

STOCHASTICKÁ ANALÝZA NESTACIONÁRNÍCH PROCESŮ V MATLABU

David Kvapil

UNIS, a.s., Brno

Abstrakt

Příspěvek popisuje stochastickou analýzu a matlabovské modelování nestacionárních procesů v technometrii. Sumarizují se známé poznatky o stabilizaci trendu i rozptylu, navrhují se možný způsob řešení situace, kdy data vykazují nestacionaritu ve střední hodnotě i v rozptylu. Empiricky zjištěná provozní teplárenská data, tvořící vysokofrekvenční časové řady, se vyznačují v čase proměnlivou variabilitou (měnlivá volatilita), což vede k vážným problémům při použití klasických lineárních (S)AR(I)MA modelů (nesplnění homoskedasticity a normality). Jsou pozorovány shluky volatility, je indikováno leptokurtické pravděpodobnostní rozdělení. Je použito GARCH modelování, pracující s (v čase se měnícími) podmíněnými statistikami – podmíněnou střední hodnotou a podmíněným rozptylem. Pro samotnou analýzu byly vytvořeny vlastní matlabovské skripty, stejně jako jsou využívány běžné matlabovské funkce statistického, optimalizačního a GARCH toolboxu. Na provozních datech je konkrétně demonstrován postup při vybudování GARCH modelu.

1 Stabilizace trendu

Při modelování reálných provozních teplárenských dat se dospělo k situaci, kdy zkoumané časové řady měřených fyzikálních veličin vykazovaly nestacionaritu ve střední hodnotě i nestacionaritu v rozptylu. Pokusíme se sumarizovat známé poznatky o stabilizaci trendu i rozptylu, navrhnut možný způsob řešení takovéto situace a na ukázkovém příkladě naznačíme konkrétní postup stochastické analýzy a modelování ARCH procesů v Matlabu.

Pro procesy nestacionární ve střední hodnotě rozlišujeme deterministický trend a stochastický trend. U deterministického trendu chápeme nestacionaritu jako funkci času, tj. k modelování používáme např. polynomický trend, příp. periodický trend

$$f(t) = \beta_0 + \beta_1 t + \dots + \beta_d t^d, \text{ resp. } f(t) = \mu + \sum_{j=1}^p (\alpha_j \cos \lambda_j t + \beta_j \sin \lambda_j t).$$

Pro modely ARMA(p, q): $\Phi(B)Y_t = \Theta(B)\varepsilon_t$ je požadováno, aby proces byl kauzální, tj. všechny kořeny polynomu $\Phi(z) = 1 - \varphi_1 z - \dots - \varphi_p z^p$ ležely vně jednotkové kružnice. Pokud nějaký kořen leží na jednotkové kružnici, mluvíme o procesu nestacionárním se stochastickým trendem, pokud leží uvnitř jednotkového kruhu, mluvíme o nestacionárním procesu explozivního typu. Nestacionární proces obsahující stochastický trend lze převést na stacionární diferencováním. Zavedeme diferenční operátor

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - B)Y_t,$$

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2} = (1 - B)^2 Y_t,$$

⋮

$$\Delta^d Y_t = Y_t - \binom{d}{1} Y_{t-1} + \binom{d}{2} Y_{t-2} - \dots + (-1)^d Y_{t-d} = (1 - B)^d Y_t.$$

Nestacionární proces se stochastickým trendem se nazývá integrovaný smíšený model ARIMA(p, d, q), formálně jej zapisujeme pomocí operátoru zpětného chodu

$$\text{ARIMA}(p,d,q): \Phi(B)(1-B)^d Y_t = \Theta(B) \varepsilon_t$$

a položíme-li $W_t = (1-B)^d Y_t$, pak je W_t stacionární ARMA(p,q) proces.

Poznamenejme, že číslo d nemusí být obecně celé. Pak jej nazýváme frakcionální parametr, operátor $\Delta^d = (1-B)^d$ vyjadřuje frakcionální diferenci a hovoříme o frakcionálně integrovaných procesech ARFIMA(p,d,q). Jde o procesy s dlouhou pamětí, jejich charakteristickou vlastností je, že hodnoty ACF neklesají s rostoucím zpožděním exponenciálně, ale hyperbolicky.

2 Stabilizace rozptylu

Procesy, kde není splněna podmínka neměnnosti rozptylu v čase (homoskedasticita), je většinou možné vhodně transformovat. Situace nestabilního rozptylu nastává především v případech, kdy náhodná veličina Y_t má rozdělení závislé na jediném parametru, který obecně nemusí mít pro všechna t stejnou hodnotu. Hledá se pak netriviální funkce g tak, aby náhodná veličina $Z_t = g(Y_t)$ měla rozptyl nezávisející na t . Tato úloha obecně nemá řešení, používá se však určitých užitečných approximací.

Při předpokladu exponenciální závislosti směrodatné odchylky na střední hodnotě $\sigma_t = \sigma_0 \mu_t^\Theta$ se nejčastěji se uvažují případy od $\Theta = 0$ (řada je homoskedastická) až po $\Theta = 1$ (pak existuje lineární závislost). Při volbě $\lambda = 1 - \Theta$ lze odstranit heteroskedasticitu (pro řadu $Y_t > 0$) např. jednou z následujících transformací:

Box – Coxova mocninná transformace

$$Z_t = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda} & \text{pro } \lambda \neq 0, \\ \ln(Y_t) & \text{pro } \lambda = 0, \end{cases}$$

příp. mocninná transformace

$$Z_t = \begin{cases} Y_t^\lambda & \text{pro } \lambda \neq 0, \\ \ln(Y_t) & \text{pro } \lambda = 0. \end{cases}$$

Pokud hodnoty náhodné veličiny nejsou kladné, můžeme použít jednu z následujících transformací:

Box – Coxova mocninná transformace s posunutím

$$Z_t = \begin{cases} \frac{(Y_t + a)^\lambda - 1}{\lambda} & \text{pro } \lambda \neq 0, \\ \ln(Y_t + a) & \text{pro } \lambda = 0, \end{cases}$$

kde a je takové reálné číslo, aby všechny realizace byly kladné,

příp. mocninná transformace se znaménky

$$Z_t = \begin{cases} \text{sgn}(Y_t) \frac{|Y_t|^\lambda - 1}{\lambda} & \text{pro } \lambda \neq 0, \\ \text{sgn}(Y_t) \ln(Y_t) & \text{pro } \lambda = 0. \end{cases}$$

Poznamenejme ale, že při nesplnění podmínky $Y_t > 0$ nebo když pozorované hodnoty jsou blízké nule, je možno řadu transformovat posunutím pouze s vědomím značného rizika znehodnocení původní řady a znevěrohodnění výsledného modelování. Navíc mocninná transformace není spojitá pro $\lambda \rightarrow 0$, proto je třeba se vyhýbat malým nenulovým hodnotám parametru λ .

Odhad transformačního parametru mocninné transformace se pak provede pomocí maxima logaritmu věrohodnostní funkce

$$l^*(\lambda) = -\frac{n}{2} \ln(s^2(\lambda)) + (\lambda - 1) \sum_{i=1}^n \ln y_i, \text{ kde } s^2 = \hat{\sigma}^2,$$

$$z_i = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \ln(y_i) & \lambda = 0, \end{cases}$$

pro ekvidistantní hodnoty $\lambda_1 < \lambda_2 < \dots < \lambda_m$ (pro dostatečně velké m) z vhodně zvoleného intervalu $(\lambda_1^*, \lambda_2^*)$, kde $\lambda_1^*, \lambda_2^* \in \mathbf{R}$, a $\lambda_1^* < \lambda_2^*$, vypočítá hodnoty $l^*(\lambda)$ a hledá argument $\hat{\lambda}$ maxima těchto hodnot. Bylo odvozeno asymptotické rozdělení statistiky

$$K = -2[l^*(\lambda) - l^*(\hat{\lambda})] \sim \chi^2(1)$$

a lze takto zkonstruovat jednostranný asymptotický interval spolehlivosti pro parametr λ

$$1 - \alpha = P(K < \chi^2_{1-\alpha}(1)) = P(-2[l^*(\lambda) - l^*(\hat{\lambda})] < \chi^2_{1-\alpha}(1)) = P(l^*(\hat{\lambda}) - \frac{1}{2}\chi^2_{1-\alpha}(1) \leq l^*(\lambda)),$$

tj. všechna λ splňující nerovnost

$$l^*(\lambda) \geq D_\alpha = l^*(\hat{\lambda}) - \frac{1}{2}\chi^2_{1-\alpha}(1)$$

leží v intervalu spolehlivosti a jsou tedy přijatelná.

