Analýza závislosti veličin sledovaných v rámci TBD

Helena Koutková
Vysoké učení technické v Brně, Fakulta stavební, Ústav matematiky a deskriptivní geometrie
e-mail: koutkova.h@fce.vutbr.cz

Abstrakt

Příspěvek se zabývá analýzou závislosti veličin, které jsou sledovány v rámci technicko-bezpečnostního dohledu přehrad (TBD). Je orientován především na některá úskalí vyskytující se při rutinním použití statistických metod, kdy nejsou ověřovány předpoklady těchto metod.

1. Úvod

Statistická analýza rozsáhlých souborů dat, které jsou získávány v rámci TBD na vodních dílech, se dosud omezuje převážně na odhad korelace (těsnosti lineární závislosti) a odhad parametrů lineární regresní funkce sledovaných veličin metodou nejmenších čtverců. Předpokladem úspěchu použití statistických metod je respektování jejich předpokladů.

Důležité je také uvědomit si, za jakým účelem se statistické hodnocení provádí. Tak například z historického chování hráze odvozená přiléhavá regresní závislost mezi nezávislými (např. poloha hladiny v nádrži, srážkový úhrn) a závislými (piezometrická výška ve vrtech, průsak) veličinami by měla umožnit kontrolu, zda se závislé veličiny pohybují v „přijatelných“ mezích. To vede k doplnění regresního modelu o pásy spolehlivosti.

Základním vstupním předpokladem pro úspěšné použití regresních modelů je homogenita dat („stejné“ vstupní podmínky v období jejich získávání). K odlišení dat získaných za jiných podmínek může sloužit mimo jiné shluková analýza.

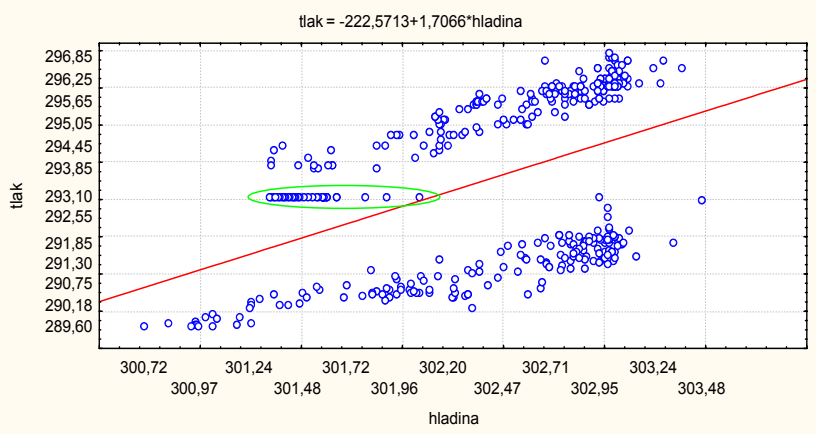
V dalším textu jsou v případě analýzy závislosti piezometrické výšky (dále označované jako „tlak“) ve vrtech (náhodná veličina Y) na hladině vody v nádrži (náhodná veličina X) demonstrovány některé postupy, které by bylo vhodné aplikovat v rámci vyhodnocení veličin naměřených v rámci TBD. K dispozici bylo 410 měření hodnot hladiny a tlaku v období od 1/3/00 do 2/26/07, měření byla prováděna průměrně 1x za 7 dní.

2. Grafická prezentace

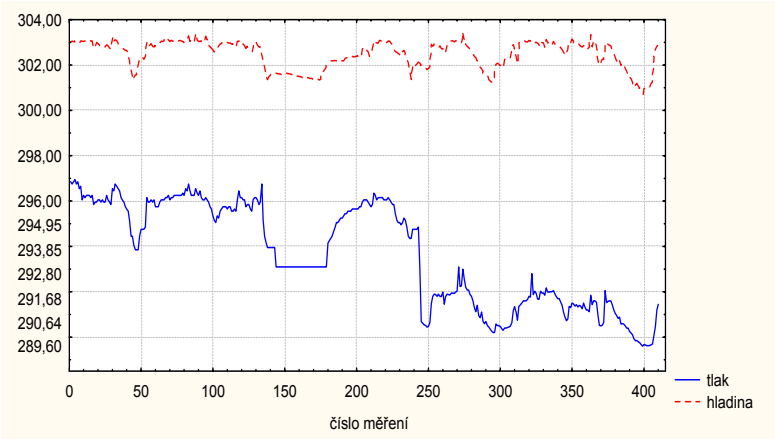
Základní představu o vlastnostech datového souboru poskytne grafická prezentace dat, která umožňuje odhalit případné zvláštnosti v datech.

Označme jako x_i, resp. y_i zjištěné hodnoty hladiny a tlaku v i-tém pozorování ($i = 1, \ldots, N$). Zapíšeme-li vektory pozorování (x_i,y_i) do řádků matice, dostaneme *datovou matici* typu $N/2$.

Dvourozměrný datový soubor lze znázornit pomocí *bodového diagramu*, kde hodnoty proměnných chápeme jako souřadnice bodů v rovině. Z bodového diagramu (obr.1) je patrné, že data netvoří homogenní soubor, což je pro statistické analýzy nutný předpoklad. Jsou zde rozpoznatelné dva shluky a navíc i „podezřelá“ měření označená v elipse. Jedná se o měření 145-179 ve dnech 11/29/02 až 1/3/03. Z obr. 2 lze usuzovat, že tyto body patrně odpovídají poruše měření v daném období, kdy byla měřena konstantní hodnota tlaku 293,10. Při další analýze tyto údaje vypustíme. Zbývající data podrobíme *shlukové analýze* za účelem identifikace příslušných shluků.



Obr. 1 Bodový diagram



Obr. 2 Průběh hodnot hladiny a tlaku v závislosti na pořadí měření

3. Shluková analýza

Cílem shlukové analýzy je roztřídění *N* pozorování do několika pokud možno homogenních shluků. Požadujeme, aby pozorování uvnitř jednotlivých shluků byla podobná co nejvíce a pozorování z různých shluků co nejméně.

Podobnost nebo *nepodobnost* pozorování měříme pomocí různých měr vzdáleností. Nejčastěji se používá *euklidovská vzdálenost*. Z datové matice víme, že:

i-té pozorování je charakterizováno vektorem pozorování (x_i, y_i),

j-té pozorování je charakterizováno vektorem pozorování (x_j, y_j).

