Y = 1 \left| X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

s jedním prediktorem. Parametr β_0 určuje posunutí logistické křivky podél osy x a parametr β_1 určuje strmost křivky v okolí bodu $\left[-\frac{a}{b}; \frac{1}{2}\right]$.

2.1.1 Metoda maximální věrohodnosti

Předpokládáme náhodný vektor $X = (X_1, \dots, X_n)'$, jehož složky tvoří náhodný výběr pocházející z rozdělení s hustotou $f(x|\theta)$, kde $x \in R^n, \theta = (\theta_1, \dots \theta_n)' \in \Omega$ je vektorem parametrů charakterizujících toto rozdělení. Kde $f(x|\theta)$ pochází z nějakého systému hustot $\{f(x|\theta); \theta \in \Omega\}$, který je vektorem parametrů $\theta \in \Omega \underline{\subset} R^m$ jednoznačně určen. Jelikož $X_1, \dots, X_n$ tvoří náhodný výběr, je sdružená hodnota pravděpodobnosti rovna:

$$f(x|\theta) = f(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Kde se jedná o funkci parametrů $x_1, \dots, x_n$, kde vektor parametrů θ je fixovaný. Definujme věrohodností funkci:

$$\boldsymbol{L}(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta),$$

kde je ovšem naší proměnnou je θ a x představuje parametr. Cílem metody maximální věrohodnosti je pro dané realizace x vektoru X (tj. pro n nezávislých naměřených hodnot stejné proměnné X) takový odhad vektoru parametrů θ, který bude maximalizovat věrohodností funkci.

Odhad $\widehat{\theta_{ML}} \in \Omega$ nazveme maximálně věrohodným odhadem právě tehdy, když pro libovolné $x \in R^n$, a pro všechna $\theta \in \Omega$ platí

$$\boldsymbol{L}(\widehat{\theta_{ML}}|x) \geq \boldsymbol{L}(\theta|x).$$

Je výhodnější místo s věrohodností funkcí pracovat s jejím přirozeným logaritmem, z tohoto důvodu zavádíme logaritmickou věrohodnostní funkci:

$$l(\theta) = \ln \boldsymbol{L}(\theta|x).$$

Z důvodu, že je logaritmická věrohodnostní funkce neměnní polohu maximálně věrohodného odhadu. Abychom nalezli maximum funkce $\boldsymbol{L}(\theta|x)$ využijeme metod známých z matematické analýzy sloužících k hledání extrémů funkcí více proměnných. Předpokládejme, že $\boldsymbol{L}(\theta|x)$ má parciální derivace alespoň druhého řádu na Ω. Parciální derivace věrohodnostní funkce je ve tvaru:

$$\frac{\partial \boldsymbol{L}(\theta|x)}{\partial \theta_j} = 0,$$

pro $j = 1, \ldots, m$, která má řešení $\theta = \hat{\theta}$. Aby funkce $\boldsymbol{L}(\theta|x)$ nabývala v bodě $\hat{\theta}$ svého maxima musí platit:

$$H(\widehat{\theta}) = \left(\frac{\partial^2 \boldsymbol{L}(\theta|x)}{\partial\theta_i \partial\theta_j} \right)_{i,j=1}^{m} | < 0$$

což znamená, že Hessova matice $H(\widehat{\theta})$ je negativně definitní.

2.1.1.1 Testy poměru věrohodností

Předpokládejme logistický regresní model **M** s odhady koeficientů **b** a podmodel $\widetilde{\boldsymbol{M}}$ s odhady koeficientů $\widetilde{\boldsymbol{b}}$. Podmodelem máme namysli původní model po vyloučení některých regresorů. Cílem je zjistit zda se model **M** a $\widetilde{\boldsymbol{M}}$ významně liší. K tomuto účelu využijeme testu poměru věrohodností, který provádíme pomocí tzv. devianci. Vezmeme původní model, který má počet parametrů stejný jako je hodnot vektorů x_i. Tento model označujeme jako saturovaný a hodnotu jeho věrohodnostní funkce označujeme l_{max}. Každý další model označujeme jako podmodel původního modelu. Přiléhavost podmodelu k původnímu modelu posuzujeme pomocí deviance:

$$D(\boldsymbol{b}) = 2\big(l_{max} - l(\boldsymbol{b})\big).$$

Čím vyšší hodnota deviance, tím je přiléhavost modelu menší. Deviance je analogií k reziduálnímu součtu čtverců u lineárního regresního modelu. Devianci modelu pro podmodel saturovaného modelu bude:

$$D(\boldsymbol{b}) = -2l(\boldsymbol{b}) = -2\sum_{i=1}^{n}\big[(y_i - 1)x_i'\beta - \ln\,(1 + e^{-x_i'\beta})\big].$$