Testování hypotéz: testujeme $H_0: \lambda = \lambda_0$ proti alternativě $H_1: \lambda > \lambda_0$. Nejdříve budeme testovat $H_0^1: \lambda = 1$. Pokud H_0^1 nezamítneme, tj. $l^*(1) \geq D_\alpha$, nemusíme data transformovat. Pokud předchozí hypotézu zamítneme, můžeme testovat další $H_0^2: \lambda = 0$. Pokud H_0^2 nezamítneme, tj. $l^*(0) \geq D_\alpha \wedge l^*(1) < D_\alpha$, transformace bude ve tvaru $z_i = \ln y_i$. Pokud však bude $l^*(0) < D_\alpha \wedge l^*(1) < D_\alpha$, provedeme transformaci $z_i = (y_i^{\hat{\lambda}} - 1)/\hat{\lambda}$.

Algoritmus pro řešení praktických úloh:

1. Zkontrolujeme vstupní data, aby byla nezáporná, jinak příp. přičteme kladnou konstantu.
2. Vektor dat rozložíme na krátké úseky o délce 4 až 12 údajů.
3. V každém úseku provedeme robustní odhad střední hodnoty $\hat{\mu}_i$ (průměr, medián) a rozptylu $\hat{\sigma}_i^2$ (max-min, mezikvartilové rozpětí).
4. Předpokládáme, že platí $\sigma(\mu) = \sigma\mu^\Theta$, logaritmováním dostaneme $\ln \sigma(\mu) = \ln \sigma + \Theta \ln \mu$ a neznámé Θ odhadneme metodou nejmenších čtverců díky hodnotám $v_i = \ln \hat{\sigma}_i$, $u_i = \ln \hat{\mu}_i$ v regresním modelu $v_i = a + \Theta u_i + \varepsilon_i$, kde $a = \ln \sigma$, $\varepsilon_i \sim WN(0, \sigma_\varepsilon^2)$.
5. Pro odhad $\hat{\Theta} = 1 - \hat{\lambda}$ pomocí t -statistiky zkonstruujeme interval spolehlivosti $I(\hat{\Theta})$. Pokud tento interval bude obsahovat nulu, tj. $0 \in I(\hat{\Theta})$, data se nebudou transformovat. Pokud $0 \notin I(\hat{\Theta}) \wedge 1 \in I(\hat{\Theta})$, volí se transformace $z_i = \ln y_i$. Jinak se volí $z_i = (y_i^{\hat{\lambda}} - 1)/\hat{\lambda}$.

3 Procesy s proměnlivými režimy

Jiným východiskem při řešení tohoto problému je přistoupení k modelování procesů s proměnlivými režimy – modely TAR (Threshold Autoregressive), modely MSW (Markov Switching), případně modely ARCH/GARCH (Generalized Autoregressive Conditional Heteroscedasticity).

Empiricky bylo zjištěno, že vysokofrekvenční časové řady se vyznačují v čase proměnlivou variabilitou, hovoří se o měnlivé volatilitě. To vede k vážným problémům při použití klasických lineárních (S)AR(I)MA modelů (Box – Jenkinsova metodologie). Bylo zjištěno, že s variabilitou může souviset úroveň a síla autokorelace v časových řadách. Charakteristické rysy takovýchto analyzovaných časových řad tedy nemohou být plně zachyceny jen lineárními modely, předpokládající pouze jeden typ závislosti – korelační závislost. Nelineární modely vycházejí z nějaké nelineární funkce řady stejně rozdělených nezávislých náhodných veličin, takže předpokládají obecnější formu závislosti než pouze korelační. Změnu volatility (tedy i autokorelace) časové řady lze chápat jako změnu v režimu chování časové řady, která může být způsobena různými faktory – deterministickými i nesystematickými a nepredikovatelnými.

V 80. letech 20. stol. (Engle) vznikla koncepce modelování volatility časových řad (modely typu ARCH), která byla postupně rozpracována a byla navržena řada dalších modelů volatility. Svůj původ mají tyto modely v ekonometrii, postupně se začínají uplatňovat i v technometrii.

Modely ARCH pracují s podmíněnou střední hodnotou a podmíněným rozptylem. Mějme např. stacionární autoregresní model prvního řádu AR(1)

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

kde $|\phi| < 1$, $\varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$, který lze vyjádřit ve formě

$$X_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \phi^3 \varepsilon_{t-3} + \dots$$

Nepodmíněná střední hodnota veličiny X_t je nulová, tedy $E(X_t) = 0$. Podmíněná střední hodnota E_{t-1} je střední hodnota veličiny X_t za předpokladu, že náhodné veličiny nabýly v časech $t-1, t-2, \dots$ konkrétních hodnot. Závisí na volbě podmíny, je její funkcí – tuto funkci označujeme jako funkci regresní. Pro proces AR(1) je podmínkou určitá hodnota náhodné veličiny v čase $t-1$,

$$E_{t-1}(X_t) = \phi X_{t-1},$$

tj. podmíněná střední hodnota veličiny X_t je závislá na čase. Podmíněnou střední hodnotu lze vyjádřit obecným ARMAX(p, q, n) modelem

$$X_t = C + \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \sum_{k=1}^n \beta_k A_{t,k},$$

kde $A_{t,k}$ je regresní matice.

Nepodmíněný rozptyl veličiny X_t pro proces AR(1) je

$$D_t(X_t) = \frac{\sigma_\varepsilon^2}{1 - \phi^2},$$

je tedy neměnný v čase. Podmíněný rozptyl se označuje jako funkce skedastická, jejíž průběh tak charakterizuje měnovost rozptylu veličiny X_t v závislosti na hodnotách veličin X_{t-1}, X_{t-2}, \dots . Podmíněný rozptyl veličiny X_t je $D_{t-1}(X_t) = \sigma_\varepsilon^2$, je tedy také neměnný v čase (je funkcí homoskedastickou).

Uvedené vlastnosti jsou charakteristické pro všechny stacionární a invertibilní procesy, tj. pro procesy typu AR, MA a ARMA.

Podívejme se nyní na nestacionární procesy, konkrétně integrované procesy. Uvažujme proces náhodné procházky

$$X_t = X_{t-1} + \varepsilon_t,$$

kde $\varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$, který lze vyjádřit ve tvaru $X_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \varepsilon_{t-3} + \dots$, takže nepodmíněná střední hodnota je nulová, tj. $E(X_t) = 0$. Podmíněná střední hodnota $E_{t-1}(X_t) = X_{t-1}$ je závislá na čase. Nepodmíněný rozptyl $D_t(X_t) = t\sigma_\varepsilon^2$ je lineární funkcí časové proměnné. Podmíněný rozptyl $D_{t-1}(X_t) = \sigma_\varepsilon^2$ je opět funkcí homoskedastickou.

Tyto vlastnosti jsou charakteristické pro všechny typy integrovaných procesů, tj. pro třídu modelů ARIMA. S uvedenými typy modelů se pracuje velmi dobře, neboť je lze jednoduše transformovat na gaussovský proces bílého šumu. Problematika bodových a intervalových odhadů parametrů tohoto procesu je ve statistické teorii dobře rozpracována.