Jejich euklidovská vzdálenost d_{ij} je:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} .$$

Čím je vzdálenost d_{ij} menší, tím jsou pozorování s indexy *i* a *j* podobnější. Vzdálenosti vypočítané pro všechna pozorování se zapisují do tzv. *matice vzdáleností*. Pro vytváření shluků existuje celá řada algoritmů. Pokud známe počet *k* shluků, můžeme např. použít tzv. *metodu k-průměrů*. Pokud počet *k* shluků neznáme, což je v aplikacích častější, používá se nejčastěji tzv. *aglomerativní hierarchický algoritmus*, kde se jednotlivé typy liší tím, jak měříme vzdálenosti mezi vytvářenými shluky.

Hierarchické shlukování spočívá v tom, že postupně slučujeme pozorování, a to nejprve nejbližší a v dalších krocích pak stále vzdálenější. Dostáváme tak postupně rozklady $S^{(1)}$,…,

$S^{(N)}$ od rozkladu na jednotlivá pozorování $S^{(1)}$ až do rozkladu $S^{(N)}$, který obsahuje jediný shluk, a to všechna pozorování. Přitom každý předchozí rozklad $S^{(m)}$ je zjemněním následujícího $S^{(m+1)}$. Při zjemnění shluky v rozkladu $S^{(m)}$ vznikají rozdělením některých shluků v rozkladu $S^{(m+1)}$. Postup lze shrnout do 3 kroků.

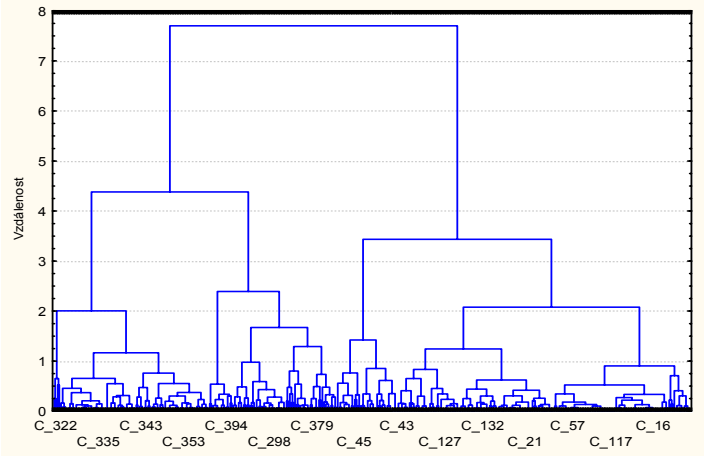
1. Každé pozorování považujeme za samostatný shluk.
2. V matici vzdáleností najdeme 2 shluky, jejichž vzdálenost je minimální.
3. Tyto 2 shluky spojíme do nového, většího shluku. Přepočítáme matici vzdáleností. Její řád se zmenší o 1. Pokud je počet shluků $k > 1$, vrátíme se na krok 2, pokud je $k = 1$, ukončíme výpočet.

Pro výpočet vzdálenosti mezi vícebodovými shluky byla použita *metoda průměrné vazby*, kdy je vzdálenost mezi 2 shluky průměrem vzdáleností mezi jejich pozorováními. Výsledky hierarchického shlukování znázorňujeme graficky pomocí tzv. *dendrogramu*. Je to graficky znázorněná posloupnost dvojic $\{(v_1,S^{(1)}),\ldots,(v_N,S^{(N)})\}$, kde $v_1, \ldots, v_N$ je neklesající posloupnost úrovní spojování a $S^{(i)}$ je roztřídění pozorování odpovídající úrovni v_i , $i = 1, \ldots, N$. Z dendrogramu pro danou úroveň vzdáleností – tzv. *řez dendrogramu*, určíme počet k shluků i složení shluků.

Metoda k-průměrů je nehierarchická metoda, vychází z následujícího algoritmu

1. Náhodně stanovíme rozklad souboru N pozorování do k shluků.
2. Určíme výběrové průměry (centroidy).
3. Pro všechna pozorování spočítáme vzdálenost od všech výběrových centroidů. Pozorování zařadíme do toho shluku, k jehož výběrovému centroidu má nejblíže. Pokud nedošlo k žádnému přesunu, považujeme aktuální shluky za definitivní, jinak se vrátíme na krok 2.

Stejně postupujeme i v případě, kdy pracujeme s vícerozměrnými daty, než jsou dvourozměrná. Na obr. 3 je zachycen výsledek hierarchického shlukování pro data z obr. 1. Pro úroveň vzdáleností 5 jsme dostali 2 výrazné shluky. Stejný výsledek byl získán pomocí metody *k-průměrů*, $k = 2$. Do jednotlivých shluků byla zařazena stejná pozorování jako v případě hierarchického shlukování. Rozklad do dvou shluků je tvořen tak, jak je již předem patrné z obr. 1, kde shluk č. 1 je pod šikmou čárou a shluk č. 2 nad ní.

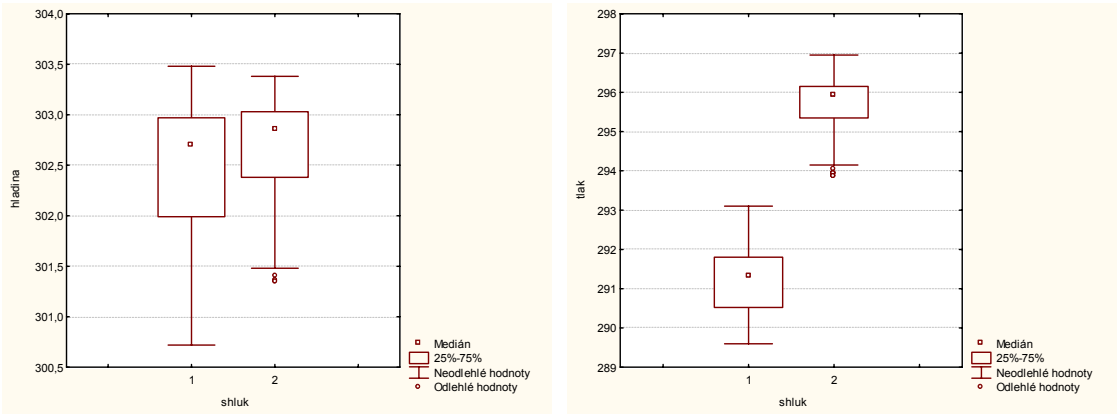


Obr. 3 Dendrogram

Ve 2. shluku je 207 pozorování, v 1. shluku 166. Popisné statistiky shluků jsou v tab.1, indexy značí shluk 1 resp. 2. Na obr. 4 jsou znázorněny krabicové grafy. Je vidět, že se shluky příliš neliší úrovní hladiny, liší se ale její variabilitou a variabilitou a úrovní tlaku. Větší variabilita tlaku v 1. shluku je způsobena větší variabilitou hladiny v 1. shluku. Ze situace na přehradě je známo, že shluk 1 přísluší stavu po technickém zásahu a shluk 2 stavu před zásahem, který nastal mezi 17.11. a 1.12.2003. Při analýze musí být každý shluk zpracováván zvlášť.