K porovnání původního modelu **M** a jeho podmodelu $\widetilde{\boldsymbol{M}}$ s využitím testu poměru věrohodností, a to pomocí deviancí. Testovaná statistika je ve tvaru:

$$2\left(l(\boldsymbol{b}) - l(\widetilde{\boldsymbol{b}})\right) = \left(2\left(l_{max} - l(\widetilde{\boldsymbol{b}})\right)\right) - \left(2\left(l_{max} - l(\boldsymbol{b})\right)\right) = \boldsymbol{D}(\widetilde{\boldsymbol{b}}) - \boldsymbol{D}(\boldsymbol{b}).$$

Testová statistika má za platnosti testovaného podmodelu asymptoty rozdělení $\chi^2(q)$, kde q je rozdíl počtu nezávislých parametrů v porovnávaných modelech. Nulovou hypotézu

$$H_0\text{: } \widetilde{\boldsymbol{M}} \textit{ je podmodel modelu } \boldsymbol{M} \textit{ zamítáme na hladině významnosti } \alpha$$

v případě kdy $\boldsymbol{D}(\widetilde{\boldsymbol{b}}) - \boldsymbol{D}(\boldsymbol{b}) \geq \boldsymbol{\chi}^2_{1-\alpha}(q)$. Využíváme 2 testů poměru věrohodnosti:

1. testování významnosti celého modelu s tzv. nulovým model

$$P(Y = 1 \,|\, X = x) = \frac{1}{1+e^{-b_0}},$$

za nulovou hypotézu bereme

$$H_0\text{: } \beta_1 = \cdots = \beta_m = 0.$$

2. ověření zda podsoubor regresorů $\beta_{i1}, \ldots, \beta_{ik}$ významně přispívá k vysvětlení variability závislé binární proměnné, testujeme nulovou hypotézu

$$H_0\text{: } \beta_{i1} = \cdots = \beta_{ik} = 0$$

2.1.1.2 Testy dobré shody, informační kritéria

Máme-li vytvořený logistický regresní model, je třeba posoudit, jak kvalitně prokládá data. K tomuto účelu se využívá řada koeficientů či testů dobré shody. Využíváme pouze skutečné realizace hodnot $y_1, \dots, y_n$ a jim příslušné modelem přiřazená skóre $s_1, \dots, s_n$. Testy dobré shody porovnávají naměřené hodnoty y_i a jim modelem přiřazené skóry s_i pomocí Pearsonova testu dobré shody. Pearsonův test dobré shody má asymptoticky χ^2 rozdělení a je ve tvaru:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - s_i)^2}{s_i(1 - s_i)}.$$

Tento test je určen Pearsonovými rezidui

$$\frac{y_i - s_i}{\sqrt{s_i(1 - s_i)}}$$

a testová statistika χ^2 má asymptoticky rozdělení $\chi^2 = (n - m - 1)$. V případě, že jsou rozdíly velké, můžeme vyvodit, že není vhodné používání těchto metod.

Informační kritéria se využívají v případě, kdy v modelu roste počet regresorů. U logistického regresního modelu roste také hodnota logaritmické věrohodnostní funkce, což může být výhodné, protože takto důvěryhodnost modelu roste. Z jiného pohledu s růstem regresorů se zvyšuje potřeba kontrolovat více proměnných. Z tohoto důvodu byla zavedena informační kritéria, která hodnotu logaritmu věrohodnostní funkce pro daný model tzv. penalizují s ohledem na počet regresorů použitých pro výstavbu modelu. Za vhodný (vyvážený) model se považuje ten, který dosáhne nejnižší hodnoty informačního kritéria. Označíme-li *l* hodnotu logaritmické věrohodnostní funkce nějakého modelu **M**, který má *m* regresorů a je vytvořen ze souboru o rozsahu *n*. Typy informačních kritérií:

Akaikeovo informační kritérium: $AIC = -2l + 2m$

Baysovo informační kritérium: $BIC = -2l + m \ln n$

Hannanovo-Quinnovo informační kritérium: $HQ = -2l + m \ln(\ln n)$

2.1.2 ROC křivka

Pojem ROC křivka je křivka operační prahové charakteristiky. V logistické regresi se využívá k měření kvality vytvořeného modelu. S ROC křivkou jsou spjaty pojmy senzitivita a specificita. Mějme celkem N objektů, které lze rozdělit do dvou skupin A a B (časová řada se ustálí a neustálí). Označíme N_A počet objektů patřící do skupiny A a N_B počet objektů patřících do skupiny B. Dále mějme klasifikační proces, který umožňuje zařadit daný objekt do skupiny A nebo B dle určitých charakteristik (v našem případě rozptyl > 0,03 časová řada průměrů se ustálí a rozptyl < 0,03 časová řada průměrů se neustálí). Zajímá nás jak přesný je tento rozhodovací algoritmus, což znamená pokud je objekt zařazen do skupiny jaká je pravděpodobnost toho, že do této skupiny opravdu patří. Z toho důvodu se definují:

- **Senzitivita** (citlivost): pravděpodobnost, že objekt, který byl zařazen do skupiny A, do skupiny A skutečně patří.

- **Specificita**: pravděpodobnost, že objekt, který byl zařazen do skupiny B (tj. nebyl zařazen do skupiny A), do skupiny B opravdu patří (nepatří do skupiny A).

Za ideální považujeme takový algoritmus kde senzitivita a specificita jsou rovny 1. Klasifikační procedura rozdělí celkový soubor N objektů do skupin, tak že do skupiny A zařadí N_A' objektů a do skupiny B jich zařadí N_B'. Z celkového počtu N_A' jich ve skutečnosti n_A patří do A a $\bar{n}_A$ je chybně zařazených do A. Senzitivitu a specificitu pak následně odhadujeme jako podíly:

$$sensitivita = \frac{n_A}{N_A}$$

$$specificita = \frac{n_B}{N_B}$$

tj. odhad senzitivity je poměr mezi počtem správně klasifikovaných objektů skupiny A vůči počtu všech objektů z A. Odhad specificity je analogický sensitivitě.

Logistickou regresí lze takto provést zpětnou klasifikaci objektů pro ohodnocení její kvality. K tomuto účelu je podstatná volba prahového bodu P_C. Pro různé hodnoty prahového bodu dosahujeme různých hodnot senzitivity a specificity. Graficky se jejich vztah zobrazuje pomocí ROC křivky. Prahový bod volíme tak, aby byla splněna některá z následujících podmínek:

- Dosažení požadované sensitivity testu
- Dosažení požadované specificity testu
- Maximalizace součtu senzitivity a specificity
- Tak aby euklidovská vzdálenost mezi levým horním bodem $[0; 1]$a ROC křivkou byla co nejmenší.

2.2 Předpověď pravděpodobnosti ustálení časových řad průměrů za využití logistické regrese

Logistická regrese je v programu R implementována funkcí *glm,* do které se jako argument zadává model. Funkce glm odhadne parametry metodou maximální věrohodnosti. Logistickou regresi spustím na datech charakteristik ($Char_1$-$Char_6$) vypočítaných z počátečního úseku časových řad průměrů. Model stanovuji následovně:

$$ln\left(\frac{P(velký\ rozptyl)}{P(malý\ rozptyl)}\right) = \beta_0 + \beta_1 Char_1 + \beta_2 Char_2 + \beta_3 Char_3 + \beta_4 Char_4 + \beta_5 Char_5 + \beta_6 Char_6$$

kde $Char_1$,…, $Char_6$ jsou charakteristiky napočítané z počátečního úseku časových řad:

- $Char_1$… minimum z počátečního úseku časových řad průměrů
- $Char_2$… maximum z počátečního úseku časových řad průměrů
- $Char_3$… průměr z počátečního úseku časových řad průměrů
- $Char_4$… počet vrcholů z počátečního úseku časových řad průměrů
- $Char_5$…průměrná vzdálenost mezi vrcholy z počátečního úseku časových řad průměrů
- $Char_6$… sklon počátečního úseku časových řad průměrů