Prakticky se při identifikaci modelu časové řady postupuje následovně: vybere se nějaký model, nejčastěji podle tvaru ACF a PACF, odhadnou se jeho parametry. Ve druhé fázi se provádí diagnostická kontrola modelu, ve které se ověřuje, zda příslušná transformace analyzovaného procesu vede k procesu gaussovského bílého šumu. Na základě odhadů parametrů se vypočítají residua a jejich prostřednictvím se testuje autokorelace, heteroskedasticita a normalita nesystematické složky modelu.

V případě většiny empirických teplárenských denních časových řad se vyskytuje situace, kdy nejsou splněny podmínky, za kterých lze použít lineární modely typu ARMA či ARIMA – není splněna především podmínka homoskedasticity a normality. Pokud provedeme grafické znázornění rozdělení hodnot časové řady, zjistíme, že má tzv. leptokurtické rozdělení pravděpodobnosti – takové rozdělení je špičatější a má „tlustší (delší) konce“ (Fat Tails) ve srovnání s rozdělením normálním. Existují dvě koncepce řešení problému nesplnění podmínek lineárního modelu třídy AR(I)MA.

První koncepce (Engle, 1984) vychází z představy, že problém je v typu modelu časové řady. Tato představa je založena na myšlence, že podíváme-li se na analyzovanou časovou řadu, všimneme si, že se její variabilita stejně jako úroveň v čase mění. Uvažované lineární modely jsou totiž založeny na podmínce, že se sice podmíněná střední hodnota v čase mění, ale podmíněný rozptyl je konstantní, což však mnohdy neodpovídá realitě. Je tedy vhodné navrhnout modely, které by splňovaly předpoklad v čase se měnícího podmíněného rozptylu (příp. podmíněné střední hodnoty a podmíněného rozptylu). Podstatné je, že tato koncepce nemění původní požadavek normality.

Druhá koncepce vychází z představy (Mandelbrot, 60.léta), že problém není v modelu časové řady, ale v požadavku normality rozdělení, který není reálný. Většina denních časových řad je charakteristická rozdělením, které je špičatější a má tlustší konce než rozdělení normální. Tato druhá koncepce má fundamentální charakter a její další rozpracování by znamenalo revoluční zásah do celé oblasti statistického modelování.

4 Modely volatility

Uvažujme stochastický proces $\{\varepsilon_t\}$, u kterého předpokládáme nulovou podmíněnou střední hodnotu

$$E(\varepsilon_t | \Omega_{t-1}) = E_{t-1}(\varepsilon_t) = 0$$

a podmíněný rozptyl

$$D(\varepsilon_t | \Omega_{t-1}) = D_{t-1}(\varepsilon_t) = E_{t-1}\{\varepsilon_t - E_{t-1}(\varepsilon_t)\}^2 = E_{t-1}(\varepsilon_t^2 | \Omega_{t-1}) = \sigma_t^2,$$

přičemž Ω_{t-1} je relevantní minulá informace až do času $t-1$. Tyto požadavky splňuje model procesu $\{\varepsilon_t\}$ ve tvaru

$$\varepsilon_t = e_t \sigma_t,$$

kde $E_{t-1}(e_t) = 0$, $D_{t-1}(e_t) = 1$, $E(e_t e_{t-i}) = 0$ pro $i = 0, \pm 1, \pm 2, \dots$ (nezávislé veličiny s nulovou střední hodnotou a jednotkovým rozptylem). Je-li rozdělení náhodné veličiny e_t za podmínky informace, která je k dispozici v čase $t-1$, normované normální, tj. $e_t \sim N_{t-1}(0, 1)$, potom je rozdělení náhodné veličiny ε_t za podmínky informace, která je k dispozici v čase $t-1$, rovněž normální, avšak s podmíněným rozptylem, který se mění v závislosti na čase, tj. $\varepsilon_t \sim N_{t-1}(0, \sigma_t)$. Dále platí

$$\frac{E(\varepsilon_t^4)}{E(\varepsilon_t^2)^2} \geq E(e_t^4),$$

tj. špičatost nepodmíněného rozdělení ε_t je větší nebo rovna špičatosti normovaného normálního rozdělení, přičemž rovnost nastane tehdy, jsou-li podmíněné rozptyly konstantní. Jednotlivé modely volatility spočívají ve formulaci vývoje podmíněného rozptylu σ_t^2 v čase.

Lineární modely volatility jsou charakteristické tím, že podmíněný rozptyl je lineární funkcí veličin $\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \dots, \varepsilon_{t-p}^2$. Nelineární modely jsou schopné postihnout různé asymetrické jevy, např. pákový efekt, kdy se kladné a záporné šoky do podmíněného rozptylu nepromítají symetricky, jak to předpokládají lineární modely volatility. Základními lineárními modely volatility jsou ARCH(q) a GARCH(p, q) modely.

Proces ARCH(q)

Podmíněný rozptyl je

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_q \varepsilon_{t-q}^2,$$

kde $\omega > 0$ a $\alpha_1, \dots, \alpha_q \geq 0$. Toto lze zapsat i pomocí operátoru zpětného posunutí B , pro který je $B^j \varepsilon_t = \varepsilon_{t-j}$, pak

$$\sigma_t^2 = \omega + (\alpha_1 B + \alpha_2 B^2 + \dots + \alpha_q B^q) \varepsilon_t^2.$$

Model ARCH(q) lze vyjádřit v autoregresním tvaru modelu AR(q) procesu $\{\varepsilon_t^2\}$

$$\varepsilon_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 + \nu_t,$$

kde $\nu_t = \varepsilon_t^2 - \sigma_t^2$. Proces ARCH(q) je slabě stacionární, leží-li kořeny polynomiální rovnice $(\alpha_1 B + \alpha_2 B^2 + \dots + \alpha_q B^q) = 0$ vně jednotkového kruhu. Nepodmíněný rozptyl procesu je

$$D(\varepsilon_t) = E(\varepsilon_t^2) = \frac{\omega}{1 - \sum_{i=1}^q \alpha_i},$$

tedy je konstantní v čase a proces $\{\varepsilon_t\}$ je nepodmíněně homoskedastický.

Zejména pak model ARCH(1) má tvar podmíněného rozptylu

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2.$$

Úpravou lze získat autoregresní tvar modelu

$$\varepsilon_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \nu_t,$$

kde $\nu_t = \varepsilon_t^2 - \sigma_t^2 = \sigma_t^2 (e_t^2 - 1)$. Jestliže $\alpha_1 = 0$, pak je $\{\varepsilon_t\}$ proces gaussovského bílého šumu (za podmínky normality). Je-li α_1 příliš vysoké, je rozptyl procesu ARCH(1) nekonečně velký. Proces je slabě stacionární, jestliže $\alpha_1 < 1$. Nepodmíněná střední hodnota, resp. nepodmíněný rozptyl jsou

$$E(\varepsilon_t) = 0, \text{ resp. } D(\varepsilon_t) = E(\varepsilon_t^2) = \frac{\omega}{1 - \alpha_1}.$$

Podmíněný rozptyl je kladný pro $\omega > 0$ a $\alpha_1 \geq 0$. Čtvrtý moment je konečný pro $3\alpha_1^2 < 1$, tehdy proces generuje data, která mají rozdělení špičatější a jeho konce jsou „tlustší“ v porovnání s rozdělením normálním,

$$K_\varepsilon = \frac{E(\varepsilon_t^4)}{E(\varepsilon_t^2)^2} = \frac{3(1 - \alpha_1^2)}{1 - 3\alpha_1^2}.$$

Model ARCH je charakteristický tím, že lze pomocí něj zachytit shluky volatility v časové řadě. To plyne z toho, že jestliže je ε_{t-1} v absolutní hodnotě vysoké, lze očekávat i ε_t v absolutní hodnotě vysoké. Vysoké (příp. nízké) hodnoty časové řady budou následovány vysokými (příp. nízkými) hodnotami jakéhokoliv znaménka.