Tab. 1 Popisné statistiky shluků

Proměnná	Počet	Průměr	Medián	Min	Max	Dolní kvartil	Horní kvartil	Rozpětí	Kvart rozpětí	Sm odch
hladina_1	166	302,418	302,695	300,720	303,480	301,990	302,970	2,760	0,980	0,658
tlak_1	166	291,161	291,335	289,600	293,100	290,520	291,800	3,500	1,280	0,764
hladina_2	207	302,684	302,850	301,350	303,380	302,380	303,030	2,030	0,650	0,449
tlak_2	207	295,695	295,950	293,850	296,950	295,350	296,150	3,100	0,800	0,715

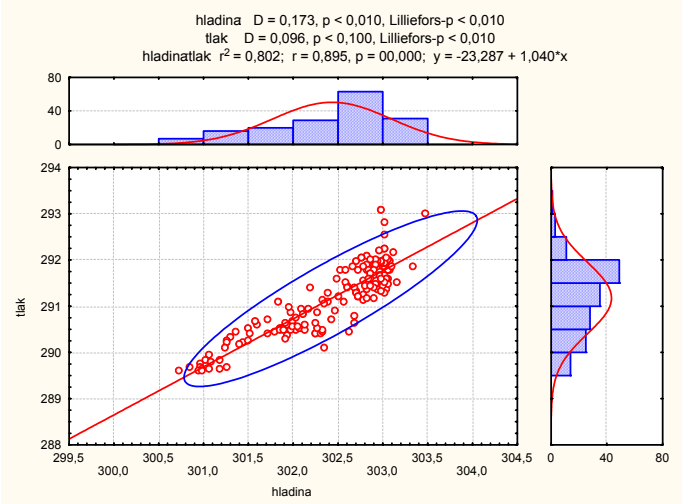


Obr. 4 Krabicové grafy

4. Regresní model

Dále se zabýváme závislostí tlaku na hladině v případě shluku 1. Bodový diagram (obr.5) indikuje silnou *lineární závislost*. Těsnost lineární závislosti dvou kvantitativních náhodných veličin měříme *výběrovým korelačním koeficientem R*, jehož druhá mocnina (*výběrový koeficient determinace* R^2) násobená 100 udává, kolik % variability jedné z proměnných lze vysvětlit lineární závislostí na druhé. V případě 1. shluku je $r = 0{,}895$. Jde tedy o silnou lineární závislost, neboť lze lineární závislostí tlaku na hladině vysvětlit 80,2% variability tlaku.

Zamítáme sice shodu marginálních rozdělení s normálním na hladině významnosti 0,01, protože *p*-hodnota Lilieforsova varianty Kolmogorovova testu shody je menší než 0,01. Tato rozdělení jsou mírně zešikmená. Dvourozměrné rozdělení je však eliptické, což vystihuje graf elipsy na obr. 5. Vzhledem k tomu, že se nejedná o výrazné odchýlení od normálního rozdělení, lze testovat hypotézy o hodnotě korelačního koeficientu a konstruovat intervaly spolehlivosti – viz [3].



Obr. 5 Analýza rozdělení sledovaných veličin ve shluku 1

Průběh závislosti hodnot tlaku (*závisle proměnné Y*) na hodnotách hladiny (*nezávisle proměnné X*) vystihuje regresní funkce veličiny Y vzhledem k veličině X, tj. *podmíněná střední hodnota* $E(Y|x)$ náhodné veličiny Y za podmínky, že náhodná veličina X nabyla hodnoty x. Na základě výše uvedených výsledů předpokládáme, že

$y = E(Y|x) = \beta_0 + \beta_1 x,$

kde β_0 a β_1 jsou neznámé konstanty, tzv. *regresní parametry*. Označme $\boldsymbol{\beta} = (\beta_0,\beta_1)^T$ vektorový parametr a $\boldsymbol{x} = (1,x)^T$, potom lze psát

$y = E(Y|x) = \boldsymbol{x}^T\boldsymbol{\beta}$.

Jedná se o speciální případ lineární regresní funkce veličiny Y vzhledem k veličině X (tj. funkce, která je lineární vzhledem k parametrům), protože je lineární i vzhledem k hodnotám nezávisle proměnné. Vektorový parametr $\boldsymbol{\beta}$ se běžně odhaduje metodou nejmenších čtverců (MNČ), aniž by se ověřovaly předpoklady, které zaručují „dobré“ vlastnosti MNČ odhadů. Ty jsou zaručeny, pokud se vektor pozorování závisle proměnné řídí tzv. *klasickým lineárním regresním modelem* (KLRM). Označme tedy

$\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$ sloupcový n-rozměrný náhodný vektor, jehož složky Y_i jsou neznámé hodnoty závisle proměnné Y za podmínky, že nezávisle proměnná X nabyla hodnoty x_i, tj.

$Y_i = Y|x_i$ pro $i = 1, \ldots, n$.

O náhodném vektoru $\boldsymbol{Y}$ říkáme, že se řídí KLRM, jestliže pro jeho střední hodnotu $E(\boldsymbol{Y})$ a kovarianční matici $cov(\boldsymbol{Y})$ platí

$E(\boldsymbol{Y}) = \boldsymbol{A\beta},\ cov(\boldsymbol{Y}) = \sigma^2\boldsymbol{I},$

kde $\boldsymbol{A}$ je tzv. *regresní matice*, což je v případě naší regresní funkce matice, jejíž i-tý řádek je $\boldsymbol{x}_i^T = (1,x_i)$, $i = 1, .., n$. $\boldsymbol{\beta}$ je k rozměrný sloupcový vektorový parametr, u nás $\boldsymbol{\beta} = (\beta_0,\beta_1)^T$. $\boldsymbol{I}$ je jednotková matice typu n/n a σ^2 je neznámý parametr. KLRM můžeme psát ve tvaru

$\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{\varepsilon}$,

kde $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$ je vektor tzv. *náhodných chyb*, který má nulový vektor středních hodnot a kovarianční matici $\sigma^2\boldsymbol{I}$ a který je výslednicí neuvažovaných náhodných vlivů.

