UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

DIPLOMOVÁ PRÁCA

2011 ZDENKA ZUBÁČOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
Fakulta matematiky, fyziky a informatiky

Spracovanie predikčných modelov pre prognózy spotreby zemného plynu v predajnom portfóliu SPP na slovenskom území

Diplomová práca

Študijný odbor: 9.1.9. Aplikovaná matematika
Študijný smer: ekonomická a finančná matematika

Vedúci diplomovej práce:
Mgr. Matej Kruškpán

Diplomantka :
Zdenka Zubáčová

Bratislava 2011

688a5775-850c-470f-ad58-a3b4b02976c6

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Zdenka Zubáčová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov : Spracovanie predikčných modelov pre prognózy spotreby zemného plynu v predajnom portfóliu SPP na slovenskom území

Cieľ : Cieľom diplomovej práce je pripraviť pre SPP nové predikčné algoritmy pre prognózy spotreby zemného plynu a tým priamo znížiť náklady SPP na odchýlku. Diplomová práca má ďalej za cieľ navrhnúť nové druhy premenných o charaktere počasia, ktoré by mohli pomôcť spresniť prognózy SPP.

Vedúci : Mgr. Matej Krušpán

Dátum zadania: 09.02.2010

Dátum schválenia: 07.04.2011

.....
prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Dátum potvrdenia finálnej verzie práce, súhlas s jej odovzdaním (vrátane spôsobu sprístupnenia)

.....
vedúci práce

Čestné prehlásenie

Vyhlasujem, že som diplomovú prácu vypracovala samostatne s použitím uvedenej odbornej literatúry.

Bratislava 14.04.2011 .....

Vlastnoručný podpis

Poďakovanie

Ďakujem svojmu vedúcemu Mgr. Matejovi Kruškánovi za ochotu, trpezlivosť a pomoc pri tejto práci. Taktiež moje poďakovanie patrí RNDr. Mgr. Beáte Stehlíkovej, PhD. za cenné rady pri tvorbe modelov.

Názov práce:	Spracovanie predikčných modelov pre prognózy spotreby zemného plynu v predajnom portfóliu SPP na slovenskom území
Pracovisko:	EFM, FMFI UK v Bratislave
Autor:	Zdenka Zubáčová
Vedúci DP:	Mgr. Matej Krušpán
Dátum:	14.04.2011
Kľúčové slová:	Časový rad. Stacionarita. Spotreba plynu. ARMA model. Predikcie.

Anotácia: Táto diplomová práca sa venuje spracovaniu predikčných modelov pre prognózy spotreby zemného plynu v predajnom portfóliu Slovenského plynárenského priemyslu (SPP) na slovenskom území. Na začiatku sa budeme venovať teoretickej časti. Spomenieme si základné pojmy používané pri ekonometrickom modelovaní, definujeme si klasický model lineárnej regresie a v tretej kapitole si špecifikujeme autokorelačné metódy pre časové rady. V praktickej časti si najskôr ukážeme rady, s ktorými budeme ďalej pracovať a pokúsime sa určiť čo najpresnejšie modely, ktorých správnosť overíme pomocou krátkodobých predikcií.

Title:	Processing of predictive models for forecasting natural gas consumption in the sales portfolio of the SPP on the Slovak territory.
Department:	EFM, FMFI UK in Bratislava
Author:	Zdenka Zubáčová
Leader :	Mgr. Matej Krušpán
Date:	14.04.2011
Key words:	Time series. Stationarity. Gas consumption. ARMA model. Prediction.

Abstract: This thesis deals with the processing of predictive models to forecast gas consumption in the sales portfolio of the Slovak Gas Industry (SPP) on the Slovak territory. At the beginning we will talk about theory. Mention the basic terms used in econometric modelling, we define the classical linear regression model. We will specify the autocorrelation method for time series in the third chapter.
In practical part, we will define basic time series, with which we will work and try to determinate models, which we will verify with using short-term predictions.

Obsah

Úvod

V prvej kapitole sa pokúsime čitateľovi priblížiť teoretické poznatky ohľadom časových radov. Väčšinu všeobecných definícií a vzorcov budeme čerpať predovšetkým z publikácie od Tomáša Cipru [1]. Vysvetlíme si, čo si budeme predstavovať pod pojmom *časový rad*, aké procesy sú považované za *stochastické* a čo musí byť splnené aby sme mohli tvrdiť, že rad je *stacionárny*. Každý časový rad sa dá rozložiť na niekoľko častí, a ktoré časti to sú si spomenieme na záver kapitoly.

Klasický model lineárnej regresie bude predmetom skúmania v druhej kapitole. Zameriame sa na to, ako správne identifikovať model. Pri určovaní správnosti modelu je potrebné overiť niektoré základné charakteristiky, akými je heteroskedasticita, autokorelácia a iné.

Špecifickým typom ekonometrických modelov, ktoré sa využívajú predovšetkým v ekonómii na tvorbu krátkodobých predikcií, venujeme poslednú teoretickú kapitolu. Vysvetlíme, čo sú to $ARMA$ modely a ako postupovať pri ich identifikácii a následnom overení správnosti pomocou predikcií.

Štvrtá a zároveň posledná kapitola je určená praktickej časti. Z reálnych dát, ktoré zachytávajú skutočnú spotrebu zemného plynu na území Slovenskej republiky, vytvoríme čo najpresnejšie modely, pomocou ktorých budeme vedieť predpovedať vývoj daného časového radu do blízkej budúcnosti. Popri tomto časovom rade budeme skúmať a iný, ktorý zobrazuje vývoj teploty na tom istom území. Keďže sa dá predpokladať, že to, koľko plynu sa spotrebúva, či už v domácnostiach alebo pri veľkých odberateľoch, značne závisí od teploty ovzdušia. Či sú naše úvahy správne si zhrnieme v závere tejto práce.

Kapitola 1

Základné pojmy

V prvej kapitole si definujeme základné pojmy, s ktorými budeme neskôr pracovať. Začneme definíciou časových radov vo všeobecnosti a ukážeme si, čo si predstavujeme pod pojmom "biely šum". Spomenieme si základnú vlastnosť týchto radov, ktorou je stacionarita a následne rozložíme rad na niekoľko častí, z ktorých si bližšie špecifikujeme sezónnu zložku. Základné definície a vzorce boli čerpané predovšetkým z kníh [1, 2].

1.1 Časový rad

Pod pojmom *časový rad* rozumieme postupnosť hodnôt, ktoré sú usporiadané v čase. Predpokladáme, že časový interval medzi jednotlivými pozorovaniami je konštantný. Ak budeme teda hovoriť o časovom rade, budeme mať na mysli *diskrétny časový rad s ekvidistatným krokom.*

Diskrétny časový rad si označíme ako postupnosť dát $(x_1, x_2, \cdots, x_n)$, kde ľubovoľná hodnota x_t je reálne číslo. Ďalej budeme predpokladať, že náš časový rad je realizáciou stochastického (náhodného) procesu $(X_1, X_2, \cdots, X_n)$. Ak náhodnú premennú so svojim rozdelením označíme ako X_t, tak x_t je jej hodnota a n označuje celkovú dĺžku časového radu.

Podľa počtu sledovaných znakov rozlišujeme časové rady na *jednorozmerné* (sledovaný je len jeden znak) a *viacrozmerné* (viac sledovaných znakov).

Jednou z najdôležitejších požiadaviek pre časové rady je aby hodnoty boli homogénne, tj. porovnateľné z hľadiska priestorového, časového a vecného.

1.2 Stochastické procesy a biely šum

Ako sme už spomenuli, pod pojmom *stochastický proces* je myslený náhodný proces $\{X(\omega, t), \omega \in \Omega, t \in T\}$, kde Ω je výberový priestor a T je indexová množina. Podľa časovej indexovej množiny T rozlišujeme 2 typy náhodných procesov a to

- náhodné procesy s diskrétnym časom (T tvoria diskrétne reálne hodnoty)
- náhodné procesy so spojitým časom (T tvorí interval)

Pre každý stochastický proces platia nasledovné charakteristiky:

- $\mu_t = E(X_t)$ (stredná hodnota)
- $\sigma_t^2 = D(X_t) = E(X_t - \mu_t)^2$ (rozptyl)
- $cov(r,s) = E[(X_r - \mu_r)(X_s - \mu_s)]$ (kovariančná funkcia)
- $\rho(r,s) = \frac{cov(X_r,X_s)}{\sqrt{D(X_r)}\sqrt{D(X_s)}}$ (korelačná funkcia)

1.3 Stacionarita

Jednou z vlastností časových radov je aj *stacionarita*. Časový rad je stacionárny, ak jeho rozdelenie pravdepodobností je nemenné, tj. invariantné v čase. V prípade nestacionárnych radov sa vhodnou transformáciou dá dosiahnuť aby bol rad stiacionárny.

Rozlišujeme 2 základné typy stacionarity a to *slabú* a *silnú stacionaritu*. Silná stacionarita hovorí o tom, že pravdepodobnostné správanie stochastického procesu je v čase invariantné. Ak rad vykazuje nemennosť strednej hodnoty, rozptylu a kovariancie, tj. štatistických mier skúmanej náhodnej premennej v čase, tak hovoríme o *slabej stacionarite*. Musí teda platiť:

$$E(X_t) = \mu \qquad \text{pre } t \in T \tag{1.1}$$

$$V(X_t) = \sigma^2 \qquad \text{pre } t \in T \tag{1.2}$$

$$cov(X_t, X_s) = cov(X_{t+k}, X_{s+k}) \qquad \text{pre } t,s \in T, t \neq s, k \in Z \tag{1.3}$$

Keď budeme hovoriť o stacionarite, budeme mať na mysli práve slabú stacionaritu.

O *bielom šume* hovoríme ak platí:

$$\begin{aligned} E(X_t) &= \mu && \text{pre } t \in T, \text{zvyčajne } \mu = 0 \\ V(X_t) &= \sigma^2 && \text{pre } t \in T \\ cov(X_t, X_s) &= 0 && \text{pre } s \neq t \end{aligned}$$

Jedným z testov na stabilitu je *Ramsey RESET test*[1] . Tento test sa používa na zistenie špecifikačných chýb, ktoré vznikli v dôsledku vynechania premenných alebo chybnou špecifikáciou analytického tvaru modelu. Nulovou hypotézou je, že model je špecifikovaný správne, avšak pri vyvrátení tejto hypotézy alternatíva neexistuje, čo znamená, že ak nie je model špecifikovaný správne, tento test nám nehovorí nič o tom, ako ho vylepšiť.