Proces GARCH(p,q)

Má podmíněný rozptyl ve tvaru

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \dots + \beta_p \sigma_{t-p}^2,$$

kde $p \geq 0$, $q > 0$, $\omega > 0$, $\alpha_i \geq 0$ pro $i = 1, \dots, q$, $\beta_j \geq 0$ pro $j = 1, \dots, p$. Pomocí operátoru zpětného posunutí jej lze vyjádřit ve tvaru

$$\begin{aligned} \sigma_t^2 &= \omega (\alpha_1 B + \alpha_2 B^2 + \dots + \alpha_q B^q) \varepsilon_t^2 + (\beta_1 B + \beta_2 B^2 + \dots + \beta_p B^p) \sigma_t^2 = \\ &= \omega + \alpha_q (B) \varepsilon_t^2 + \beta_p (B) \sigma_t^2. \end{aligned}$$

Model GARCH(p,q) je tak možné psát ve tvaru

$$\varepsilon_t^2 = \omega + \sum_{i=1}^m (\alpha_i + \beta_i) \varepsilon_{t-i}^2 - \sum_{i=1}^p \beta_i \nu_{t-i} + \nu_t$$

nebo pomocí operátoru zpětného posunutí

$$\varepsilon_t^2 = \omega + [\alpha(B) + \beta(B)] \varepsilon_t^2 + \nu_t - \beta(B) \nu_t,$$

kde $\nu_t = \varepsilon_t^2 - \sigma_t^2$. Jedná se tedy o model ARMA(m,q) pro ε_t^2 , kde $m = \max\{p, q\}$. Jestliže je $p = 0$, model se transformuje na ARCH(q). Je-li $p = 0 = q$, pak je $\{\varepsilon_t\}$ proces gaussovského bílého šumu. Model GARCH(p,q) lze vyjádřit jako ARCH(∞), tedy model ARCH s mnoha zpožděními lze approximovat modelem GARCH s méně zpožděními, resp. s menším počtem parametrů. Model GARCH(p,q) je slabě stacionární, jestliže $\alpha_p(1) + \beta_q(1) < 1$. Nepodmíněný rozptyl procesu je

$$D(X_t) = E(X_t^2) = \frac{\alpha_0}{1 - \alpha_p(1) - \beta_q(1)}.$$

Zejména pak podmíněný rozptyl procesu GARCH(1,1) má tvar

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1},$$

kde $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$. Pomocí operátoru zpětného posunutí píšeme

$$(1 - \beta_1 B) \sigma_t = \omega + \alpha_1 \varepsilon_{t-1}^2.$$

Úpravou jej můžeme přepsat do tvaru modelu ARMA(1,1)

$$\varepsilon_t^2 = \omega + (\alpha_1 + \beta_1) \varepsilon_{t-1}^2 + v_t - \beta_1 v_{t-1},$$

kde $v_t = \varepsilon_t^2 - \sigma_t^2$. Proces je slabě stacionární pro $\alpha_1 + \beta_1 < 1$. Nepodmíněný rozptyl procesu je

$$D(\varepsilon_t) = E(\varepsilon_t^2) = \frac{\omega}{1 - \alpha_1 - \beta_1},$$

tedy je konstantní v čase a proces $\{\varepsilon_t\}$ je nepodmíněně homoskedastický. Proces GARCH(1,1) stejně jako procesy ARCH generují řady hodnot, které mají rozdělení špičatější a s „tlustšími konci“ v porovnání s normálním rozdělením. Špičatost rozdělení náhodných veličin ε_t má pro $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ tvar

$$K = \frac{E(\varepsilon_t^4)}{E(\varepsilon_t^2)^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2},$$

jinak je ∞ . Autokorelační funkce procesu $\{\varepsilon_t^2\}$ je

$$\rho_1 = \alpha_1 + \frac{\alpha_1^2 \beta_1}{1 - 2\alpha_1 \beta_1 - \beta_1^2}, \quad \rho_k = \rho_1 (\alpha_1 + \beta_1)^{k-1} \text{ pro } k = 2, 3, \dots$$

Hodnoty autokorelační funkce s rostoucím zpožděním k exponenciálně klesají, rychlosť poklesu závisí na velikosti součtu $\alpha_1 + \beta_1$. Blíží-li se tento součet hodnotě jedna, je pokles autokorelační funkce velmi pozvolný. Hodnoty parciální autokorelační funkce rovněž s rostoucím zpožděním exponenciálně klesají. Obecně lze tvrdit, že tvar ACF a PACF procesu $\{\varepsilon_t^2\}$ odpovídá tvaru těchto funkcí modelu ARMA(p,q).

Proces GARCH(1,1) je z praktického hlediska zajímavý tím, že jím lze approximovat proces ARCH s mnoha zpožděními.

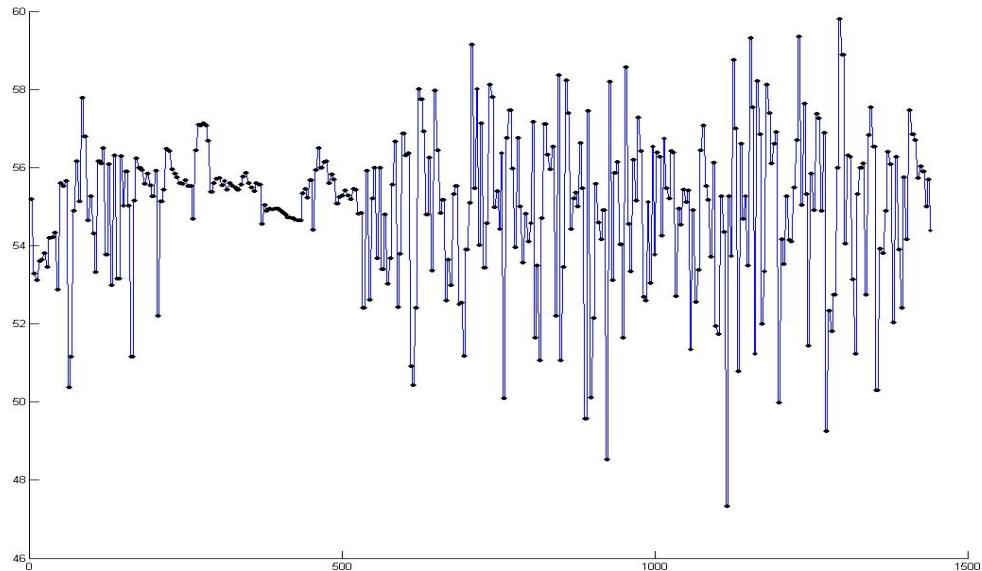
Dalšími lineární modely jsou IGARCH, FIGARCH, GARCH-M, nelineární modely jsou pak EGARCH, IEGARCH, FIEGARCH, GJR-GARCH, QGARCH, SV model aj.

Proces výstavby modelů volatility lze rozdělit do následujících kroků:

- Určení vhodného lineárního nebo nelineárního úrovňového modelu pro danou časovou řadu.
- Testování nulové hypotézy podmíněné homoskedasticity proti alternativní hypotéze podmíněné heteroskedasticity lineárního nebo nelineárního typu.
- Odhadování parametrů zvoleného lineárního nebo nelineárního modelu podmíněné heteroskedasticity.
- Ověření vhodnosti daného modelu diagnostickými testy.
- Modifikace modelu, je-li to třeba.
- Užití modelu pro popisné nebo predikční účely.

5 Analýza provozních teplárenských dat

Stručně naznačíme průběh stochastické analýzy získaných reálných provozních teplárenských dat. Na následujícím obrázku je graficky znázorněn časový průběhu hodnot teplot výstupní vody na výstupu z výměníkové stanice (Teplo Přerov, a.s., Palackého ulice) během jednoho dne (leden, 2007).