Výsledky odhadu parametrů jsou na následující **Tabulce 2. 1.**

Tabulka 2. 1. Výsledky odhadu parametrů logistické regrese

	Estimate	Std. Error	t value	Pr (>|t|)	
β_0	57.152748	21.216688	2.694	0.00895	**
β_1	0.020500	0.007788	2.632	0.01056	*
β_2	-0.241744	0.293288	-0.824	0.41276	
β_3	-0.407840	0.077601	-5.256	1.7e-06	***
β_4	0.013176	0.222784	0.059	0.95302	
β_5	-0.011754	0.007292	-1.612	0.11176	
β_6	-0.006781	0.005763	-1.177	0.24356	

Značení významnosti parametrů: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Z výsledků vyplývá, že významné nejsou všechny parametry, z toho důvodu budu původní model redukovat. Model redukuji pomocí Akaikeovo informačního kritéria. V programu R na redukci modelu využiji funkci *step*, která model zredukuje využitím Akaikeova informačního kritéria. Funkce step zredukovala původní model se 6 proměnnými na model s 2 proměnnými.

Redukovaný model je ve tvaru:

$$ln\left(\frac{P(velký\ rozptyl)}{P(malý\ rozptyl)}\right) = \beta_0 + \beta_1 Char_1 + \beta_3 Char_3$$

Výsledky odhadů parametrů pro redukovaný model, metodou maximální věrohodnosti, jsou na následující **Tabulce 2. 2.**

Tabulka 2. 2.: Výsledky odhadu parametrů redukovaného modelu

	Estimate	**Std. Error**	**t value**	**Pr (>|t|)**	
β_0	33.65865	6.44878	5.219	1.75e-06	***
β_1	0.01451	0.00219	6.626	5.97e-09	***
β_3	-0.37729	0.07240	-5.212	1.81e-06	***

Značení významnosti parametrů: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pomocí testu poměru věrohodností porovnám redukovaný model s původním modelem. Výsledky jsou v **Tabulce 2. 3.:**

$$Původní\ model: Rozptyl \sim Char_1 + Char_2 + Char_3 + Char_4 + Char_5 + Char_6$$

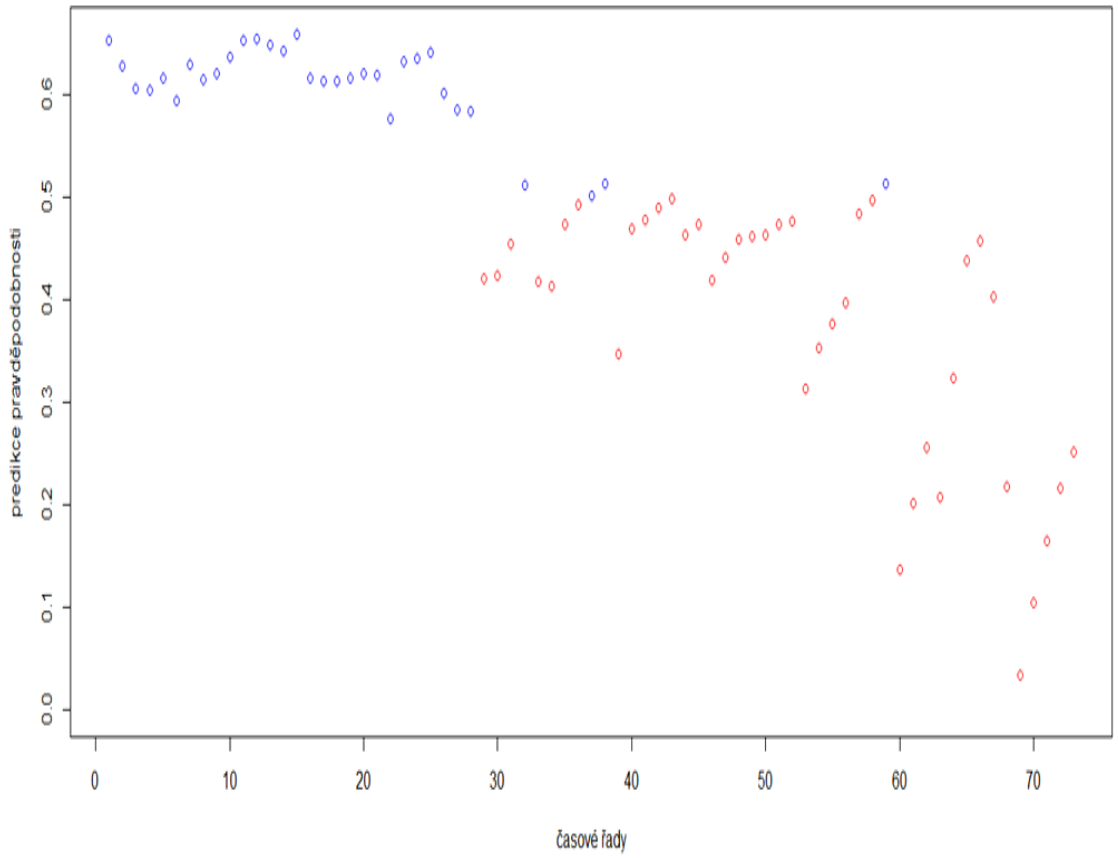
$$Redukovaný\ model: Rozptyl \sim Char_1 + Char_3$$