Testovacie kritérium na overenie nulovej hypotézy má tvar:

$$F = \frac{\frac{R_*^2 - R^2}{v_1}}{\frac{1-R_*^2}{v_2}}$$

[1] Z anglického **Re**gression **S**pecification **E**rror **T**est

kde R^2 je Koeficient determinácie, v_1, v_2 sú stupne voľnosti pre F rozdelenie a R^2_* je Koeficient determinácie pre upravený daný model, a to tak, že do modelu pridáme nové vysvetľujúce premenné, ktoré budú mocninami vyrovnaných hodnôt vysvetľovanej premennej (v_1 je počet nezávislých pridaných premenných, v_2 je počet parametrov v upravenom modeli).

1.4 Dekompozícia časového radu

Dekompozičné metódy kladú dôraz na prácu so systematickými zložkami časových radov (trendovou, sezónnou a cyklickou) a jednotlivé pozorovania berú ako navzájom nekorelované. Typickým nástrojom je tu regresná analýza.

Časové rady môžeme rozdeliť na niekoľko častí:

1. **Trend** - dlhodobé zmeny v rade. Táto zložka vzniká v dôsledku dlhodobého pôsobenia zmien. Ak je trend lineárny, tak prvé diferencie rad stacionarizujú.
2. **Cyklická zložka** - zachytáva periodické zmeny s premennou periódou.
3. **Sezónna zložka** - periodické zmeny v rade v nejakom časovom itervale.
4. **Reziduálna zložka** - ostáva v rade po odstránení trendu, cyklickej a sezónnej zložky (náhodné zmeny, ktoré sa väčšinou označujú ako biely šum[2]).

Rozloženie radu na tieto zložky nám môže pomôcť k identifikovaniu pravidelného spávania sa radu. Dekompozícia radu môže byť *aditívna* (časový rad = trend + cyklická zložka + sezónna zložka + biely šum), kde sú všetky zložky v absolútnych číslach, alebo *multiplikatívna* (časový rad = trend * cyklická zložka * sezónna zložka * biely šum), kde trend je absolutnou hodnotou a ostatné zložky sú voči nemu relatívne. Každý časový rad väčšinou obsahuje biely šum ale nemusí obsahovať ostatné zložky.

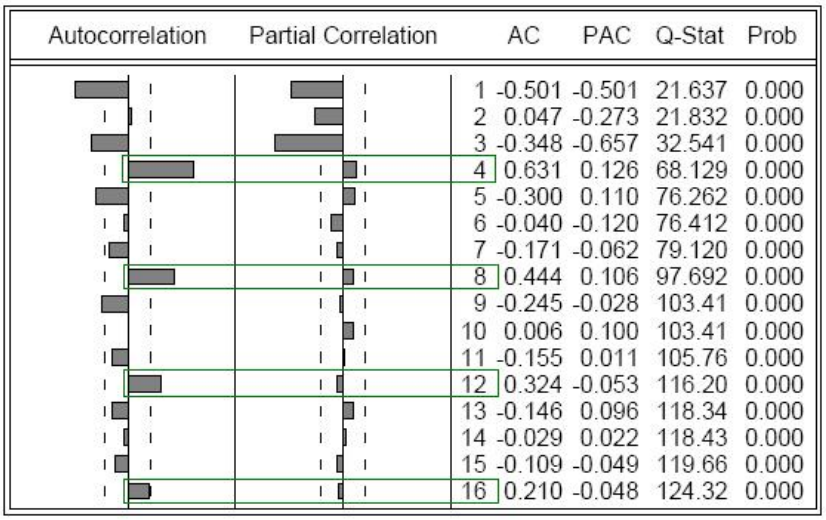
Sezónnosť

V tejto časti sa budeme zaoberať sezónnou zložkou modelu. Jedná sa o zložku, ktorá sa v časovom rade pravidelne opakuje. Väčšinou sa zmeny dejú počas jedného kalendárneho roka (napríklad štvrťročne) a každý rok sa pravidelne opakujú. Pri tvorbe modelov je potrebné časový rad najskôr sezónne očistiť. Či sa takáto zložka v rade nachádza vieme odhaliť pomocou priebehu korelogramu[3].

Z priebehu vidíme sezónny charakter radu. Výrazné korelácie sú pre lagy 4,8,12 a 16, a keďže dáta boli štvrťročné, znamená to ročnú sezónnosť. Pri analýze by sa mali separovať *sezónne faktory* označované ako $I_1, I_2, \cdots, I_s$ kde s znamená dĺžku sezóny. Rozdiel medzi aditívnou a multiplikatívnou dekompozíciou z hľadiska sezónnosti je v tom, že pri multiplikatívnej sa hodnoty sezónnych výkyvov zväčšujú pre rastúci trend, resp. zmenšujú pre klesajúci. Sezónna a trendová zložka nie sú navzájom určené jednoznačne.

[2] Biely šum je stacionárny a centrovaný náhodný proces

[3] Obr. 1.1 je zo stránky http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09/cv4.html



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.501	-0.501	21.637	0.000
		2	0.047	-0.273	21.832	0.000
		3	-0.348	-0.657	32.541	0.000
		4	0.631	0.126	68.129	0.000
		5	-0.300	0.110	76.262	0.000
		6	-0.040	-0.120	76.412	0.000
		7	-0.171	-0.062	79.120	0.000
		8	0.444	0.106	97.692	0.000
		9	-0.245	-0.028	103.41	0.000
		10	0.006	0.100	103.41	0.000
		11	-0.155	0.011	105.76	0.000
		12	0.324	-0.053	116.20	0.000
		13	-0.146	0.096	118.34	0.000
		14	-0.029	0.022	118.43	0.000
		15	-0.109	-0.049	119.66	0.000
		16	0.210	-0.048	124.32	0.000

Obr. 1.1: Sezónny priebeh dát

Riešenie sezónnosti

Ak sa v časovom rade vyskytuje sezónna zložka, je potrebné ju eliminovať a to buď dekompozíciou radu, alebo pomocou Holtovej-Wintersonovej metódy. Dekompozícia môže byť buď aditívna, alebo multiplikatívna. Pri aditívnej sa postupuje pomocou zostrojenia kĺzavých priemerov a centrovania sezónnych faktorov. Ukážeme si príklad na mesačných dátach.

$$I_j = I_j^* - \overline{I}^* = I_j^* - \frac{I_1^* + \cdots + I_{12}^*}{12}$$

kde $I_1^*, \cdots, I_{12}^*$ sú naše odhady, $j = 1, 2, \cdots, 12$ pretože predpokladáme mesačné dáta a $\overline{I}^*$ je aritmetický priemer. Potom výsledný rad upravíme na tvar

$$\widehat{y}_t^{(12)} = y_t - I_j$$

kde $\widehat{y}_t^{(12)}$ je centrovaný kĺzavý priemer a t zodpovedá j-temu mesiacu v roku.

Pri multiplikatívnej dekompozícii sa postupuje rovnako, len s tým rozdielom, že sezónne faktory vyjadríme ako

$$I_j = \frac{I_j^*}{\overline{I}^*} = \frac{I_j^*}{\sqrt[12]{I_1^* . \cdots . I_{12}^*}}$$

a výsledný rad bude vyzerať nasledovne:

$$\widehat{y}_t^{(12)} = \frac{y_t}{I_j}$$

Kapitola 2

Klasický model lineárnej regresie

Pri tvorbe ekonometrického modelu sa vo väčšine prípadov postupuje iteratívne, tj. metódou "pokus, omyl" postupne zistíme, ktorý z modelov je najvhodnejší. Ak sú nezávisle od seba študované dva totožné rady, tak výsledné modely môžu byť odlišné avšak s veľmi podobnou resp. rovnakou interpretáciou záverov.

V tejto kapitole sa zameriame na klasický všeobecný model lineárnej regresie. Ukážeme si ako správne odhadnúť parametre modelu. Keď máme odhadnutý model, je potrebné otestovať niektoré vlastnosti ako je napríklad heteroskedasticita, autokorelácia a iné, o ktorých si povieme viac na najbližších stranách. Vzorce a definície budú predovšetkým zo zdrojov [1, 2].

2.1 Všeobecný model

Jedným z najdôležitejších ekonometrických nástrojov je *regresná analýza*, ktorá slúži na popis vzťahu medzi veličinami. Tieto veličiny sa nazývajú *premenné*. Úlohou tejto analýzy je určiť a vysvetliť zmeny hodnôt jednej premennej pomocou zmien iných premenných. *Vysvetľovanú premennú* (závislá premenná) označujeme ako y a *vysvetľujúce premenné* (nezávislá premenná) ako $x_{1,2..,k}$.

Ak sa rozhodneme skúmať vzťah len medzi jednou vysvetľovanou a jednou vysvetľujúcou premennou, tak sa miera lineárnej závislosti skúma pomocou *korelačného koeficientu* ($\rho = corr(x, y)$). Tento koeficient nadobúda hodnoty od -1 po 1, pričom ak je blízko hodnoty 1, znamená to, že sa medzi premennými nachádza *kladná korelácia*. Čím je vyššia hodnota jednej premennej, tým je vyššia hodnota aj druhej premennej. Ak je blízko hodnoty -1, tak je tam *záporná korelácia*. Čím je vyššia hodnota jednej premennej, tým je nižšia hodnota druhej premennej. Ak je hodnota blízka nule, tak premenné vzájomne nekorelujú. Vo všeobecnosti platí, že $E(var(x|y)) \leq var(y)$.

Lineárny regresný model sa dá formálne zapísať ako:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \epsilon_t$$

kde čas $t = 1, 2, \cdots, T$, y_t je hodnota vysvetľovanej premennej v čase t, $x_1, \cdots, x_k$ sú hodnoty vysvetľujúcich premenných pozorovaných v čase t, β_0 je absolútny člen regresie, $\beta_1, \cdots, \beta_k$ sú neznáme parametre modelu, ϵ_t je reziduálna (náhodná) zložka modelu.

Poznámka: Rozdiel medzi počtom pozorovaní a počtom parametrov modelu sa nazýva *počet stupňov voľnosti modelu* a musí platiť $n > k + 1$, kde $k + 1$ je počet parametrov $\beta_0, \beta_1, \cdots, \beta_k$ a n je počet pozorovaní.