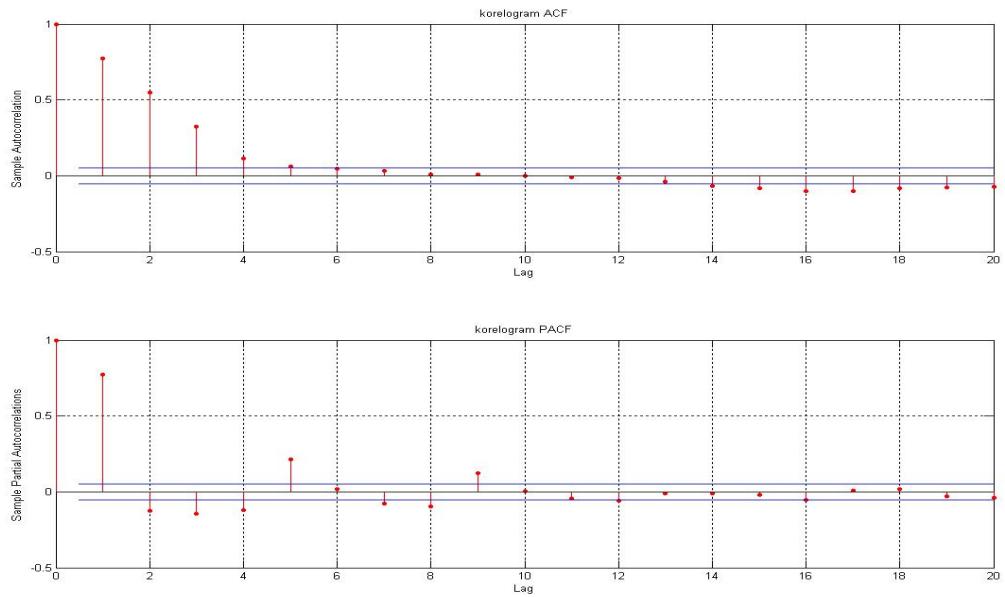


Obr.1: Časový průběhu hodnot teplot výstupní vody na výstupu z výměníkové stanice

Pro vektor dat (označíme jej A) vytvoříme matlabovský skript `BJmodel.m`, který provede identifikaci Box – Jenkinsova modelu. Konkrétně z daného vektoru dat spočítá výběrový průměr, výběrový rozptyl, zobrazí korelogramy ACF a PACF, zobrazí periodogram a spočítá matici tvořenou hodnotami FPE kriteria pro jednotlivé ARMA modely – prvek s minimální hodnotou na pozici $[i, j]$ indikuje model $\text{ARMA}(i-1, j-1)$ (indexuje se od nuly). Dostáváme

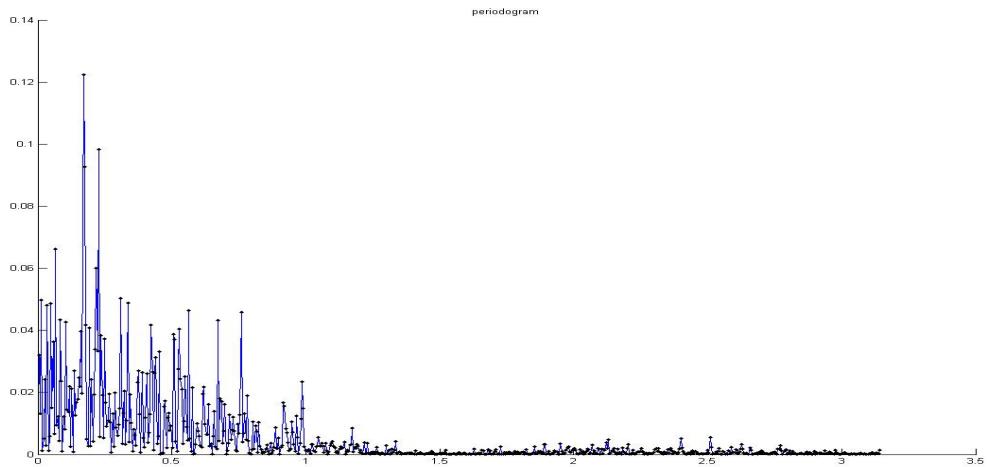
```
>> BJmodel(A)
vyberovy prumer: 55.0193
vyberovy rozptyl: 3.60564
hodnoty FPE kriteria pro jednotlive ARMA modely:
1.0e+003 *
3.030731706022    0.909115365313    0.324244052345    0.166364044939
0.001631776099    0.001634188044    0.001636436964    0.001638666986
0.001634048368    0.001636455241    0.001638704943    0.001637861626
0.001637473969    0.001639879685    0.001641482159    0.001642407768
```

Z této matice se jako nejvhodnější jeví model $\text{ARMA}(1,0)$, neboť prvek na pozici $(2,1)$ má nejmenší hodnotu. Korelogramy na následujícím obrázku také naznačuje, že by se mohlo jednat o proces $\text{ARMA}(1,0)$.



Obr.2: Koreogramy ACF a PACF

Periodogram na následujícím obrázku neindikuje existenci nějakých význačných frekvencí.



Obr.3: Periodogram

Vytvoříme matlabovský skript `ARMAmodel.m`, který může pomoci v ověření správnosti volby modelu ARMA(1,0). Je založen na ověřování nekorelace residuů pomocí Portmonteau testu, přičemž by mělo platit, že hodnota Portmonteau statistiky Q by měla být menší než kritická hodnota testu. Navíc spočítá hodnoty střední residuální chyby, střední kvadratické chyby a střední absolutní chyby. Dostáváme tak:

```
>> ARMAmodel(A,1,0)
Discrete-time IDPOLY model: A(q)y(t) = e(t)
A(q) = 1 - 0.9731 (+-0.01949) q^-1
Estimated using ARMAX from data set yc
Loss function 1.58739 and FPE 1.5896
Sampling interval: 1
```

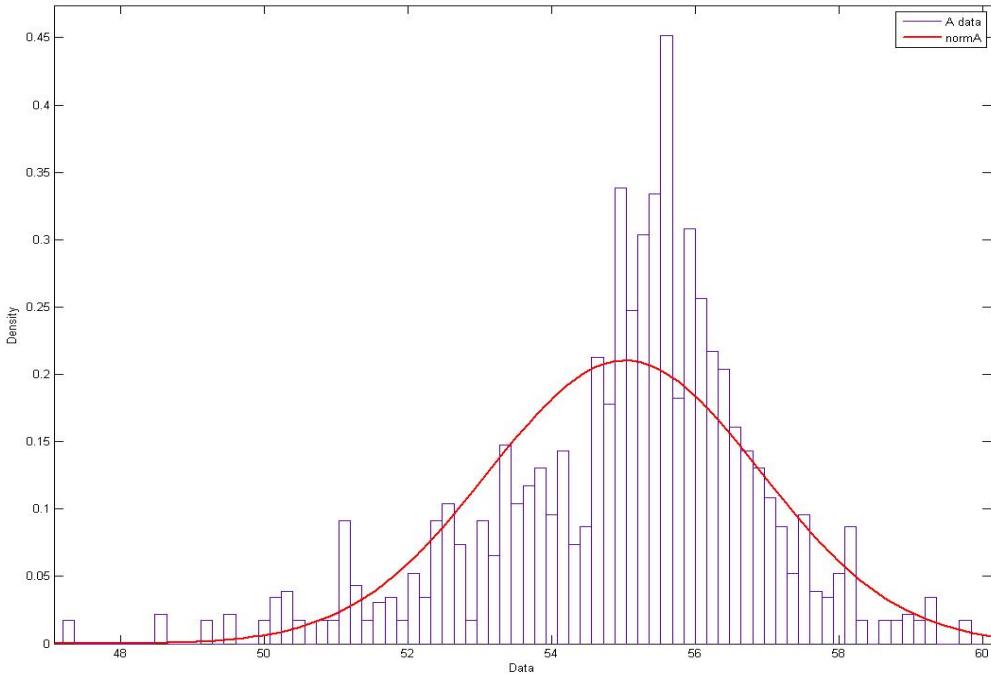
hodnota Portmonteau statistiky: $Q = 251.591$

```

kriticka hodnota : krit = 52.1923
prumerna rezidualni chyba: ME = -0.000543812
prum. kvadraticka chyba: MSE = 1.58627
prum. absolutni chyba: MAE = 0.457015

```

Vidíme, že hodnota Portmonteau statistiky $Q = 251.591$ je podstatně vyšší než kritická hodnota testu $k = 52.1923$. Sestavíme histogram naměřených dat a porovnáme jej s normálním rozdělením.