V běžně používané MNČ se předpokládá, že regresní parametry mohou nabývat libovolných hodnot, tj. nejsou na ně kladeny žádné omezující požadavky. Jednoznačnost MNČ pak zaručuje regulárnost matice $\boldsymbol{A}$.

Pokud je regresní matice $\boldsymbol{A}$ stochastická, což v našem případě je, protože se nejedná o plánovaný a řízený laboratorní experiment, ale o pozorování dvourozměrného rozdělení náhodného vektoru (X,Y), můžeme využít všech postupů regresních modelů s tím, že požadujeme, aby vysvětlující proměnná a náhodná chyba byly nezávislé náhodné veličiny. Matici $\boldsymbol{A}$ můžeme také považovat za deterministickou, pokud jsou hodnoty nezávisle proměnných měřeny s větší přesností než hodnoty závisle proměnných.

Pro konstrukci *intervalů spolehlivosti* a *testy hypotéz* se předpokládá vícerozměrné normální rozdělení vektoru náhodných chyb $\boldsymbol{\varepsilon}$ (tedy i náhodného vektoru $\boldsymbol{Y}$) nebo dostatečně velký rozsah souboru a nevelké odchylky od normálního rozdělení.

Za výše uvedených předpokladů je MNČ odhad $\hat{\boldsymbol{\beta}}$ vektorového parametru $\boldsymbol{\beta}$, tj. statistika, ve které funkce

$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{A\beta})^T(\boldsymbol{Y} - \boldsymbol{A\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$

nabude absolutní minimum, *nejlepším nestranným lineárním odhadem* parametru $\boldsymbol{\beta}$. Tedy odhady složek vektorového parametru $\boldsymbol{\beta}$ jsou lineárními kombinacemi složek vektoru $\boldsymbol{Y}$, jejich realizace kolísají okolo jejich skutečné hodnoty a mezi všemi nestrannými lineárními odhady mají nejmenší rozptyl.

Hledání MNČ odhadu $\hat{\boldsymbol{\beta}}$ pak vede na řešení *soustavy normálních rovnic*

$\boldsymbol{A}^T\boldsymbol{A\beta} = \boldsymbol{A}^T\boldsymbol{Y}$.

Jedná se o soustavu k lineárních rovnic pro neznámý vektor parametrů $\boldsymbol{\beta}$, která má v případě regulární regresní matice $\boldsymbol{A}$ právě jedno řešení

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{Y} .$$

Toto řešení je přímo hledaným odhadem vektoru $\boldsymbol{\beta}$ získaným metodou nejmenších čtverců.

Bodové odhady *parametrických funkcí*, tj. funkcí parametru $\boldsymbol{\beta}$, pak dostaneme tak, že za parametr $\boldsymbol{\beta}$ dosadíme jeho odhad $\hat{\boldsymbol{\beta}}$. Tak např. pro odhad $\hat{\boldsymbol{Y}}$ vektoru $\boldsymbol{Y}$ používáme odhad $\hat{E}(\boldsymbol{Y})$ jeho střední hodnoty $E(\boldsymbol{Y})$, tj.

$$\hat{\boldsymbol{Y}} = \hat{E}(\boldsymbol{Y}) = \boldsymbol{A}\hat{\boldsymbol{\beta}} .$$

Odhad $\hat{\boldsymbol{\varepsilon}}$ vektoru chyb $\boldsymbol{\varepsilon}$ je

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} .$$

Složky tohoto vektoru se nazývají (*klasická*) *rezidua*. *Bodovým odhadem regresní funkce* $y=E(Y|x)= \beta_0 + \beta_1 x = \boldsymbol{x}^T\boldsymbol{\beta}$ je statistika

$$\hat{E}(Y \mid x) = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}.$$

Tuto statistiku používáme i pro *předpověď* $\hat{Y} \mid x$ *hodnoty* $y|x$ veličiny $Y|x= \mathrm{E}(Y|x)+ \varepsilon |x$.

Všechny výše uvedené odhady jsou za předpokladu, že se náhodný vektor $\boldsymbol{Y}$ řídí KLRM, opět nejlepší nestranné lineární odhady.

Nestranným odhadem rozptylu σ^2 je statistika

$$S^2 = S_e /(n-k) ,$$

kde k je počet regresních parametrů v regresní funkci a

$$S_e = S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}$$

je tzv. *reziduální součet čtverců*. Statistiku S nazýváme *směrodatná chyba modelu*.

4.1 Adekvátnost modelu

Statistickým kritériem kvality modelu je reziduální součet čtverců S_e i směrodatná chyba modelu S, které měří rozptýlenost hodnot náhodné veličiny Y okolo regresní funkce. Čím jsou obě statistiky menší, tím je model adekvátnější. Nevýhodou obou statistik je, že nejsou omezeny shora a hodí se tudíž spíše pro porovnávání kvality různých modelů. V případě, že má regresní funkce absolutní člen, lze variabilitu závisle proměnné Y vyjádřenou tzv. *celkovým součtem čtverců* S_c rozložit na část, která není vysvětlená regresním modelem (tj. *reziduální součet čtverců* S_e) a část, která je regresním modelem vysvětlena, tj. tzv. *teoretický (regresní) součet čtverců* S_t, *tj.*

$$S_c = S_e + S_t, \;\; S_c = \sum_{i=1}^{n} (Y_i - M_Y)^2 , \; S_t = \sum_{i=1}^{n} (\hat{Y}_i - M_Y)^2 , \; M_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i .$$