Ak chceme predchádzajúcu rovnicu zapísať v maticovom tvare, dostaneme nasledovné:

$$y = X\beta + \epsilon$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{T2} & \cdots & x_{Tk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}$$

Kde X je matica $(T \times k)$, β je vektor $(k \times 1)$ a ϵ je vektor $(T \times 1)$. β_1 sa nazýva aj *absolútny člen* (intercept). Ak sa hodnota premennej x_i zvýši o jednotku, tak sa zmení hodnota premennej y o hodnotu β_i (za predpokladu, ak sa nič iné nezmenilo[1]). ϵ_t je *reziduálna* (náhodná) *zložka modelu.* Táto zložka v sebe zahŕňa:

- Súhrn vplyvov, ktoré nie sú v modeli zahrnuté
- Chyby v meraní
- Nekorektnú voľbu regresného vzťahu

2.2 Odhad parametrov

Odhad parametrov lineárneho modelu spadá do problematiky ekonometrických metód a je podmienený splnením predpokladov o náhodnej zložke. Týmito metódami sa budeme zaoberať neskôr. Medzi najčastejšie spôsoby odhadu parametrov patrí *metóda najmenších štvorcov*[2].

Metóda najmenších štvorcov

Základom tejto metódy je minimalizácia súčtu štvorcov odchýlok, určených ako rozdiel medzi pozorovanými hodnotami vysvetľovanej premennej a jej vypočítanými hodnotami. Je založená na dvoch princípoch:

[1] Takzvaný princíp *ceteris paribus*

[2] Označovaná aj OLS z anglického Ordinary Least Squares

1. $\sum_{t=1}^{n} e_t^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \rightarrow min$ (minimalizácia súčtu štvorcov odchýlok)
2. $\sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0$ (nulové odchýlky, tj. reziduá sú rovné nule)

kde e je rezíduum, y je empirická hodnota závislej premennej a $\widehat{y}$ predstavuje teoretickú hodnotu tej istej premennej, za ktorú dosadzujeme zvolený funkčný predpis. Vytvorením parciálnych derivácií podľa jednotlivých premenných

$$\frac{\sigma F(b_0, b_1, \cdots, b_k)}{\sigma b_i}$$

získame sústavu $k+1$ rovníc s $k+1$ neznámymi $b_0, b_1, \cdots, b_k$. Ako príklad si zoberieme jednoduchú lineárnu závislosť s funkciou danou ako

$$Y' = B_0 + B_1 X$$

Úpravou dostaneme tvar:

$$\sum y = nb_0 + b_1 \sum x$$

$$\sum xy = b_0 \sum x + b_1 \sum x^2$$

a následne použijeme parciálne derivácie

$$b_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \qquad b_1 = \frac{n \sum xy \sum y - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

Koeficient b_0 je konštanta, ktorá určuje, kde priamka pretína os y a b_1 hovorí, o koľko sa zmení závislá premenná, ak sa nezávislá zmení o jedna.

2.3 Verifikácia modelu

Verifikáciu modelu si rozdelíme na 2 časti a to

1. štatistická
   - koeficient determinácie (testovanie ako celku)
2. ekonometrická
   - autokorelácia
   - heteroskedasticita
   - multikolinearita

2.3.1 Koeficient determinácie

Koeficient determinácie (R Square) ozn. aj ako R^2 vyjadruje, akú časť celkovej variability závislej premennej vysvetľuje model. Je to miera kvality vyrovnania empirických hodnôt závislej premennej modelovanými hodnotami. Hodnota tohto koeficientu sa vyjadruje v percentách a teda je v rozmedzí od 0

po 1. Čím viac sa približuje k číslu 1,tým lepší je daný model. Vypočítame ho podľa vzťahu:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

kde RSS je *reziduálny súčet štvorcov* (residual sum of squares), ktorý sa vypočíta podľa vzťahu:

$$RSS = \sum_{t=1}^{T} \widehat{\epsilon}_t^2 = \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2$$

O správnosti modelu hovorí jeho nezáporná hodnota. Čím je menšia, tým je model lepší.

$$TSS = \sum_{t=1}^{T} (y_t - \overline{y})^2$$

je *úplný súčet štvorcov* (total sum of squares).

$$ESS = \sum_{t=1}^{T} (\widehat{y}_t - \overline{y})^2$$

je *vysvetlený súčet štvorcov* (explained sum of squares).
Podľa Pytagorovej vety platí:

$$TSS = ESS + RSS$$

V praxi sa často používa *korigovaný koeficient determinácie* (adjusted coefficient of determination):

$$\overline{R}^2 = 1 - [\frac{T-1}{T-k}(1 - R^2)]$$

kde k je počet vysvetľujúcich premenných.

2.3.2 Heteroskedasticita

Medzi podmienky klasického lineárneho regresného modelu patrí aj požiadavka konečného a konštantného rozptylu náhodných zložiek, tým pádom aj rezíduí modelu, ktorý označujeme ako *homoskedasticitu*. Jej opakom je heteroskedasticita. Heteroskedasticita znamená, že je porušený klasický predpoklad $D[\epsilon_i] = \sigma^2$, to znamená, že disperzia náhodných chýb nie je konštantná a konečná u všetkých pozorovaní. Jedným z najpoužívanejších testov na heteroskedasticitu je *Whiteov test* (White heteroskedasticity test). Princíp spočíva v tom, že ak máme napríklad model:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \epsilon_t, \quad t = 1, \cdots, T$$

tak sa vytvorí pomocný model:

$$\widehat{\epsilon_t} = \alpha_1 + \alpha_2 x_{t2} + \alpha_3 x_{t3} + \alpha_4 x_{t2}^2 + \alpha_5 x_{t3}^2 + \alpha_6 x_{t2} x_{t3} + u_t$$

Našim cieľom je zistiť, že či sa rozptyl pôvodných chýb (ľavá strana pomocného modelu) systematicky mení v závislosti na všetkých vysvetľujúcich premenných pôvodného modelu.

Riešenie heteroskedasticity

Rozlišujeme 2 typy heteroskedasticity a to *aditívnu* a *multiplikatívnu*. (multiplikatívna dáva kladný rozptyl).

Heteroskedasticitu môže spôsobiť napríklad chybná špecifikácia modelu (vynechanie niektorej dôležitej vysvetľujúcej premennej) alebo ak použijeme k odhadu parametrov modelu namiesto pôvodných pozorovaní ich skupinové priemery. Pri väčšine prípadov sú príčiny heteroskedasticity neznáme. Jej odstránenie sa rieši logaritmickou (čím sa stláča stupnica, v ktorej sú premenné merané) alebo inou transformáciou.

2.3.3 Autokorelovanosť reziduí

Aby boli reziduá *nekorelované*, musí platiť :

$$cov(\epsilon_s, \epsilon_t) = 0 \quad pre \quad s \neq t$$

Autokorelácia je sériová závislosť náhodných porúch, poprípade reziduí. Je to typický jav časových radov. K autokorelácii dochádza keď je reziduálna zložka ϵ_t korelovaná so svojimi opozdenými a budúcimi hodnotami $\epsilon_{t+k}, (k \neq 0)$. Predpona „auto“ znamená, že korelácia je v rámci jedného radu.

Dôvody autokorelácie

- Chýbajú niektoré vysvetľujúce premenné (regresory)
- Nesprávna špecifikácia matematickej formy modelu
- Údaje vykazujú zotrvačnosť vo vývoji
- Funkcionálny regresný vzťah je nelineárny

Najjednoduchší typ autokorelácie sa modeluje pomocou reziduálnych zložiek ϵ_t.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

kde parameter ρ sa označuje ako *biely šum* a nadobúda hodnoty od -1 po 1. Hodnoty blízke 1 znamenajú *pozitívnu koreláciu*, hodnoty blízke -1 *zápornú koreláciu* a hodnoty okolo 0 *nekorelovanosť*. Tento model sa nazýva aj *autoregresný model prvého rádu*. Jedným z najznámejších testov na overenie autokorelácie prvého rádu sa používa **Durbin-Watsonov test**. Za nulovú hypotézu uvažujeme

$$H_0 : \rho = 0$$

Testovacia štatistika má tvar

$$DW = \frac{\sum_{t=2}^{T}(\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\widehat{\epsilon}_t^2}$$

Pri pozitívnej autokorelácii sú hodnoty malé a naopak pri zápornej autokorelácii sú hodnoty DW testu vysoké. Vieme to aproximovať nasledovne

$$DW = 2(1 - \widehat{\rho}),$$

$$\widehat{\rho} = \frac{\sum_{t=2}^{T} \widehat{\epsilon}_{t-1}\widehat{\epsilon}_t}{\sum_{t=1}^{T} \widehat{\epsilon}_t^2}$$

Pričom platí:

- Ak $\widehat{\rho} = 1$ tak $DW = 0$
- Ak $\widehat{\rho} = 0$ tak $DW = 2$
- Ak $\widehat{\rho} = -1$ tak $DW = 4$

Pri dynamických modeloch[3] hrajú významnú úlohu nasledujúce triedy:

- *Lineárny regresný model s autokorelovanými reziduami* (bez oneskorených premenných, ale s oneskorením v reziduálnej zložke)
- *Model rozložených časových oneskorení* (oneskorené vysvetľujúce premenné ale nie oneskorená vysvetľovaná premenná) – DL-model (distributed lag model)
- *Autoregresný model rozložených časových oneskorení* (oneskorená vysvetľovaná premenná; môžu byť oneskorené aj vysvetľujúce premenné)

2.3.4 Multikolinearita

Multikolinearita znamená vysokú vzájomnú korelovanosť vysvetľujúcich premenných (regresorov) a teda matica týchto regresorov (ozn. X) nemá plnú hodnosť. Vtedy sa jedná o *perfektnú multikolinearitu*. Ak hodnota determinantu matice $X^T X$ je blízka nule, nie je ľahké skonštruovať inverznú maticu. Vieme to urobiť len za cenu veľkých štatistických chýb pri odhade parametrov v regresnom modeli.

$$det(X^T X)^{-1} = 0$$

Najjednoduchší spôsob ako zistiť, či je prítomná, je podľa *výberového korelačného koeficientu* medzi dvomi vysvetľujúcimi premennými. Vysoké hodnoty (nezáleží na tom, či kladné alebo záporné) signalizujú multikolinearitu.