Tabulka 2. 3.: Výsledky porovnání původního modelu s modelem redukovaným pomocí testu poměru věrohodností

	#Df	LogLik	Df	Chisq	Pr (>Chisq)
Model původní	8	-19.578			
Model Redukovaný	4	-21.983	-4	4.809	0.3075

Značení významnosti modelu: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Z tabulky vyplývá, že redukovaný model se moc neliší od původního modelu. K predikci pravděpodobnosti ustálení časových řad průměrů využiji redukovaný model. Dosazením hodnot do redukovaného modelu získám předpovědi toho, že se časová řada průměrů ustálí. Výsledky predikce pravděpodobnosti ustálení časových řad průměrů jsou na následujícím **Obrázku 2. 2.**

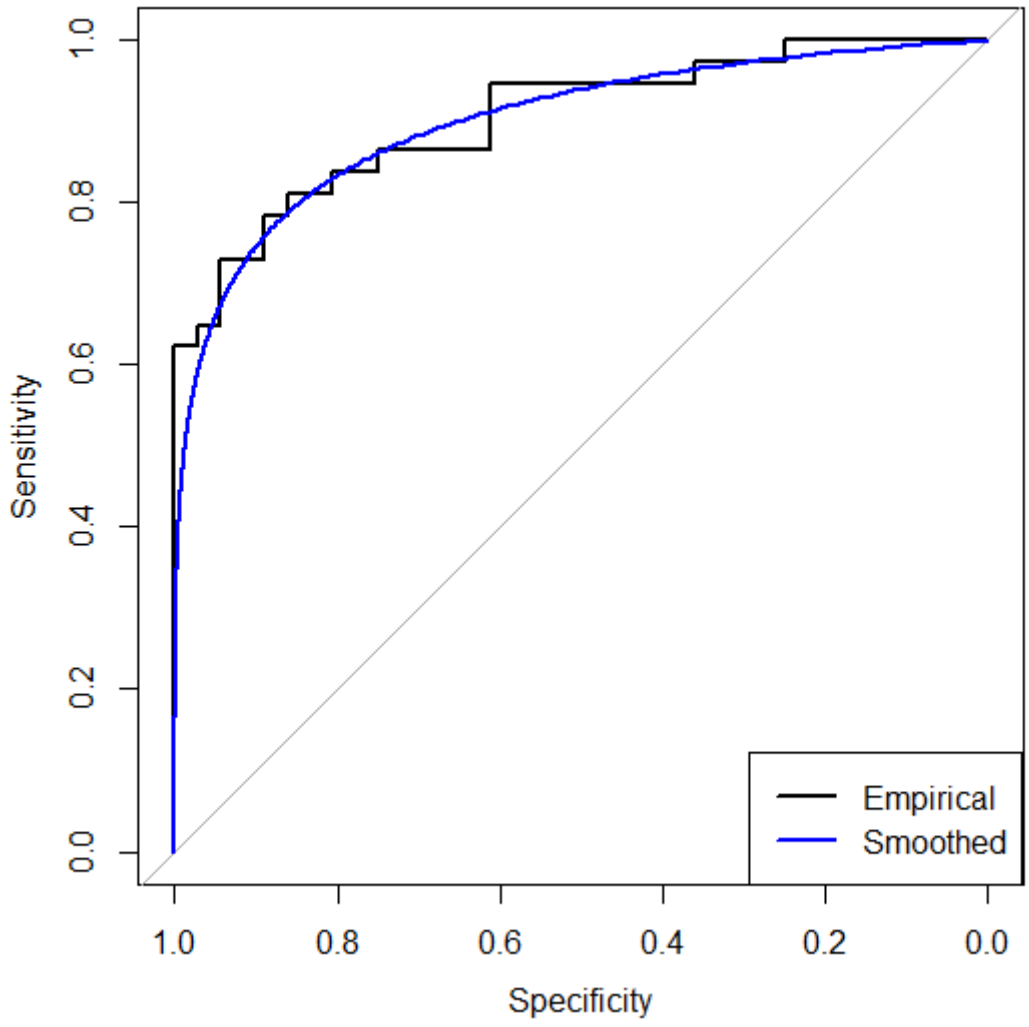


Obrázek 2. 2.: Výsledky predikce pravděpodobnosti ustálení časových řad průměrů

Na **Obrázku 2. 2.** jsou modře vyjádřeny časové řady s vysokou pravděpodobností ustálení tedy nad 50%. Červeně jsou na **obrázku 2. 2.** vyjádřeny časové řady, které mají pravděpodobnost ustálení pod 50%. Z výsledků predikce pravděpodobnosti ustálení časových řad průměrů vyplývá, že většina časových řad průměrů se neustálí.

2.3 Posouzení kvality logistické regrese

K posouzení kvality logistické regrese využiji ROC křivku. ROC křivka je do programu R naimplementována pomocí funkce *roc*. Do funkce *roc* se jako argumenty zadávají původní proměnné logistické regrese a vektor předpovědí. Vektor předpovědí je v mém případě rozptyl, který je zadán binárně. Následně funkce vypočítá sensitivitu a specificitu společně s prahovým bodem a vykreslí ROC křivku. Pomocí argumentu *smooth* vykreslí funkce vedle původní ROC křivky, novou vyrovnanou ROC křivku. Na následujícím **Obrázku 2. 3.** je vykreslena původní ROC křivka a i vyrovnaná ROC křivka k logistické regresi.