Lze ukázat, že

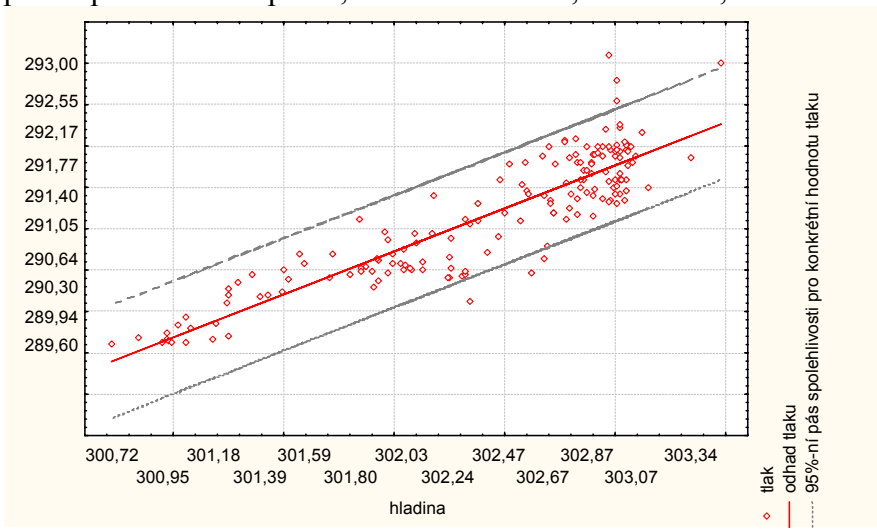
$$S_t / S_c = 1 - S_e / S_c = R^2,$$

kde R^2 je výběrový koeficient determinace. Kritériem shody modelu s daty je tedy v tomto případě i výběrový koeficient determinace R^2, který je omezen shora číslem 1 a jehož interpretaci známe. O kvalitě modelu svědčí i „délka“ *intervalů spolehlivosti* pro konkrétní hodnoty $Y|x$. V případě velkého rozsahu souboru je 95 resp. 99%-ní interval spolehlivosti pro konkrétní hodnotu $Y|x$ (tj. interval, který ji s pravděpodobností aspoň 0,95 resp. 0,99 překryje) interval, jehož krajní meze jsou přibližně rovny $\hat{Y} \mid x \pm 2S$ resp. $\hat{Y} \mid x \pm 3S$.

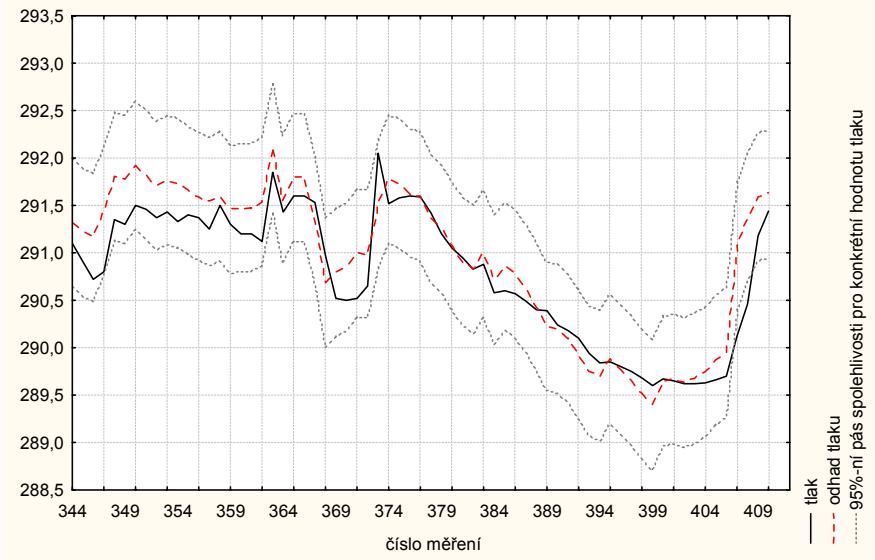
Vztahy pro výpočet přesného intervalu spolehlivosti pro konkrétní hodnotu $Y|x$, regresní funkci, tj. $E(Y|x)$ i regresní parametry lze najít např. v [5]. Meze intervalu spolehlivosti pro konkrétní hodnotu $Y|x$ při spojitě se měnícím x vytvoří tzv. *pás spolehlivosti* okolo regresní

funkce. V aproximativním případě jsou meze pásu rovnoběžné s regresní funkcí. Při menším rozsahu dat je pás nejužší v bodě, jehož souřadnice jsou průměry nezávisle a závisle proměnné a směrem k větším i menším hodnotám nezávisle proměnné se rozšiřuje.

Odhad regresní funkce ve shluku 1 je $\hat{y} = -23{,}2873+1{,}0398x$. Směrodatná chyba modelu, která se používá ke konstrukci intervalů spolehlivosti, je $s = 0{,}341$. Odhad koeficientu determinace je 0,802. Z hlediska shody pozorovaných a modelovaných hodnot (odhadů historických dat) se jedná o kvalitní model. Realizace 95%-ního intervalu spolehlivosti pro konkrétní hodnotu $Y|x$ je přibližně ($1{,}040x$ -23,605,$1{,}040x$-22,605). Regresní model včetně 95%-ního pásu spolehlivosti je na obr.6, výsek tímto modelem v závislosti na pořadí měření je na obr.7. Příslušné hodnoty pro měření 400 – 410 jsou uvedeny v tab. 2. Tak např. odhad hodnoty tlaku a odhad střední hodnoty tlaku ve dni 1/1/07 je 289,636, naměřená hodnota byla 289,62. Hodnota tlaku je s pravděpodobností aspoň 0,95 v rozmezí 288,949 až 290,320 .



Obr. 6. Regresní model v případě 1. shluku



Obr. 7. Výsek regresním modelem včetně pásů spolehlivosti

Tab. 2 Odhady tlaku a střední hodnoty tlaku získané metodou nejmenších čtverců

Čís.m.	Datum	Hladina	Tlak	Odhad tlaku	Klasická rezidua	-95%PI	+95%PI	Odhad AR
400	12/18/06	300,97	289,67	289,656	0,014	288,970	290,341	289,808
401	12/25/06	300,98	289,65	289,666	-0,016	288,981	290,351	289,723
402	1/1/07	300,95	289,62	289,635	-0,015	288,949	290,320	289,629
403	1/8/07	300,99	289,62	289,676	-0,056	288,991	290,361	289,664
404	1/15/07	301,06	289,63	289,749	-0,119	289,065	290,433	289,714
405	1/22/07	301,18	289,66	289,874	-0,214	289,191	290,557	289,794
406	1/29/07	301,25	289,70	289,947	-0,247	289,265	290,629	289,798
407	2/5/07	302,34	290,12	291,080	-0,960	290,405	291,756	290,891
408	2/12/07	302,62	290,46	291,371	-0,911	290,695	292,047	290,774
409	2/19/07	302,83	291,18	291,590	-0,410	290,913	292,266	290,849
410	2/26/07	302,87	291,44	291,631	-0,191	290,955	292,308	291,183