Za multikolinearitu sa nepovažuje vzájomná závislosť medzi vysvetľovanou a vysvetľujúcimi premennými. Model s multikolinearitou je citlivý aj na malé zmeny. Riešenie multikolinearity:

- Ignorácia
- Vynechanie vysvetľujúcich premenných, ktoré ju spôsobujú
- Transformácia vysvetľujúcich premenných, ktoré ju spôsobujú

[3] Keď prejdeme na závislosť medzi diferenciami, hovoríme o dynamických modeloch

Kapitola 3

Autokorelačné metódy pre ČR

V predposlednej kapitole sa budeme venovať Box-Jenkinsonovej metodológii, ktorá berie za základ konštrukciu modelu časového radu reziduálnej zložky (tj. zložky náhodného charakteru) a zaoberá sa analýzou časových radov na základe špeciálnych stochastických modelov ako sú ARMA resp. ARIMA[1] modely. Túto metodológiu objavili George Box a Gwilym Jenkins v roku 1970 a zaoberali sa tým ako nájsť nejlepšie predpovede do budúcna pomocou minulých hodnôt. Základným zdrojom budú knihy [1, 2, 10].

3.1 Box-Jenkinson metodológia

Box-Jenkinsonova metodológia sa vyznačuje tým, že trend a sezónnosť sú modelované stochasticky a kladie dôraz na (auto)korelačnú analýzu. Lineárne autoregresné modely s kĺzavým priemerom označujeme ako ARMA, čisto autoregresné modely AR a modely s kĺzavým priemerom MA. Základným predpokladom týchto modelov je, že hodnota náhodnej premennej X_t v čase závisí len od stochastickej (náhodnej) zložky a od predchádzajúcich náhodných premenných. Závislosť od predchádzajúcich hodnôt je lineárna.

3.1.1 AR(p) proces

Autoregresný proces $AR(p)$ môžeme zapísať v rovnicovom tvare ako

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + \epsilon_t \tag{3.1}$$

kde ϵ_t je biely šum a $\Phi_1, \Phi_2, \cdots, \Phi_p$ sú AR-koeficienty.

[1] Ak pre zdiferencovaný rad vytvoríme model, tak pre pôvodný rad sme vytvorili ARIMA model

Zavedieme si operátor spätného posunutia, ktorý označíme ako B, pre ktorý platí $B^i Y_t = Y_{t-i}$. AR proces vieme pomocou tohto operátora prepísať na tvar

$$(1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p) Y_t = \Phi_p(B) Y_t = \epsilon_t$$

Aby bol proces AR(p) stacionárny, korene rovnice

$$(1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p) = \Phi_p(B) = 0$$

musia ležať mimo jednotkového kruhu.

3.1.2 MA(q) proces

Model s kĺzavým priemerom $MA(q)$ môžeme zapísať v rovnicovom tvare ako

$$Y_t = \mu + \epsilon_t + \Theta_1 \epsilon_{t-1} + \Theta_2 \epsilon_{t-2} + \cdots + \Theta_q \epsilon_{t-q} \tag{3.2}$$

Po prepísaní pomocou operátora spätného posunutia B, dostaneme nasledovnú rovnicu

$$Y_t = (1 + \Theta_1 B + \Theta_2 B^2 + \cdots + \Theta_q B^q) \epsilon_t = \Theta_q(B) \epsilon_t$$

MA procesy sú vždy stacionárne. Pri týchto modeloch overujeme, či sú *invertovateľné*. Model je invertovateľný ak korene rovnice

$$\Theta_q(B) = 0$$

ležia mimo jednotkového kruhu.

3.1.3 Zmiešaný ARMA(p,q) proces

Proces $ARMA(p, q)$ vzniká zmiešaním $AR(p)$ a $MA(q)$ modelov. Všeobecný tvar pre tento proces je nasledovný

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \epsilon_t + \Theta_1 \epsilon_{t-1} + \Theta_2 \epsilon_{t-2} + \cdots + \Theta_q \epsilon_{t-q} \tag{3.3}$$

Pomocou spätného operátora

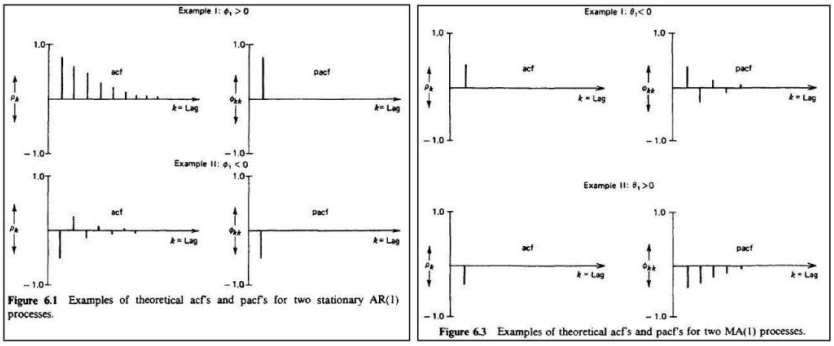
$$(1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p) Y_t = (1 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_q B^q) \epsilon_t$$

$ARMA(p, q)$ je stacionárny podľa $AR(p)$ časti a invertovateľný podľa $MA(q)$ časti modelu.

3.2 Konštrukcia modelu

3.2.1 Identifikácia modelu

Najskôr sa snažíme model identifikovať pomocou grafu autokorelačnej funkcie. Z neho vieme zistiť, či sa jedná o proces $AR(p)$ alebo $MA(q)$. Ak ACF klesá k nule (resp. striedavo klesá) a PACF je nenulová tak sa jedná o $AR(p)$ proces, kde p je rovné počtu nenulových členov parciálnej autokorelácie. Naopak, ak PACF klesá k nule alebo striedavo klesá, tak sa jedná o $MA(q)$ proces, kde q je dané počtom nenulových členov autokorelácie.



Obr. 3.1: Ukážka ACF a PACF procesov prvého rádu

Poznámka:[2] Ak členy autokorelačnej funkcie klesajú len veľmi pomaly, tak sa jedná o nestacionárny rad a treba ho stacionarizovať.

Ďalšou z možností ako identifikovať správny typ modelu je pomocou informačných kritérií ako je *Akaikeho informačné kritérium* - AIC

$$AIC(k,l) = ln\widehat{\sigma}^2_{k,l} + \frac{2(k+l+1)}{n}$$

alebo *Bayesovo informačné kritérium* - BIC

$$BIC(k,l) = ln\widehat{\sigma}^2_{k,l} + \frac{l\mathrm{nn}(k+l+1)}{n}$$

kde $\widehat{\sigma}^2_{k,l}$ je odhadnutý rozptyl bieleho šumu a k je počet parametrov.

3.2.2 Odhad modelu

Odhad modelu robíme väčšinou pomocou iteračných metód avšak ako pomocné nám môžu slúžiť momentové odhady. Napríklad pre model $AR(1)$ sú momentové odhady

$$\widehat{\varphi}_1 = r_1, \widehat{\sigma}^2 = \widehat{\sigma}^2_y(1 - \widehat{\varphi}_1 r_1)$$

kde musí platiť pre parameter $|r_1| < 1$. Pre model $MA(1)$ sú momentové odhady

$$\widehat{\theta}_1 = \frac{1 - \sqrt{1 - 4r_1^2}}{2r_1}, \widehat{\sigma}^2 = \frac{\widehat{\sigma}^2_y}{1 + \widehat{\theta}^2_1}$$

kde pre parameter r_1 musí byť splnená podmienka $|r_1| < 1/2$. Čo sa týka smerodajných odchýlok pre tieto odhady, tak pre proces $AR(1)$ je momentovým odhadom

$$\sigma(\widehat{\varphi}_1) \simeq \left(\frac{1 - \widehat{\varphi}^2_1}{n}\right)^{1/2}$$

a pre proces $MA(1)$ je smerodajná odchýlka

$$\sigma(\widehat{\theta}_1) \simeq \left(\frac{1 - \widehat{\theta}^2_1}{n}\right)^{1/2}$$

[2]Obr. 3.1 je zo stránky http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09/cv1.html

Pre zmiešané ARMA modely sa najčastejšie používajú NLS-odhady[3], ktoré využívajú predovšetkým Gauss-Newtonove algoritmy zamerané na minimalizáciu súčtu štvorcov.

3.2.3 Verifikácia modelu

Diagnostika modelu spočíva v overovaní, či je nami zvolený model správny. Jedným z aspektov na overenie správnosti sú testy stacionarity. Dôležité je overiť, či model spĺňa podmienku stacionarity, o ktorej sme hovorili v predchádzajúcej kapitole.

Dickey-Fullerov test

V praxi je väčšina časových radov nestacionárna. Buď sa jedná o *deterministickú nestacionaritu* alebo *stochastickú nestacionaritu*. Deterministickú spôsobuje trend. Po jeho odstránení pomocou regresii sa rad stáva stacionárnym. Stochastickej nestacionarity sa vieme zbaviť prechodom na prvé diferencie Δy_t. Príkladom je náhodná prechádzka s driftom

$$y_t = \alpha + y_{t-1} + \epsilon_t$$

ktorá sa dá jednoducho stacionarizovať prvými diferenciami

$$\Delta y_t = \alpha + \epsilon_t$$

čím dostávame posunutý biely šum, ktorý je stacionárny. Tento druh stacionarizovania svedčí o prítomnosti jednotkového koreňa. O prítomnosti jednotkového koreňa nám hovorí aj postupné pomalé klesanie korelogramu, avšak je možné to overiť aj pomocou rôznych testov. Jedným z najpoužívanejších je **Dickey-Fullerov test**.

Budeme predpokladať jednoduchý AR(1) proces daný vzťahom

$$y_t = \rho y_{t-1} + \epsilon_t$$

kde ρ je koeficient a ϵ_t biely šum. Ak koeficient $\rho = 1$, tak sa jedná o jednotkový koreň a teda model je nestacionárny. Regresný model vieme prepísať pomocou prvých diferencií

$$\Delta y_t = (\rho - 1) y_{t-1} + \epsilon_t = \delta y_{t-1} + \epsilon_t$$

kde Δ označuje 1. diferencie. Keď budeme testovať prítomnosť jednotkového koreňa, budeme testovať nulovú hypotézu

$$H_0 : \quad \Delta y_t = \delta y_{t-1} + \epsilon_t \quad \text{pre} \quad \delta = 0$$

a alternatívu zapíšeme ako

$$H_1 : \quad \Delta y_t = \alpha + \beta t + \delta y_{t-1} + \epsilon_t \quad \text{pre} \quad \delta < 0$$

kde α je absolútny člen a β smernica. Testujeme významnosť regresného parametra δ v modeli.

$$DF = \frac{\widehat{\delta}}{\widehat{\sigma}(\widehat{\delta})}$$

[3]NLS = Nonlinear Least Squares, viac v publikácii [13]

Jedná sa o t-rozdelenie avšak s "ťahšími" koncami ako pri klasickom t-rozdelení, tj. aby sme mohli zamietnuť nulovú hypotézu, potrebujeme významejšie hodnoty t-pomeru.