Obrázek 2. 3.: ROC křivka výsledků logistické regrese

Prahový bod vypočítala funkce *roc* jako 50%. Plocha pod křivkou z **Obrázku 2. 3.** je rovna 0,898 z čehož vyplývá, že zvolený redukovaný logistický model je vhodný pro předpověď zda se časová řada průměrů v čase ustálí či zůstane kolísavá.

Závěr

Cílem mé práce bylo využít počáteční úseky 73 dlouhých CFD simulací jedné sady tvarové optimalizace čerpací pumpy, a prozkoumat zda se dají použít jako prediktory koncových úseků těchto 73 dlouhých CFD simulací. Těchto 73 CFD simulací považuji za časové řady. Za počáteční úsek považuji prvních 80 datových bodů vyrovnaných časových řad (b). Chci predikovat průměr posledních deseti členů vyrovnaných časových řad (b). Této predikci jsem se věnovala v první kapitole této práce. Průměr posledních deseti členů vyrovnaných časových řad je zvolen, protože je to robustnější ukazatel účinnosti než pouhá konečná hodnota. Vzhledem k tomu, že mám k dispozici pouze 73 vzorků časových řad, nebylo možné jako prediktory finálního průměru použít přímo všech 80 počátečních bodů vyrovnaných časových řad (b). Proto jsem se pokusila chování časových řad na jejich počátečním úseku zachytit 6 charakteristikami a to:

- Rozdílem prvních dvou datových bodů
- Minimální hodnotou počátečního úseku vyrovnaných časových řad
- Maximální hodnotnou z počátečního úseku vyrovnaných časových řad
- Průměrnou hodnotou z počátečního úseku vyrovnaných časových řad
- Počtem vrcholů na počátečním úseku vyrovnaných časových řad
- Průměrnou vzdáleností mezi vrcholy na počátečním úseku vyrovnaných časových řad

K získání predikce průměru posledních deseti hodnot vyrovnaných časových řad jsem použila dvě metody: shlukovací metodu K-means a metodu vážené lineární regrese.