Obr.4: Histogram naměřených dat

Vidíme, že se zřejmě jedná o data pocházející z leptokurtického rozdělení, tedy špičatější rozdělení s „tlustějšími“ konci oproti rozdělení normálnímu.

Test normality: matlabovský `jbtest` provede Jarque – Berův test nulové hypotézy, že data pocházejí z normálního rozdělení s neznámou střední hodnotou a neznámým rozptylem, proti alternativě, že data nepocházejí z normálního rozdělení. Test je založen na myšlence současného testování šíkmosti a špičatosti. Kromě nenormality může k zamítnutí nulové hypotézy vést i fakt, že hodnoty nejsou homoskedastické. Testovací statistika má tvar

$$JB = \frac{n}{6} \left(s^2 + \frac{(k-3)^2}{4} \right),$$

kde n je počet dat ve vzorku, s , resp. k je výběrová šíkmost, resp. špičatost. Za platnosti nulové hypotézy má χ^2 rozdělení se dvěma stupni volnosti.

```

>> [h,p,jbstat,krit] = jbtest(A, 0.05);
>> [h,p,jbstat,krit]
ans =
    1.00000000000          0.00100000000        227.04309899435      5.94951665714

```

Tedy nulovou hypotézu na 5%-ní hladině významnosti zamítáme, data nepocházejí z normálního rozdělení.

Vytvoříme matlabovký skript `testnezprvku.m` (test nezávislosti prvků výběru) a provedeme pomocí něj test autokorelace. Tento skript provádí test časové závislosti prvků výběru, případně závislost související s pořadím jednotlivých měření, pomocí koeficientu autokorelace prvního řádu. Nulová hypotéza tvrdí, že není přítomna autokorelace. Test rozhodne o jejím zamítnutí či nezamítnutí na zvolené hladině významnosti. Navíc určí hodnotu testovací statistiky a rozpětí kritického oboru. Aplikujeme-li jej na naši časovou řadu A , výsledkem je:

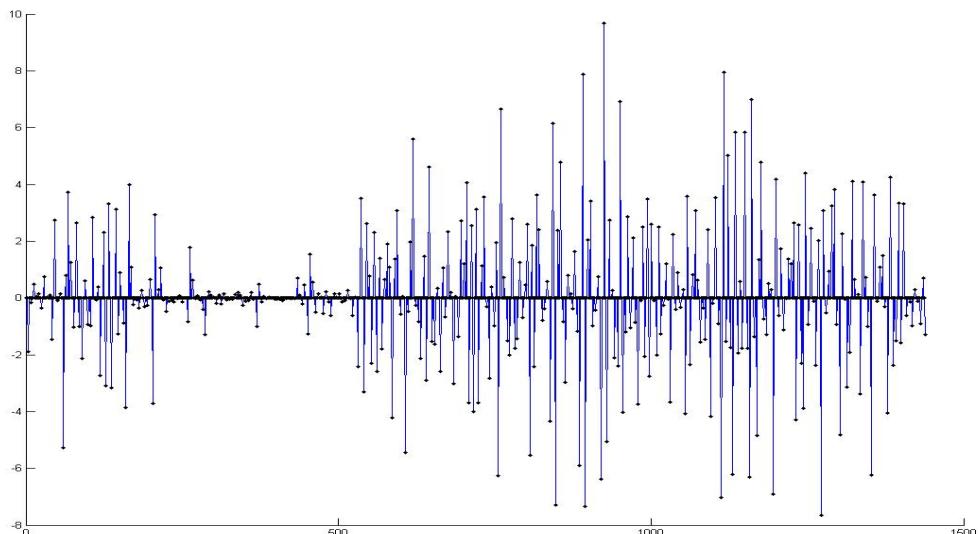
```
>> testnezprvku(A)
*** H0: není autokorelace ***
H0 zamítame - existuje autokorelace
testovaci statistika: 61.8415
hranice kritického oboru:
-1.961611612000120    1.961611612000120
```

Tedy na 5%-ní hladině významnosti nulovou hypotézu zamítáme, v datech existuje autokorelace.

Provedeme diferencování časové řady a pracujeme dále s řadou prvních differencí $A1$

```
>> A1=diff(A);
```

Průběh prvních differencí je na následujícím obrázku, ze kterého jsou jasně patrné shluky volatility.



Obr.5: První difference

Pomocí matlabovské procedury `archtest` provedeme Engelův ARCH-test, který zjišťuje přítomnost podmíněné heteroskedasticity (tedy ARCH efektu). Nulová hypotéza tvrdí, že časová řada residuů sestává z nezávislého identicky rozděleného gaussovského šumu $IID(0, \sigma_\varepsilon^2)$, tedy že neexistuje ARCH efekt. Konstruuje se regresní model

$$\hat{\varepsilon}_t^2 = \omega + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \alpha_2 \hat{\varepsilon}_{t-2}^2 + \dots + \alpha_q \hat{\varepsilon}_{t-q}^2 + u_t,$$

na jehož základě se získá index determinace R^2 . Testové kriterium ve tvaru TR^2 , kde T je počet dat, má za předpokladu platnosti nulové hypotézy asymptoticky rozdělení $\chi^2(q)$.

```
>> [H,p,stat,krit] = archtest(A1-mean(A1), [1 2 3 4 5 10 15 20]', 0.05);
>> [H,p,stat,krit]
ans =
1.0000000000000000    0.0257413771745    4.97330754969816    3.8414588206942
```

1.000000000000	0.0050510169332	10.5763313661064	5.9914645471080
1.000000000000	0.0101482495566	11.3130222684653	7.8147279032512
1.000000000000	0	139.055709112265	9.4877290367812
1.000000000000	0	152.312778733859	11.070497693516
1.000000000000	0	155.153360928627	18.307038053275
1.000000000000	0	159.156255103441	24.995790139728
1.000000000000	0	161.922848731419	31.410432844230

Tedy na 5%-ní hladině významnosti zamítáme nulovou hypotézu, pro $q=1,2,3,4,5,10,15$ i 20 je v modelu přítomna podmíněná heteroskedasticita.

GARCH modelování

Matlabovský test pomocí věrohodnostního poměru lratiotest pomůže odhalit parametry modelu GARCH(p,q).

```
>> spec11 = garchset('P',1,'Q',1,'Display','off');
>> spec21 = garchset('P',2,'Q',1,'Display','off');
>> spec12 = garchset('P',1,'Q',2,'Display','off');
>> spec22 = garchset('P',2,'Q',2,'Display','off');

>> [coeff11,errors11,LLF11] = garchfit(spec11,A1);
>> [coeff12,errors12,LLF12] = garchfit(spec12,A1);
>> [coeff21,errors21,LLF21] = garchfit(spec21,A1);
>> [coeff22,errors22,LLF22] = garchfit(spec22,A1);
```