Poznámka: Klasický Dickey-Fullerov test sa používa za predpokladu, že ϵ_t (reziduálna zložka) je nezávislým bielym šumom. Ak Δy_t obsahuje autokorelovanosť, tak musíme použiť rozšírený Dickey-Fullerov test (ADF test). Rozdiel spočíva v tom, že za nulovú hypotézu budeme brať

$$H_0 : \quad \Delta y_t = \delta y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \epsilon_t \quad \text{pre} \quad \delta = 0$$

Poznámka: Pri deterministickej nestacionarite sa neodporúča používať diferencovanie, pretože to vedie k vzniku neinvertovateľného MA procesu.

Ako ďalšie potrebujeme overiť *normalitu* reziduí. Použijeme Jarqueho-Berov test, ktorý je založený na súčasnom testovaní tretieho normovaného momentu[4] (šikmosti) a štvrtého momentu(špicatosti). Vychádza zo skutočnosti, že šikmosť normálneho rozdelenia je rovná nule a špicatosť je tri. Ak si reziduá označíme ako ϵ_t, tak j-ty normovaný moment vypočítame pomocou vzťahu

$$\widehat{m}_j = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^j, \qquad j = 2, 3, 4$$

Šikmosť vychádza z predpokladu, že medzi kvartilmi je rovnaký počet prvkov a nadobúda hodnoty $< -1, 1 >$. Testovacie kritérium pre šikmosť je

$$SK = \sqrt{\frac{n}{6} * \frac{\widehat{m}_3^2}{\widehat{m}_2^3}}$$

Špicatosť sú hodnoty, ktoré charakterizujú sústredenie početnosti [5] okolo nejakej hodnoty znaku. Testovacím kritériom v tomto prípade je

$$SP = \sqrt{\frac{n}{24} \left(\frac{\widehat{m}_4}{\widehat{m}_2^2} - 3 \right)}$$

Pri Jarque-Beryho teste za nulovú hypotézu chápeme normalitu reziduí. Testovacia štatistika je

$$JB = SK^2 + SP^2$$

Ak je splnená nulová hypotéza[6], tak šikmosť aj špicatosť sú asymptoticky normované z normálneho rozdelenia $N(0, 1)$ a štatistika JB má rozdelenie $\chi^2(2)$.

Takisto je potrebné overiť biely šum, tj. či má nulovú strednú hodnotu, konštantný rozptyl, či nekoreluje a je z normálneho rozdelenia. Často sa používajú takzvané Q-testy[7], ktoré testujú významnosť prvých K autokorelácií odhadnutého bieleho šumu. Konštanta $K \simeq \sqrt{n}$, kde n je dĺžka časového radu. V praxi

[4] Normované momenty sú momenty smerodajnej premennej

[5] Čím sú početnosti viac sústredené okolo konkrétnej hodnoty, tým je vrchol špicatejší

[6] Zamietnutie nulovej hypotézy je často z dôvodu, že reziduá nemajú konštantný rozptyl (je tam prítomná heteroskedasticita)

[7] Q-testy sa používajú na nájdenie a odstránenie "outlierov", čo sú hodnoty, ktoré sa výrazne odlišujú od ostatných hodnôt

sa pre preocesy $ARMA(p,q)$ najčastejšie využíva *Boxova-Pierceova štatistika*[8]

$$Q = n\sum_{k=1}^{K}(r_k(\widehat{\epsilon}_t))^2 \geq \chi^2_{1-\alpha}(K-p-q)$$

3.2.4 Aplikácia modelu

Ak máme overené, že nami zvolený model je správny, tak je viacero spôsobov ako ho prakticky aplikovať. Tento model vieme použiť na analýzu minulého vývoja alebo prognostickú aplikáciu. My sa budeme zaoberať určovaním prognóz do blízkej budúcnosti. Pre model vypočítame hodnoty vysvetľovaných premenných pre $t = 1, 2, \cdots, n$ a prognózy overíme pre obdobie od $t = n + 1, n+2, \cdots, n+m$. Dôležitým predpokladom pre prognózy je nemennosť štruktúry modelu a jeho stacionarita.

3.3 Predpovede

Pri Box-Jenkinsonovej metodológii je jednoduché určiť predpovede. Často sa jedná o *bodové predpovede*, ktoré je možné stanoviť za predpokladu ak sú parametre modelu stabilné a máme k dispozícii hodnoty vysvetľujúcich premenných.

Pre jednoduchosť si ukážeme lineárnu predpoveď pre všeobecný ARMA(p,q) model (3.3) s nulovou strednou hodnotou. Základný vzťah pre skutočný výpočet predpovedí je

$$\widehat{Y}_{t+k}(t) = \Phi_1\widehat{Y}_{t+k-1}(t)+\cdots+\Phi_p\widehat{y}_{t+k-p}(t)+\widehat{\epsilon}_{t+k}(t)+\Theta_1\widehat{\epsilon}_{t+k-1}(t)+\cdots+\Theta_q\widehat{\epsilon}_{t+k-q}(t)$$

kde $t+k$ označuje predpoveď v čase t posunutú o k krokov dopredu.

Chybu predpovedi určíme ako

$$e_t = y_t - \widehat{y}_t$$

kde $\widehat{y}_t$ predstavuje nami odhadnutú hodnotu a y_t skutočnú hodnotu. Z tohto dôvodu sa predpovede často robia na známich dátach, aby sme mohli overiť správnosť predpovedí. Chyby spôsobuje predovšetkým reziduálna zložka. Medzi najčastejšie používané miery merania kvality predpovedí sa používa:

- *suma štvorcov náhodnej chyby* (Sum of Squared Errors - SSE)

$$SSE = \sum_{t=1}^{n}(y_t - \widehat{y}_t)^2 = \sum_{t=1}^{n} e_t^2$$

- *priemerná reziduálna odchýlka* (Mean of Squared Errors - MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y}_t)^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

[8] Viac v publikácii [14]

- *priemerná percentuálna odchýlka* (Mean Percent Error - MPE)

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - \widehat{y}_t)^2}{y}.100\% = \frac{1}{n}\sum_{t=1}^{n}\frac{e_t^2}{y}.100\%$$

- *priemerná absolútna reziduálna odchýlka* (Mean Absolute Error - MAE)

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y - \widehat{y}| = \frac{1}{n}\sum_{t=1}^{n}|e_t|$$

Poznámka: SSE hovorí o tom, akú časť variability časového radu nemožno vysvetliť pôsobením časovej premennej. Čím viac sa MSE blíži k nule, tým je prognóza presnejšia.

Kapitola 4

Praktická časť

V praktickej časti tejto práce sa zameriame na reálne dáta, na ktorých si ukážeme konkrétny model. Budeme využívať 2 časové rady s dennými hodnotami (vrátane víkendov), konkrétne spotrebu plynu na Slovenskom území a spriemerovaný vývoj teploty na tom istom území. Vzhľadom na denný charakter dát za obdobie jedného roku, si rady rozdelíme na štyri časti podľa ročných období[1]. Správnosť modelu overíme pomocou rôznych testov a následne urobíme krátkodobé predikcie, ktoré aj overíme graficky. Všetky modely, testy a predikcie budeme robiť v programe EViews [2] (verzia číslo 5).

Spotreba plynu

Ako prvé sa zameriame na dáta sledujúce dennú spotrebu plynu v m^3 na Slovenskom území, podľa predajného portfólia Slovenského plynárenského priemyslu, a.s. (ozn. SPP), za obdobie od 20.12.2008 do 20.12.2009, ktoré záhŕňajú v sebe spotrebu domácností aj veľkých odberateľov, ako sú rôzne továrne. Graf pre toto obdobie je nasledovný:

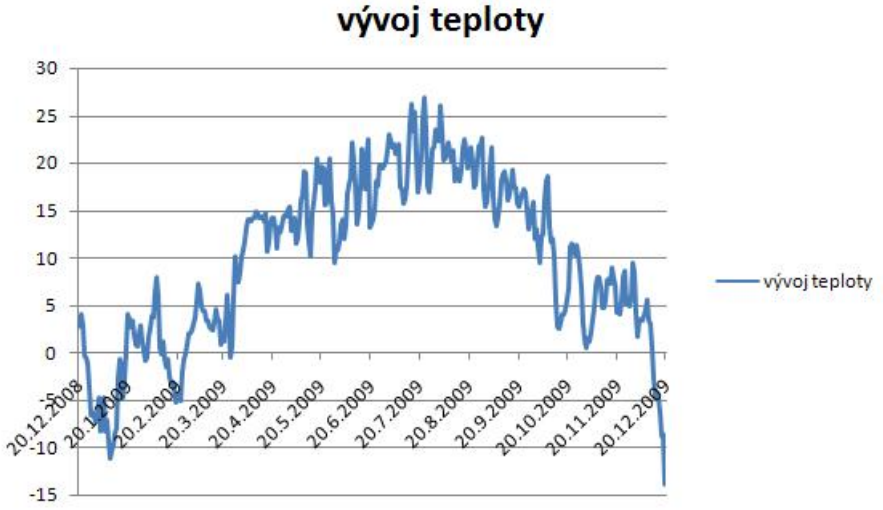
[1] Keby sa jednalo o 1/4 ročné dáta, ktoré by sme mali za dlhšie obdobie, bolo by potrebné sa zbaviť sezónnosti.