Legenda: -95%PI a +95%PI značí dolní a horní mez 95%-ního intervalu spolehlivosti konkrétní hodnoty tlaku, odhad AR značí odhad hodnoty tlaku pomocí autoregresního modelu řádu 1

4.2 Analýza reziduí

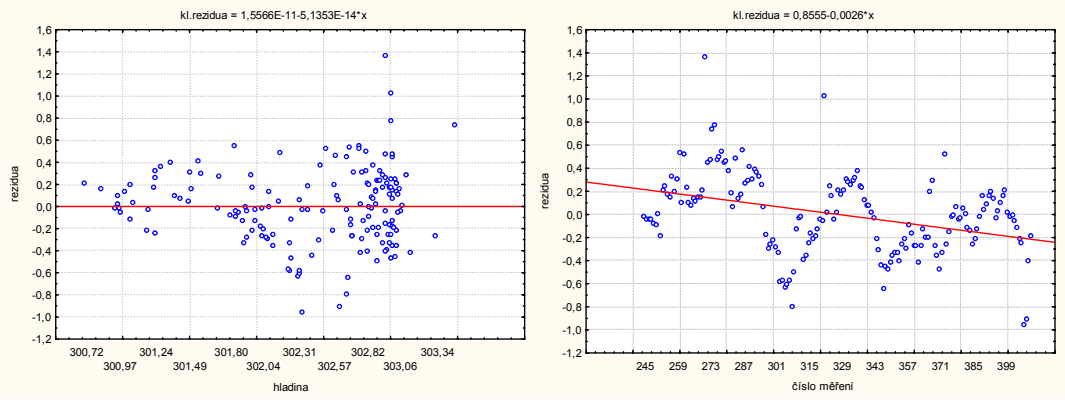
K ověřování předpokladů o náhodné chybě, kvalitě dat ale i k vylepšování modelu využíváme analýzu reziduí. Obecně lze říci, že jakákoliv nenáhodnost zjištěná u reziduí naznačuje určité nedostatky modelu.

Při analýze reziduí se vychází z *klasických reziduí*, která jsou nestrannými odhady náhodných chyb. Na rozdíl od nich jsou ale korelovaná a mají nekonstantní rozptyl. Místo klasických reziduí se používají další typy reziduí, které mají některé lepší vlastnosti pro statistickou analýzu. Dále se používají různé grafy, zejména grafy reziduí proti hodnotám predikce, hodnotám nezávisle proměnné nebo proti pořadovému číslu pozorování.

Graf reziduí proti nezávisle proměnné hladina (obr. 8) mírně indikuje *heteroskedastický model*. Rozptyl podmíněných rozdělení se nejeví konstantní, má tendenci slabě růst s růstem hodnot nezávisle proměnné. Statistickými testy (Glejser a Goldfeld viz [3]), ale nezamítáme hypotézu o konstantním rozptylu na hladině významnosti 0,01.

Z grafu reziduí proti pořadí měření (obr. 8) je patrné nenáhodné cyklické kolísání, což indikuje, že navržený model není správný. Nenáhodné kolísání může být způsobeno nezařazením proměnné čas do modelu nebo autokorelací náhodných chyb. Pro test nulové hypotézy o nekorelovanosti náhodných chyb proti alternativní hypotéze o korelovanosti sousedních chyb lze použít *Durbin - Watsonovu* statistiku

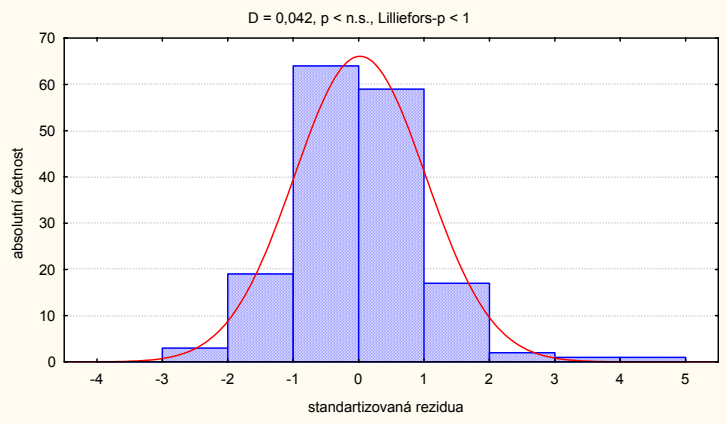
$$D = \frac{\sum_{i=2}^{N} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{N} \varepsilon_i^2}.$$



Obr. 8 Graf klasických reziduí proti nezávisle proměnné a pořadí měření

Proti nulové hypotéze svědčí hodnoty statistiky D vzdálené od čísla 2. Korelovanost náhodných chyb byla potvrzena tímto testem na hladině významnosti 0,01.

Na obr. 9 je histogram rozdělení standardizovaných reziduí, z něhož je patrné nepříliš velké odchýlení od normálního rozdělení. Lilieforsova varianta Kolmogorovova testu shody normální rozdělení na hladině významnosti 0,05 nezamítá.



Obr. 9 Histogram standardizovaných reziduí a jeho porovnání s hustotou normálního rozdělení.

4.3 Postupy při porušení předpokladů KLRM

Pokud lineární model není základní, tj. kovarianční matice náhodné chyby $\boldsymbol{\varepsilon}$ není rovna $\sigma^2\boldsymbol{I}$, ale lze ji vyjádřit ve tvaru $\sigma^2\boldsymbol{W}$, kde matice $\boldsymbol{W}$ není jednotková, mluvíme o tzv. *zobecněném lineárním regresním modelu*. Pokud je matice $\boldsymbol{W}$ regulární, hledáme odhad $\hat{\boldsymbol{\beta}}_Z$ vektoru $\boldsymbol{\beta}$ tzv. *zobecněnou metodu nejmenších čtverců*, tj. hledáme statistiku $\hat{\boldsymbol{\beta}}_Z$, ve která funkce

$$S_Z(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{A\beta})^T \boldsymbol{W}^{-1} (\boldsymbol{Y} - \boldsymbol{A\beta}) = \boldsymbol{\varepsilon}^T \boldsymbol{W}^{-1} \boldsymbol{\varepsilon}$$

nabude absolutního minima.