Predikce průměru posledních deseti hodnot vyrovnaných časových řad s využitím shlukovací metody K-means probíhala takto: z každé vyhlazené časové řady jsem extrahovala mnou definovaných 6 charakteristik. Následně jsem na takto vytvořený 6-ti rozměrných vektorech našla shluky pomocí metody k-means. Predikce probíhala takto: když přišla nová vyrovnaná časová řada (b), dle jejího počátečního úseku jsem ji zařadila do správného shluku. Následně konec této časové řady jsem předpověděla typickým koncem vyrovnaných časových řad v tomto shluku. Předpověď typickým koncem

znamená, že v rámci každého shluku jsem se podívala na posledních deset hodnot vyrovnané časové řady a udělala jsem její medián. Tento medián je hodnota, kterou jsem predikovala.

Rozumný počet shluků k predikci posledních deseti hodnot vyrovnaných časových řad jsem volila dvěma variantami a to:

(1) Pomocí vykreslení závislostí vnitřního skupinového součtu čtverců na počtu shluků.

(2) Nalezením počtu shluků, kde celková chyba predikce je co nejmenší a zároveň shluků není příliš mnoho. Celková chyba predikce je definovaná takto:

$$J = \sum_{i=1}^{73} váha\ |predikce - skutečnost|.$$

U obou metod vycházel rozumný počet shluků přibližně stejně. Z toho důvodu jsem pro predikci posledních deseti hodnot vyrovnaných časových řad využila rozumný počet shluků 5. Výsledky predikce posledních deseti hodnot vyrovnaných časových řad oproti skutečnosti s využitím metody k-means jsou na **Obrázku 1. 10.**

Predikce průměru posledních deseti hodnot vyrovnaných časových řad s využitím metody vážené lineární regrese probíhala následovně: stanovila jsem pomocí charakteristik vypočítaných z počátečních úseků vyrovnaných časových řad lineární regresní rovnici. Ve, které jsem chtěla predikovat průměr posledních deseti hodnot vyrovnaných časových řad pomocí charakteristik vypočítaných z počátečních úseků těchto řad. Následně jsem odhadla neznámé parametry váženou metodou nejmenších čtverců. Podle významnosti regresorů jsem původní lineární regresi zredukovala a opět odhadla neznámé parametry váženou metodou nejmenších čtverců. Následně jsem do lineární regresní rovnice dosadila hodnoty. Výsledky predikce průměru posledních deseti hodnot vyrovnaných časových řad jsou na **Obrázku 1. 12.**

Obě metody výpočtu predikce průměru posledních deseti hodnot vyrovnaných časových řad jsem porovnala na **Obrázku 1. 13.** Lepší metoda pro predikci průměru posledních deseti hodnot vyrovnaných časových řad je metoda lineární regrese což potvrzuje i porovnání dle reziduálních součtů čtverců.

Ve druhé kapitole jsem predikovala, zda se vyrovnané časové řady ustálí na nějaké finální hodnotě nebo se neustálí a na svém konci budou kolísavé. K této predikci jsem využila logistického regresního modelu. Za klasifikátor jsem uvažovala rozptyl posledních deseti hodnot vyrovnaných časových řad. Stanovila jsem logistický regresní model a vypočítala neznámé parametry. Dle významnosti parametrů jsem původní model redukovala a znovu odhadla neznámé parametry. Nakonec jsem porovnala redukovaný model s nulovým model a původním modelem. Dle výsledků vyšel vyhovující redukovaný model, do kterého jsem dosadila hodnoty a predikovala, zda se vyrovnaná časová řada ustálí či nikoliv. Pro posouzení kvality logistické regrese jsem vykreslila ROC křivku, která je na **Obrázku 2. 3.** Je již na rozhodnutí společnosti Sigma Lutín, která z chyb má pro ně větší váhu, zda specificita či senzitivita.

Seznam použitých zdrojů:

[1] J. Anděl: *Základy matematické statistiky*. MATFYZPRESS, Praha 2007

[2] K. Zvára: *Regrese.* MATFYZPRESS, Praha 2008

[3] Alena Lukasová, Jana Šarmanová: Metody shlukové analýzy. SNTL, Praha 1985

[4] Tomáš Cipra, *Analýza časových řad s aplikacemi v ekonomii,* SNTL, Praha 1986

[5] Kelly H. Zou, W. J. Hall and David E.Shapiro (1997),*Smooth non-parametric receiver operating charakteristic (ROC) curves fot continuous diagnostic tests*