[2] pri práci s programom EViews využijeme príručky [4, 5]

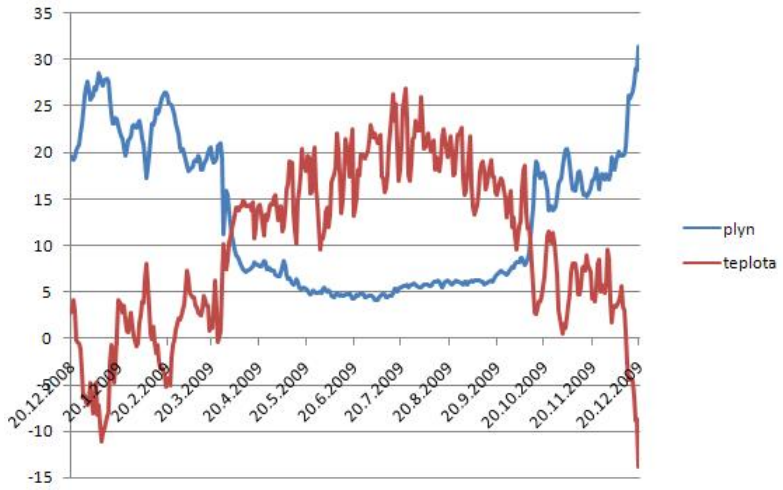
Na grafe je vidieť, že priližne od apríla do októbra je výrazne nižšia spotreba zemného plynu, čo sa aj dalo očakávať, keďže v lete nie je potrebné vykurovanie domácností, kdežto v zimných mesiacoch je spotreba oveľa vyššia.

Vývoj teploty

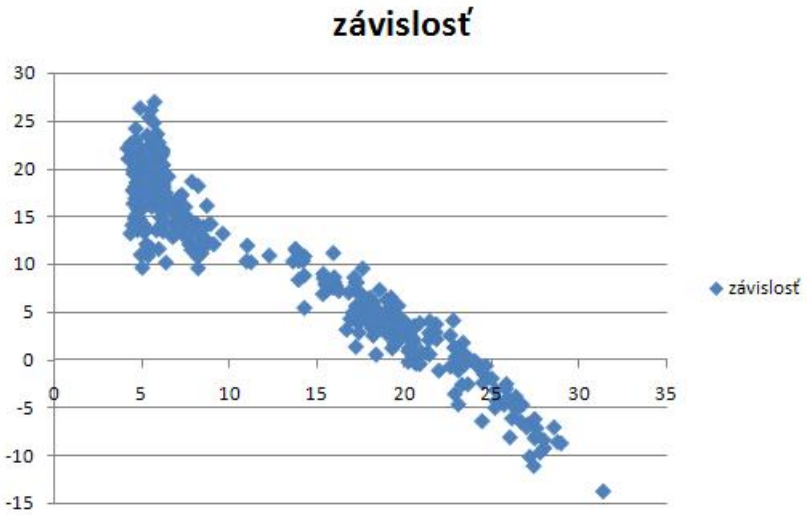
Dáta pre vývoj teploty ovzdušia predstavujú spriemerované hodnoty na celé územie Slovenska a sú získané zo Slovenského hydrometeorologického ústavu. Hodnoty radu budeme brať za rovnaké obdobie ako pre spotrebu plynu a taktiež budú denného charakteru. Graf pre teplotu:



Na grafe pre vývoj teploty je vidieť, že v období približne od apríla do októbra nastalo oteplenie, čo výrazne ovplyvnilo aj spotrebu plynu. Spoločný graf pre teplotu a plyn (kde sme plyn predelili 1 000 000):



Teraz môžeme vidieť jasnú zápornú koreláciu medzi teplotou a spotrebou plynu. Čím je vyššia teplota ovzdušia, tým je nižšia spotreba plynu. Koeficient korelácie je -0,954324527 a ukážeme si aj graf závislosti:

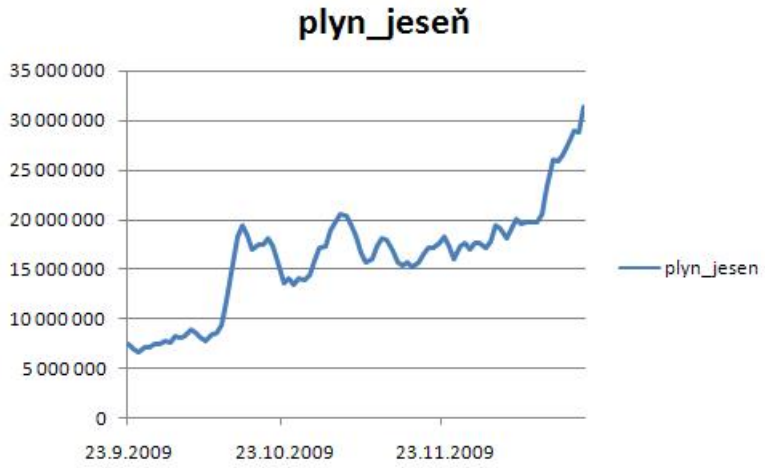


Pre jednoduchosť si dáta rozdelíme podľa ročných období na zimu (20/12/2008-20/3/2009), jar (20/3/2009-21/6/2009), leto (21/6/2009-23/9/2009) a jeseň (23/9/2009-20/12/2009). Modely budeme robiť najskôr len pre spotrebu plynu a neskôr pridáme aj vplyv teploty. Správnosť modelov overíme pomocou rôznych testov a pomocou predikcií na priližne najbližší mesiac v tom ktorom ročnom období. Čo sa týka sezónnosti, tak vzhľadom na rozdelenie radov na menšie úseky kvôli denným dátam sa so sezónnosťou pri tvorbe modelov nestretneme.

4.1 Model pre plyn - JESEŇ

Ako prvé si zoberieme dáta pre obdobie od 23/9/2009 do 20/12/2009, teda jeseň. Model budeme robiť len z dát po 1/12/2009 a na zvyšných si overíme správnosť predikcií. Najskôr sa pokúsime urobiť model pomocou ARMA procesov len s využitím dát pre spotrebu plynu, časom do modelu pridáme vplyv teploty a overíme správnosť výsledného modelu pomocou predikcií na najbližšie obdobie.

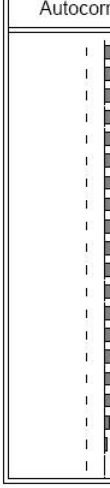
Priebeh spotreby plynu pre toto obdobie vyzerá nasledovne:



Dáta postupne stúpajú od hodnoty približne 5 miliónov po 30 miliónov. Je to spôsobené predovšetkým poklesom teploty ovzdušia. Korelogram pre toto obdobie:

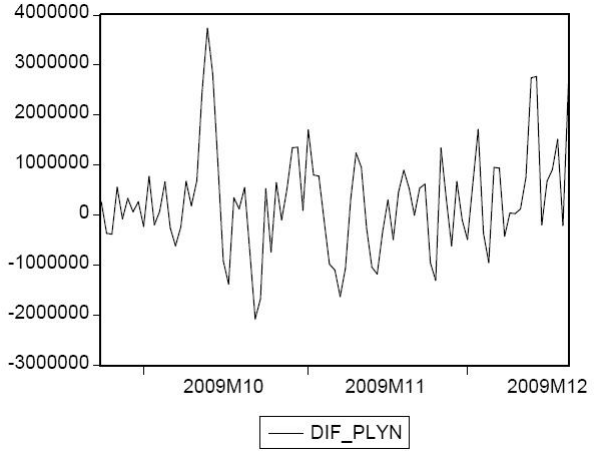
Correlogram of PLYN

Date: 04/09/11 Time: 19:53
Sample: 9/23/2009 12/20/2009
Included observations: 89



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.922	0.922	78.255	0.000
		2	0.840	-0.071	143.88	0.000
		3	0.749	-0.098	196.73	0.000
		4	0.667	0.008	239.12	0.000
		5	0.589	-0.025	272.52	0.000
		6	0.518	-0.001	298.73	0.000
		7	0.444	-0.075	318.21	0.000
		8	0.384	0.044	332.98	0.000
		9	0.344	0.089	344.94	0.000
		10	0.317	0.042	355.24	0.000
		11	0.291	-0.033	364.01	0.000
		12	0.244	-0.166	370.28	0.000
		13	0.207	0.056	374.85	0.000
		14	0.180	0.055	378.33	0.000
		15	0.148	-0.077	380.74	0.000
		16	0.124	0.021	382.45	0.000
		17	0.098	-0.018	383.53	0.000
		18	0.063	-0.051	383.98	0.000
		19	0.036	0.006	384.13	0.000
		20	0.011	-0.043	384.15	0.000

Ako bolo spomenuté v kapitole 3.2.1., pomalý pokles hodnôt korelogramu signalizuje nestacionaritu. Jedná sa o stochastickú nestacionaritu, ktorej sa zbavíme prechodom na prvé diferencie. Po tomto prechode sa nám priebeh dát zmení na nasledujúci tvar:



Tento rad je už stacionárny a môžeme prejsť k identifikácii modelu. Ako už bolo spomenuté, model určíme pomocou Box-Jenkinsonovej metodológie s tým, že najskôr nebudeme brať do úvahy vplyv teploty. Pri tomto ročnom období to bude autoregresný model s kĺzavými priemermi ARIMA(2,1,1). Po odstránení nesignifikantných hodnôt (tj. parametrov, ktorých p-hodnota je vyššia ako 5%) dostaneme následný výstup z Eviews:

Dependent Variable: DIF_PLYN
Method: Least Squares
Date: 04/09/11 Time: 19:12
Sample (adjusted): 9/25/2009 12/01/2009
Included observations: 68 after adjustments
Convergence achieved after 17 iterations
Backcast: 9/24/2009

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	1.396481	0.161572	8.643069	0.0000
AR(2)	-0.556346	0.108849	-5.111192	0.0000
MA(1)	-0.826262	0.153721	-5.375081	0.0000

R-squared	0.313010	Mean dependent var	149180.1
Adjusted R-squared	0.291872	S.D. dependent var	1022199.
S.E. of regression	860184.0	Akaike info criterion	30.21080
Sum squared resid	4.81E+13	Schwarz criterion	30.30871
Log likelihood	-1024.167	Durbin-Watson stat	1.900554

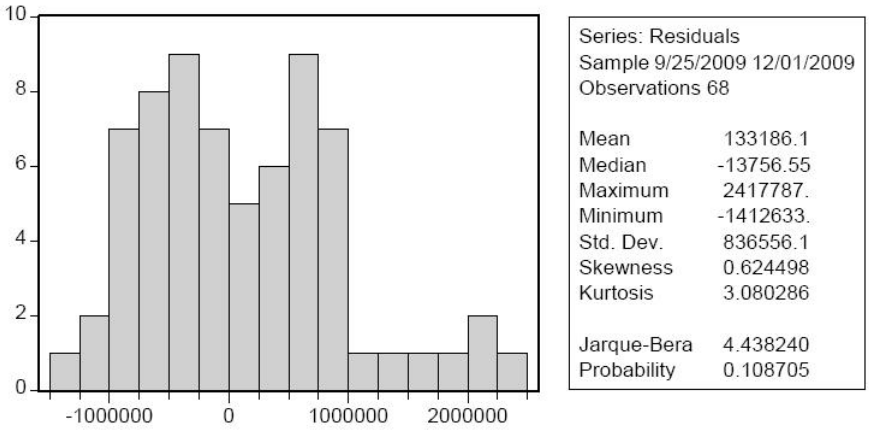
Inverted AR Roots	.70-.26i	.70+.26i
Inverted MA Roots	.83	

Z tabuľky je možné vyčítať hodnoty jednotlivých koeficientov, hodnoty rôznych testov ako je napríklad Koeficient determinácie, Durbin-Watsonova štatistika, Akaikeho kritérium a iné. Model máme odhadnutý a môžeme prejsť k jeho verifikácii.