Z teoretického hlediska je zobecněný lineární regresní model triviálním zobecněním KLRM, na který jej lze převést lineární transformací. Z praktického hlediska je to ale horší, protože matici $\boldsymbol{W}$ obvykle neznáme a musíme najít její odhad. Pokud bychom odhad vektoru $\boldsymbol{\beta}$ v zobecněném modelu hledali stejně jako v KLRM, dostaneme obecně méně přesné odhady regresních parametrů a vychýlený a méně přesný odhad směrodatné odchylky σ modelu. Zobecněný lineární model dostaneme např. v případě nekonstantnosti rozptylu náhodné chyby, tj. v případě tzv. heteroskedastického modelu, ale i v případě, že jsou náhodné chyby korelované, tj. v případě autoregresního modelu. Pomocí zobecněného modelu lze řešit i případy, kdy jsou na regresní parametry kladeny omezující podmínky. My se zde omezíme pouze na případ autokorelace, tj. korelovanosti náhodných chyb. Ostatní případy lze nalézt např. v [3]. S autokorelací se setkáváme především v případech, kdy se pozorování vztahují k různým časovým okamžikům nebo intervalům. Pak se může stát, že náhodné chyby ε_i závisí na předchozích hodnotách, tj. není splněn předpoklad o jejich nekorelovanosti. S autokorelací se můžeme setkat i v případě, že do regresního modelu nejsou zařazeny všechny významné vysvětlující proměnné.

Předpokládejme, tedy, že jsou náhodné chyby ε_i korelovány a řídí se *autoregresním modelem AR*(p) *řádu* p, tj.

$$\varepsilon_i = \varphi_1 \varepsilon_{i-1} + \ldots + \varphi_p \varepsilon_{i-p} + \tau_i \text{ pro } i = 1,\ldots,n,$$

kde φ_1, …, φ_p jsou neznámé parametry a τ_i je jiná náhodná chyba, která splňuje stejné předpoklady jako náhodná chyba v KLRM. V případě $AR(1)$, který se vyskytuje nejčastěji, lze psát

$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i = \varphi_1 \varepsilon_{i-1} + \tau_i$ pro $i = 1,...,n$,

kde

$\varphi_1 = \rho = cor(\varepsilon_i, \varepsilon_{i-1})$, $E(\varepsilon_i) = E(\tau_i) = 0$, $D(\varepsilon_i) = \sigma_\varepsilon^2$, $D(\tau_i) = \sigma_\tau^2$.

Lze ukázat, že

$$\sigma_\varepsilon^2 = \frac{\sigma_\tau^2}{1-\rho^2},\ cov(\varepsilon_i, \varepsilon_{i-j}) = \rho^j \frac{\sigma_\tau^2}{1-\rho^2} = \rho^j \sigma_\varepsilon^2 .$$

Tedy

$cov(\boldsymbol{\varepsilon}) = \sigma_\tau^2 \boldsymbol{W}$,

kde

$$\boldsymbol{W} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}.$$

Model $AR(1)$ lze zapsat ve tvaru

$\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{\varepsilon},\ E(\boldsymbol{\varepsilon}) = \boldsymbol{0},\ cov(\boldsymbol{\varepsilon}) = \sigma_\tau^2 \boldsymbol{W}$.

Inverzní matici $\boldsymbol{W}^{-1}$ k matici $\boldsymbol{W}$ lze vyjádřit ve tvaru $\boldsymbol{W}^{-1} = \boldsymbol{P}^T \boldsymbol{P}$, kde

$$\boldsymbol{P} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Vynásobíme-li $AR(1)$ zleva maticí $\boldsymbol{P}$, převedeme jej na KLRM:

$\boldsymbol{Z} = \boldsymbol{Q\beta} + \tau,\ E(\tau) = \boldsymbol{0},\ cov(\tau) = \sigma_\tau^2 \boldsymbol{I},\ \boldsymbol{Z} = \boldsymbol{PY},\ \boldsymbol{Q} = \boldsymbol{PA},\ \tau = \boldsymbol{P}\varepsilon$.

Zobecněný odhad $\hat{\boldsymbol{\beta}}_{\boldsymbol{Z}}$ parametru $\boldsymbol{\beta}$ pak můžeme hledat v transformovaném KLRM, podobně tam můžeme konstruovat intervalové odhady a testovat hypotézy. Výsledky transformujeme zpět do původního modelu.

Pro test nulové hypotézy o nekorelovanosti náhodných chyb proti alternativní hypotéze o korelovanosti $\varepsilon_i, \varepsilon_{i-1}$ lze použít již zmíněnou *Durbin - Watsonovu* statistiku D. Za odhad ρ lze vzít odhad $\hat{\rho}$ korelačního koeficientu veličin $\varepsilon_i, \varepsilon_{i-1}$

$\hat{\rho} = 1 - 0{,}5D$.

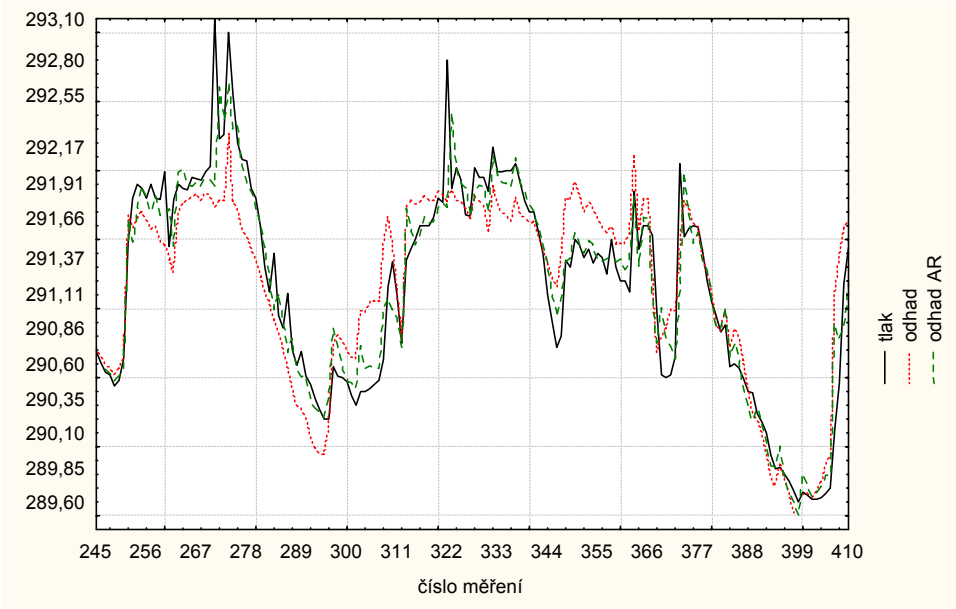
Potom pro odhad $\widetilde{Y}_i$ hodnoty veličiny Y_i pomocí $AR(1)$ dostaneme

$\widetilde{Y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}} + \hat{\rho}(Y_{i-1} - \boldsymbol{x}_{i-1}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{Z}})$ pro $i = 2,...,n$.