Postupně provedeme několik testů, z nichž je patrné, že model GARCH(1,2) má nejlepší výsledky.

```
>> [H,pValue,Stat,CriticalValue] = lratiotest(LLF21,LLF11,1,0.05);
>> [H,pValue,Stat,CriticalValue]
ans =
0    0.999722138415920    0.000000121276571    3.841458820694152

>> [H,pValue,Stat,CriticalValue] = lratiotest(LLF12,LLF11,1,0.05);
>> [H,pValue,Stat,CriticalValue]
ans =
1.00000000    0.00000000001612   49.907257904396829    3.841458820694152

>> [H,pValue,Stat,CriticalValue] = lratiotest(LLF12,LLF21,1,0.05);
>> [H,pValue,Stat,CriticalValue]
ans =
1.00000000    0.00000000001612   49.907257783120258    3.841458820694152
```

```
>> [H,pValue,Stat,CriticalValue] = lratiotest(LLF12,LLF22,1,0.05);
>> [H,pValue,Stat,CriticalValue]
ans =
0    1.0000000000000000    -1.877927843629550    3.841458820694152
```

```

>> [H,pValue,Stat,CriticalValue] = lratiotest(LLF22,LLF12,1,0.05);
>> [H,pValue,Stat,CriticalValue]
ans =
0    0.170569848830138    1.877927843629550    3.841458820694152

```

Vytvoříme matlabovský skript GARCHmodel.m , který sestrojí matici hodnot AIC, resp. BIC kriteria pro rozhodování o rádu modelu GARCH – prvek s minimální hodnotou na pozici $[i, j]$ indikuje model GARCH(i, j). V našem případě dostaneme:

```

>> GARCHmodel(A1,2)
hodnoty AIC kriteria pro jednotlivé GARCH modely:
1.0e+003 *
4.220753025055129    4.172845767150732
4.222753024933852    4.172967839307103

hodnoty BIC kriteria pro jednotlivé GARCH modely:
1.0e+003 *
4.241839839882679    4.199204285685170
4.249111543468290    4.204598061548427

```

I odtud se jeví model GARCH(1,2) jako nejlepší volba. Připravíme si matlabovskou strukturu tohoto modelu.

```

>> spec = garchset('P', 1, 'Q', 2)
spec =
    Comment: 'Mean: ARMAX(0,0,?); Variance: GARCH(1,2) '
    Distribution: 'Gaussian'
    C: []
    VarianceModel: 'GARCH'
    P: 1
    Q: 2
    K: []
    GARCH: []
    ARCH: []

```

Provedeme odhad parametrů užitím matlabovské procedury garchfit.

```
>> [coeff,errors,LLF,eFit,sFit] = garchfit(spec,A1);
```

```
%%%%%%%%%%%%%
```

```
Diagnostic Information
```

```
Number of variables: 5
```

```
Functions
```

```
Objective: garchllfn
```

```
Gradient: finite-differencing
```

Hessian: finite-differencing (or Quasi-Newton)
 Nonlinear constraints: armanlc
 Gradient of nonlinear constraints: finite-differencing

Constraints

Number of nonlinear inequality constraints:	0
Number of nonlinear equality constraints:	0
Number of linear inequality constraints:	1
Number of linear equality constraints:	0
Number of lower bound constraints:	5
Number of upper bound constraints:	5

Algorithm selected

medium-scale

%%%%%%%%%%%%%

End diagnostic information

Iter	F-count	f(x)	Line search Directional First-order					Procedure
			constraint	steplength	derivative	optimality		
0	6	2334.52	-0.025					
1	19	2332.87	-0.0248	0.00781	266	448		
2	29	2332.13	-0.02325	0.0625	481	666		
3	41	2331.84	-0.02289	0.0156	143	420		
4	48	2287.97	-0.01145	0.5	-80.9	479		
5	54	2169.63	0	1	2.78e+003	3.04e+004		
6	64	2157.44	0	0.0625	257	7.15e+003		
7	72	2150.12	0	0.25	402	1.21e+004		
8	81	2132.51	0	0.125	54.6	4.3e+003		
9	87	2101.07	0	1	38.7	9.59e+003		
10	93	2091.36	0	1	1.39	5.14e+003		
11	99	2085.66	0	1	-4.74	3.23e+003		
12	106	2082.62	0	0.5	11.6	8.13e+003		
13	112	2082.12	0	1	2.8	3.76e+003		
14	118	2081.51	0	1	-0.0778	1.12e+003		
15	124	2081.44	0	1	0.107	641		
16	130	2081.42	0	1	0.00345	64.1		
17	136	2081.42	0	1	-1.34e-005	2.75		
18	142	2081.42	0	1	2.76e-009	0.0107	Hessian modified	

Optimization terminated: magnitude of search direction less than $2 * \text{options.TolX}$

and maximum constraint violation is less than options.TolCon .

```
Active inequalities (to within options.TolCon = 1e-007):
```

lower	upper	ineqlin	ineqnonlin
4			1

```
Boundary Constraints Active: Standard Errors May Be Inaccurate.
```

```
>> garchdisp(coeff,errors)
```

```
Mean: ARMAX(0,0,0); Variance: GARCH(1,2)
```

```
Conditional Probability Distribution: Gaussian
```

```
Number of Model Parameters Estimated: 5
```

Parameter	Value	Standard Error	T Statistic
C	-0.011966	0.01614	-0.7413
K	0.0015184	8.0975e-005	18.7519
GARCH(1)	0.96139	0.0012091	795.1219
ARCH(1)	0	0.010684	0.0000
ARCH(2)	0.03861	0.01047	3.6876

Dostáváme tak GARCH(1,2) model:

$$Y_t = -0.011966 + \varepsilon_t,$$

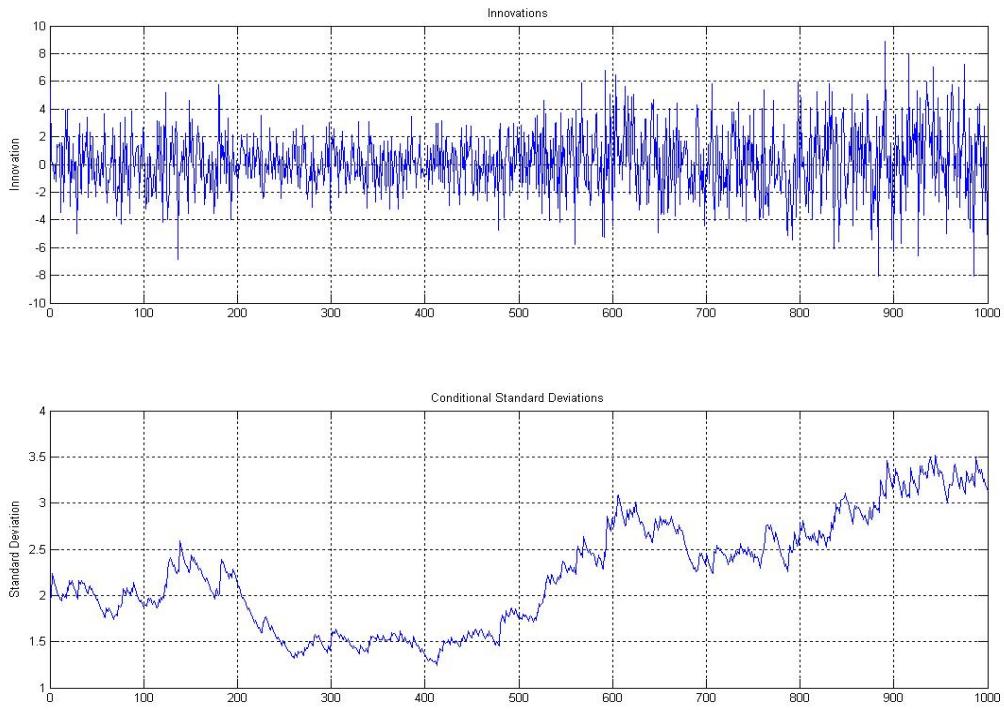
$$\sigma_t^2 = 0.0015184 + 0.96139\sigma_{t-1}^2 + 0.03861\varepsilon_{t-2}^2.$$

Provedeme simulaci dat daných výše uvedeným modelem:

```
>> [e,s] = garchsim(coeff,1000);
```

```
>> garchplot(e,s)
```

Průběhy residuí a podmíněné směrodatné odchylky simulovaného procesu jsou na následujícím obrázku.