Testovanie modelu

Ako prvé otestujeme, či sa v tomto, nami vytvorenom modeli nachádza biely šum. Najskôr je potrebné overiť normalitu, aby sme následne mohli pomocou Q-štatistiky určiť či sa v modeli nachádza alebo nenachádza biely šum.

Overenie normality:



Normalitu môžeme potvrdiť pomocou grafu resp. pomocou p-hodnoty v pravej časti tabuľky. Hodnoty vyššie ako 5% signalizujú normálne rozdelenie reziduí. V našom prípade je táto p-hodnota približne 10%, takže môžeme overiť biely šum pomocou Q-testov.

Correlogram of Residuals

Date: 04/09/11 Time: 19:14
Sample: 9/25/2009 12/01/2009
Included observations: 68
Q-statistic probabilities adjusted for 3 ARMA term(s)

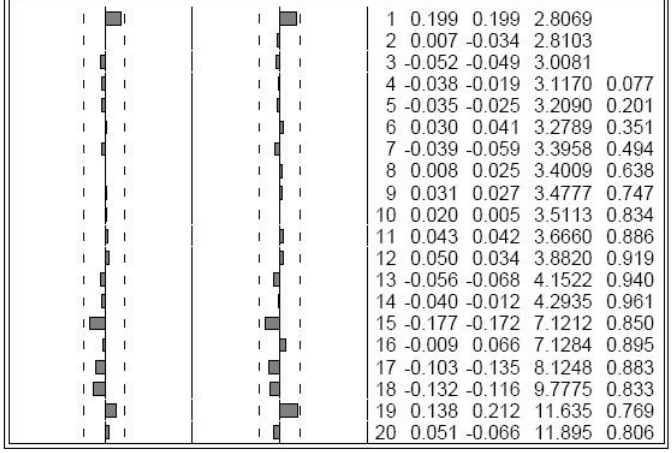
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.021	0.021	0.0327	
		2	-0.098	-0.098	0.7243	
		3	-0.025	-0.020	0.7689	
		4	-0.015	-0.024	0.7849	0.376
		5	-0.099	-0.104	1.5221	0.467
		6	0.167	0.171	3.6667	0.300
		7	-0.072	-0.107	4.0679	0.397
		8	0.028	0.068	4.1298	0.531
		9	-0.146	-0.175	5.8462	0.441
		10	-0.063	-0.048	6.1673	0.520
		11	0.036	0.045	6.2770	0.616
		12	-0.068	-0.154	6.6655	0.672
		13	-0.053	0.013	6.9119	0.734
		14	-0.020	-0.122	6.9479	0.803
		15	0.019	0.081	6.9819	0.859
		16	0.087	0.066	7.6749	0.864
		17	0.114	0.081	8.8775	0.839
		18	-0.017	0.022	8.9045	0.882
		19	0.000	-0.034	8.9045	0.917
		20	0.061	0.139	9.2747	0.931

Pri Q-štatistike p-hodnota nižšia ako 5% signalizuje, že sa v modeli nenachádza biely šum. V tabuľke pre korelogram pre reziduály sú všetky p-hodnoty

vyššie ako 5%, a teda sa do rádu jedna nachádza biely šum, a môžeme sa pozrieť do rádu dva.

Correlogram of Residuals Squared

Date: 04/09/11 Time: 19:14
Sample: 9/25/2009 12/01/2009
Included observations: 68
Q-statistic probabilities adjusted for 3 ARMA term(s)



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.199	0.199	2.8069	
		2	0.007	-0.034	2.8103	
		3	-0.052	-0.049	3.0081	
		4	-0.038	-0.019	3.1170	0.077
		5	-0.035	-0.025	3.2090	0.201
		6	0.030	0.041	3.2789	0.351
		7	-0.039	-0.059	3.3958	0.494
		8	0.008	0.025	3.4009	0.638
		9	0.031	0.027	3.4777	0.747
		10	0.020	0.005	3.5113	0.834
		11	0.043	0.042	3.6660	0.886
		12	0.050	0.034	3.8820	0.919
		13	-0.056	-0.068	4.1522	0.940
		14	-0.040	-0.012	4.2935	0.961
		15	-0.177	-0.172	7.1212	0.850
		16	-0.009	0.066	7.1284	0.895
		17	-0.103	-0.135	8.1248	0.883
		18	-0.132	-0.116	9.7775	0.833
		19	0.138	0.212	11.635	0.769
		20	0.051	-0.066	11.895	0.806

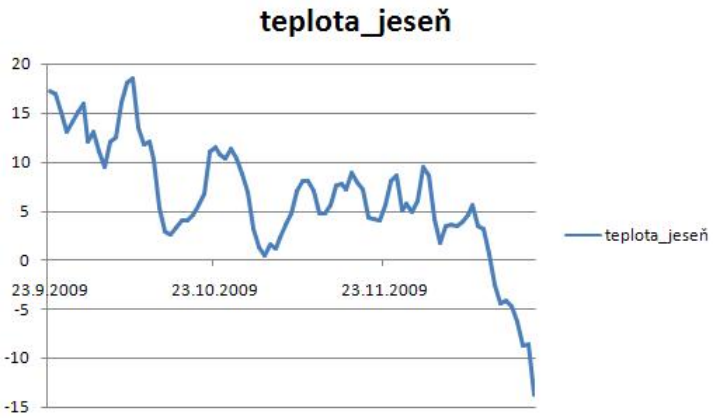
Aj pre druhé mocniny reziduí sa nám potvrdilo, že biely šum je prítomný. Ako ďalší z testov použijeme Ramseyho test na stabilitu modelu.

Ramsey RESET Test:			
F-statistic	0.241127	Probability	0.786476
Log likelihood ratio	0.526876	Probability	0.768405

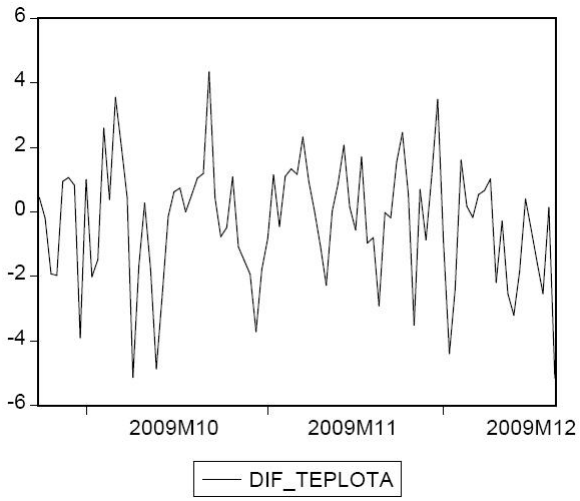
Na základe p-hodnoty môžeme tvrdiť, že model je dobre špecifikovaný, t.j. nepodarilo sa nám zamietnuť nulovú hypotézu o dobrej špecifikácii modelu.

Posledným z testov bude Whiteov test na zistenie prítomnosti heteroskedasticity, ale z technických príčin nebolo možné tento test vykonať, a preto budeme predpokladať, že heteroskedasticita nie je prítomná.

Po overení všetkých potrebných testov môžeme pridať vplyv teploty. Graf teploty pre jeseň:



Priebeh grafu ukazuje, že v tomto ročnom období sa ochladilo z približne 15 stupňov až na takmer -15 stupňov v decembri, čo ako sme už spomínali, výrazne ovplyvnilo spotrebu plynu. Podobne ako pri plyne, aj tento rad nie je stacionárny. Jedná sa taktiež o stochastickú nestacionaritu, ktorej sa zbavíme prechodom na prvé diferencie. Graf pre zdiferencovaný rad je nasledovný.



Keď si doplníme k vyššie spomenutému modelu vplyv teploty, ktorú sme stacionarizovali, zmenia sa nám hodnoty v tabuľke a po upravení a odstránení nesignifikantných parametrov dostávame výslednú tabuľku.

Dependent Variable: DIF_PLYN
Method: Least Squares
Date: 04/09/11 Time: 19:36
Sample (adjusted): 9/26/2009 12/01/2009
Included observations: 67 after adjustments
Convergence achieved after 38 iterations
Backcast: 9/24/2009 9/25/2009

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	95616.94	57318.78	1.668161	0.1004
DIF_TEPLOTA	-247213.3	51218.73	-4.826619	0.0000
AR(1)	0.397604	0.114060	3.485925	0.0009
AR(2)	0.848346	0.071984	11.78514	0.0000
AR(3)	-0.474466	0.111818	-4.243215	0.0001
MA(2)	-0.941035	0.064718	-14.54060	0.0000

R-squared	0.534559	Mean dependent var	157082.2
Adjusted R-squared	0.496408	S.D. dependent var	1027819.
S.E. of regression	729383.3	Akaike info criterion	29.92307
Sum squared resid	3.25E+13	Schwarz criterion	30.12051
Log likelihood	-996.4229	F-statistic	14.01172
Durbin-Watson stat	1.805590	Prob(F-statistic)	0.000000

Inverted AR Roots	.69-.13i	.69+.13i	-.97
Inverted MA Roots	.97	-.97	

Je vidieť, že napríklad Koeficient determinácie sa výrazne zvýšil (z hodnoty približne 0,31 na hodnotu 0,53), čo znamená, že model sa zlepšil. Pridanie vplyvu teploty neovplyvnilo normalitu ani Q-štatistiku do prvého a druhého rádu. Nakoniec si ukážeme rovnicový tvar a overíme predikcie na tomto výslednom modeli[3].