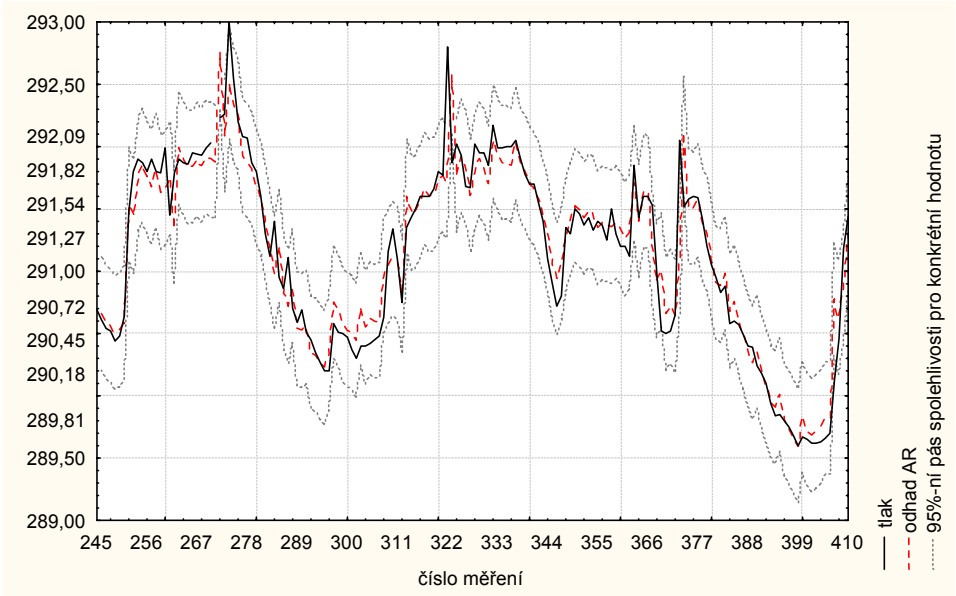
Korelovanost náhodných chyb našich dat byla prokázána. Průběh závislosti tlaku na hladině lépe než KLRM vystihuje $AR(1)$. V případě 1. shluku dostáváme odhad ρ 0,735 a odhad

$\widetilde{y}_i = -6{,}171 + 1{,}040 x_i - 0{,}764 x_{i-1} + 0{,}735 y_{i-1}$ pro $i = 2,...,n$.

Směrodatná chyba autoregresního modelu je 0,224, zatímco pro KLRM je 0,341. Šířka 95%-ního pásu spolehlivosti pro konkrétní hodnotu tlaku je 0,896 oproti šířce 1,364 tohoto pásu v KLRM. Šířka pásu je tedy o 0,234 menší, tj. představuje 65,7% šířky pásu v KLRM. Hodnoty odhadů ve shluku 1 pomocí modelu $AR(1)$ jsou uvedeny v posledním sloupci tab. 2. Na obr. 10 je výsek autoregresním modelem spolu s regresními odhady. Na obr. 11 je pak autokorelační model včetně pásů spolehlivosti pro konkrétní hodnotu tlaku.



Obr. 10 Základní a autoregresní model



Obr. 11 Autoregresní model s pásy spolehlivosti pro konkrétní hodnotu tlaku

5. Závěr

Při statistické analýze závislosti dvou veličin je vhodné vždy vycházet z grafické prezentace dat. Ta umožní odhalit případné defekty v datech, jako jsou např. hrubé chyby měření a heterogenita dat. Hrubé chyby je zapotřebí odstranit a pokud jsou data heterogenní, je zapotřebí je rozložit do homogenních skupin a analýzu provádět s každým shlukem zvlášť. Identifikaci shluků realizujeme shlukovou analýzou nebo na základě znalosti poměrů na vodním díle. Při samotné analýze závislosti sledujeme dvě její vlastnosti - těsnost a průběh. Orientační informace o těchto aspektech získáme z bodového diagramu dat. Pro měření těsnosti závislosti dvou veličin se nejčastěji používá korelační koeficient, který měří těsnost lineární závislosti. Průběh závislosti modelujeme pomocí regresní funkce, u které se předpokládá, že je znám její tvar. Regresní parametry se v aplikacích běžně odhadují metodou nejmenších čtverců. Běžně

se ale neověřují podmínky její použitelnosti, které zaručují dobré vlastnosti získaných odhadů, ani se nekonstruují intervalové odhady konkrétních hodnot závisle proměnné, jejichž šířka mimo jiné svědčí také o kvalitách odhadů. Je tedy třeba znát předpoklady metody nejmenších čtverců a umět je ověřit. V tomto směru je důležitá analýza reziduí, která umožní nejen ověřit dané předpoklady, ale i naznačí, jak model opravit v případě, že tyto předpoklady nejsou splněny.

Pozn. *Příspěvek byl zpracován v rámci řešení grantového projektu GAČR 103/05/2391.*

Literatura

[1] Anděl, J. *Matematická statistika*. MATFYZPRESS Praha, 1993.
[2] Budíková, M. *Aplikace shlukové analýzy v ekologii*. Sborník prací 11. letní školy ROBUST 2000.
[3] Hebák, P. - Hustopecký, J. *Vícerozměrné statistické metody s aplikacemi*. SNTL Praha, 1987.
[4] Hendl, J. *Přehled statistických metod zpracování dat*. PORTÁL Praha, 2006.
[5] Meloun, M. - Militký, J. Statistické zpracování experimentálních dat. ARS MAGNA Praha, 1998.
[6] STATISTICA for Windows. StatSoft, Inc. 2000.