Obr.6: Residua a podmíněné směrodatné odchylky

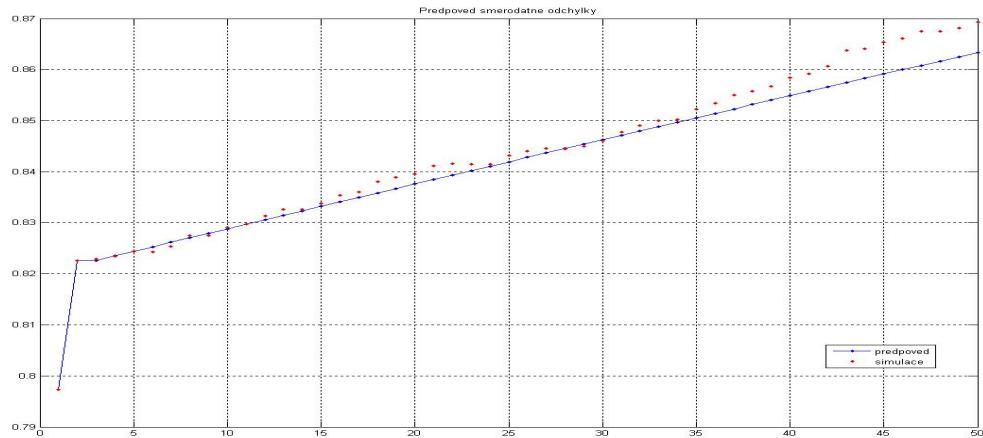
Na závěr vytvoříme předpovědi (v horizontu 50 pozorování), které porovnáme s výsledky získané simulací našeho modelu.

```

>> hor=50;
>> [sigmaForecast,meanForecast,sigmaTotal,meanRMSE]=garchpred(coeff,A1,hor);
>> nPaths=1000;
>> [eSim,sSim,ySim] = garchsim(coeff,hor,nPaths,0,[],[],eFit,sFit,A1(end));

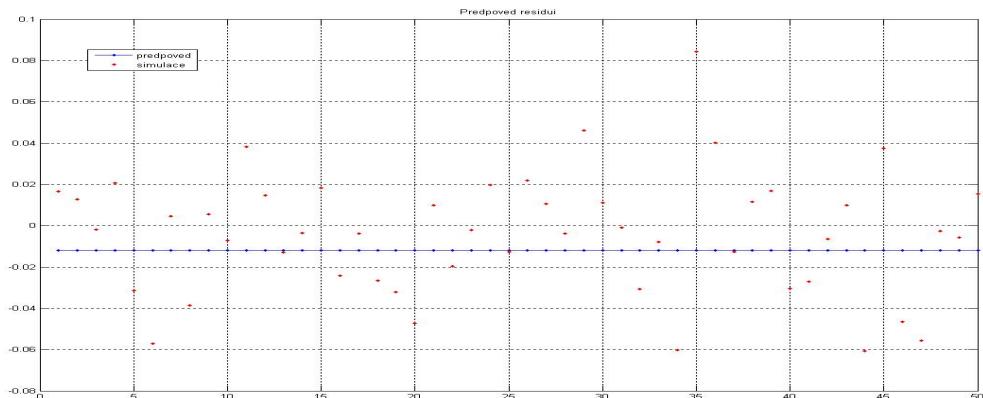
>> figure
>> plot(sigmaForecast,'.-b')
>> hold('on')
>> grid('on')
>> plot(sqrt(mean(sSim.^2,2)),'.r')
>> title('Predpoved smerodatne odchylky')
>> legend('predpoved','simulace')

```



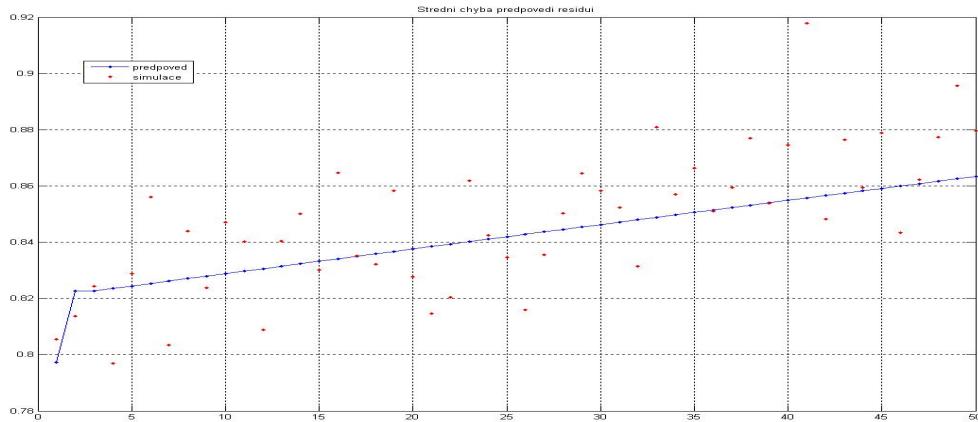
Obr.7: Předpověď směrodatné odchylky

```
>> figure(2)
>> plot(meanForecast,'.-b')
>> hold('on')
>> grid('on')
>> plot(mean(ySim,2),'.r')
>> title('Predpoved residui')
>> legend('predpoved','simulace',4)
```



Obr.8: Předpověď residuů

```
>> figure(3)
>> plot(meanRMSE,'.-b')
>> hold('on')
>> grid('on')
>> plot(std(ySim)', '.r')
>> title('Stredni chyba predpovedi residui')
>> legend('predpoved','simulace')
```



Obr.9: Střední chyba předpovědi residuů

6 Závěr

Naše výsledky nejsou produktem ukončeného a uzavřeného projektu, jsou pouze jednou z částí běžícího úkolu modelování a predikce spotřeby tepelné energie. Modely GARCH jsou běžně používané a často aplikované v ekonometrii. Našim záměrem bylo demonstrovat užití těchto stochastických modelů v technometrii, přičemž samotná analýza probíhala v prostředí MATLABu.

Literatura

- [1] Anděl, J.: *Matematická statistika*, SNTL/ALFA, Praha, 1978
- [2] Anděl, J.: *Statistická analýza časových řad*, SNTL, Praha, 1976
- [3] Arlt, J. – Arltová, M.: *Finanční časové řady*, Grada, Praha, 2003
- [4] Arlt, J. – Arltová, M.: *Konstrukce předpovědí na základě modelů GARCH*, Acta oeconomica pragensia 10: (7), VŠE, Praha, 2002
- [5] Arlt, J.: *Moderní metody modelování ekonomických časových řad*, Grada, Praha, 1999
- [6] Arlt, J. – Radkovský, Š.: *Význam modelování a předpovídání volatility časových řad pro řízení ekonomických procesů*, Politická ekonomie 48: (1), VŠE, Praha, 2000
- [7] Budíková, M. – Lerch, T. – Mikoláš, Š.: *Základní statistické metody*, skriptum MU, Brno, 2005
- [8] Cipra, T.: *Analýza časových řad s aplikacemi v ekonomii*, SNTL, Praha, 1986
- [9] Forbelská, M.: *Stochastické modelování jednorozměrných časových řad*, skriptum MU, Brno, 2009
- [10] Hebák, P. a kol.: *Vícerozměrné statistické metody I, 2, 3*, Informatorium, Praha, 2005, 2007
- [11] Likeš, J. – Machek, J.: *Matematická statistika*, MVŠT – sešit IV, SNTL, Praha, 1983
- [12] Likeš, J. – Machek, J.: *Počet pravděpodobnosti*, MVŠT – sešit X, SNTL, Praha, 1981
- [13] Meloun, M. – Militký, J.: *Statistická analýza experimentálních dat*, Academia, Praha, 2004
- [14] Rao, R. C.: *Lineární metody statistické indukce a jejich aplikace*, Academia, Praha, 1978

David KVAPIL

UNIS, a.s.

Jundrovská 33, 624 00 Brno, CZ

dkvapil@unis.cz

Článek byl vypracován na základě výsledků dosazených z grantového projektu 2C06007 MSMT