Rovnicový tvar

Pomocou hodnôt z tabuliek si napíšeme rovnicový tvar pre daný model. Ak si spotrebu plynu označíme ako y_t, teplotu ako x_t potom vieme, že platí:

$$\Delta y_t = 95616.94 + u_t$$

$$u_t = 0.397604u_{t-1} + 0.848346u_{t-2} - 0.474466u_{t-3} + \epsilon_t - 0.941035\epsilon_{t-2}$$

$$(1 - 0.397604L - 0.848346L^2 + 0.474466L^3)u_t = (1 - 0.941035L^2)\epsilon_t$$

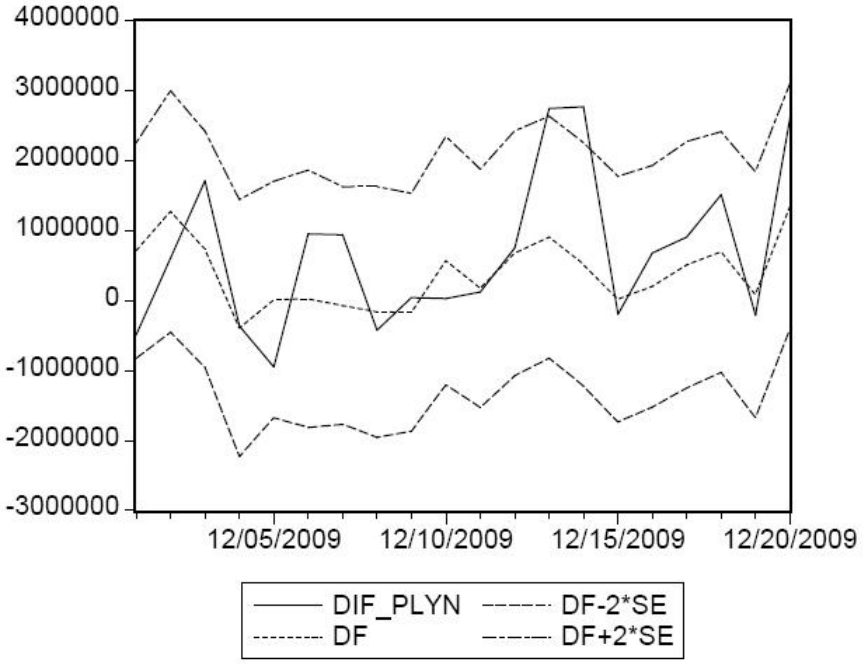
$$((1-0.397604L-0.848346L^2+0.474466L^3)(\Delta y_t - 95616.94) = (1-0.941035L^2)\epsilon_t$$

$$\Delta y_t - 0.397604\Delta y_{t-1} - 0.848346\Delta y_{t-2} + 0.474466\Delta y_{t-3} - C = \epsilon - 0.941035\epsilon_{t-2}$$

$$C = \frac{95616.94}{1 - 0.397604 - 0.848346 + 0.474466}$$

$$\Delta y_t = -247213.3\Delta x_t + C + 0.397604\Delta y_{t-1} + 0.848346\Delta y_{t-2} - 0.474466\Delta y_{t-3} + \epsilon - 0.941035\epsilon_{t-2}$$

[3] V tomto prípade sme sa rozhodli nechať aj nesignifikantnú konštantu C, pretože po jej odstránení a následnej úprave modelu sme dostali rovnaký predikčný graf.

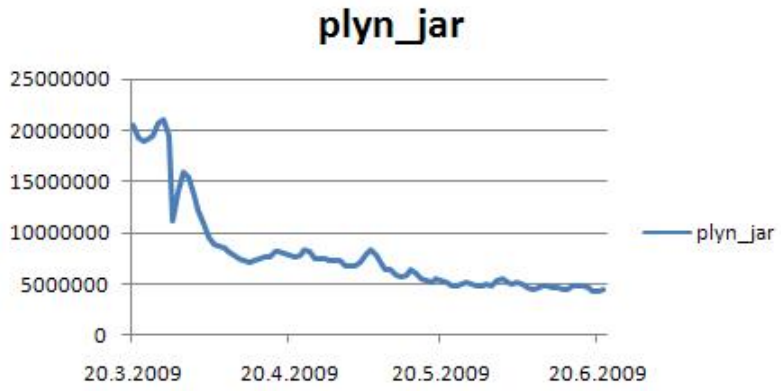


Neprerušovaná čiara zodpovedá skutočnému priebehu dát v období od 1/12/2009 do 20/12/2009, čiara označená ako DF je náš odhad, od ktorého je odpočítaná a pripočítaná dvakrát hodnota smerodajnej odchýlky. Pokiaľ sa skutočné hodnoty nachádzajú medzi smerodajnými odchýlkami, tak sme overili, že je náš model správny. Okolo 14/12/2009 skutočná hodnota spotreby plynu nespadá do nášho odhadu, čo je spôsobené náhlym zvýšením spotreby v období od 12/12/2009 do 14/12/2009 až o približne 6 miliónov. Napriek tomuto neočakávanému nárastu sa dá náš model považovať ako postačujúci a správne určený pre dané obdobie.

4.2 Model pre plyn - JAR

Ako ďalšie si zoberieme dáta pre prvé ročné obdobie. Jar berieme za obdobie od 20/3/2009 do 21/6/2009, ale ako v predchádzajúcom prípade, model spravíme len z hodnôt po 1/6/2009 a na zvyšných si overíme správnosť predikcií. Určovanie modelu a následné overovanie bude rovnaké ako pri jeseni.

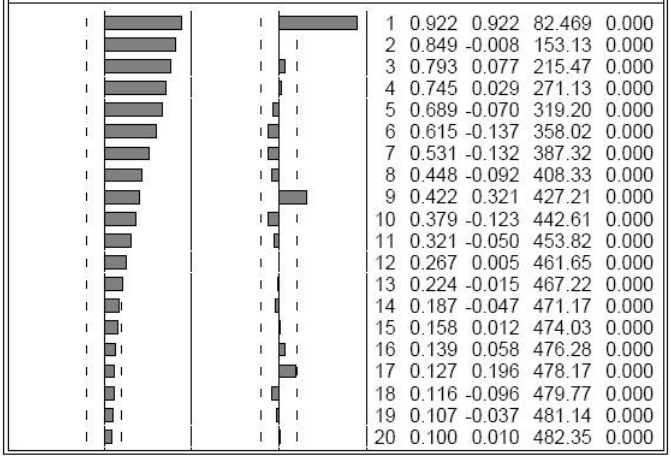
Priebeh spotreby plynu pre toto obdobie vyzerá nasledovne:



Dáta postupne klesajú a ustaľujú sa okolo hodnoty 5 000 000. Je to spôsobené predovšetkým zvyšovaním teploty ovzdušia, čo znamená, že plyn už nebol potrebný napríklad na vykurovanie domácností. Korelogram pre toto obdobie:

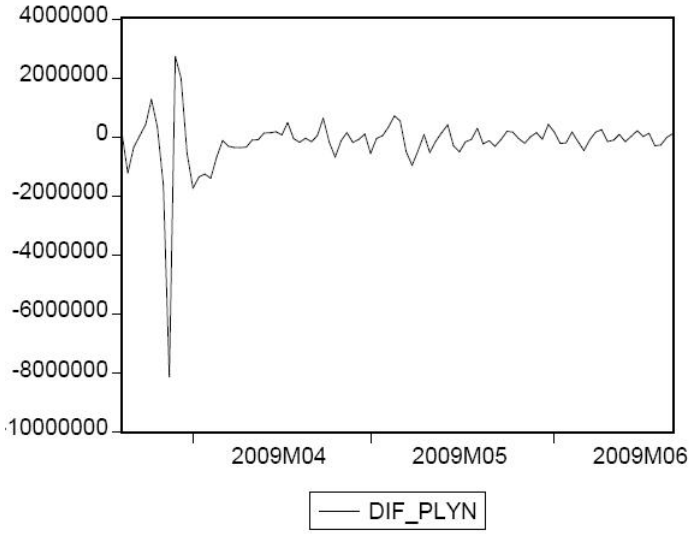
Correlogram of PLYN

Date: 03/19/11 Time: 11:31
Sample: 3/20/2009 6/21/2009
Included observations: 94



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.922	0.922	82.469	0.000
		2	0.849	-0.008	153.13	0.000
		3	0.793	0.077	215.47	0.000
		4	0.745	0.029	271.13	0.000
		5	0.689	-0.070	319.20	0.000
		6	0.615	-0.137	358.02	0.000
		7	0.531	-0.132	387.32	0.000
		8	0.448	-0.092	408.33	0.000
		9	0.422	0.321	427.21	0.000
		10	0.379	-0.123	442.61	0.000
		11	0.321	-0.050	453.82	0.000
		12	0.267	0.005	461.65	0.000
		13	0.224	-0.015	467.22	0.000
		14	0.187	-0.047	471.17	0.000
		15	0.158	0.012	474.03	0.000
		16	0.139	0.058	476.28	0.000
		17	0.127	0.196	478.17	0.000
		18	0.116	-0.096	479.77	0.000
		19	0.107	-0.037	481.14	0.000
		20	0.100	0.010	482.35	0.000

Aj tu je pomalý pokles hodnôt korelogramu signalizujúci nestacionaritu. Je to stochastická nestacionarita, ktorú odstránime prechodom na prvé diferencie, čím sa nám priebeh dát zmení na nasledujúci tvar:



Takýto rad je už stacionárny a môžeme prejsť k identifikácii modelu. Zo začiatku si budeme všímať len spotrebu plynu bez vplyvu teploty. Pri tomto ročnom období nám stačí autoregresný model bez kĺzavých priemerov. Po odstránení nesignifikantných hodnôt dostaneme následný výstup:

Dependent Variable: DIF_PLYN
Method: Least Squares
Date: 04/10/11 Time: 01:22
Sample (adjusted): 3/22/2009 6/01/2009
Included observations: 72 after adjustments
Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(2)	-0.254006	0.113981	-2.228500	0.0290

R-squared	0.039217	Mean dependent var	-191986.4
Adjusted R-squared	0.039217	S.D. dependent var	1155662.
S.E. of regression	1132775.	Akaike info criterion	30.73203
Sum squared resid	9.11E+13	Schwarz criterion	30.76365
Log likelihood	-1105.353	Durbin-Watson stat	1.899623

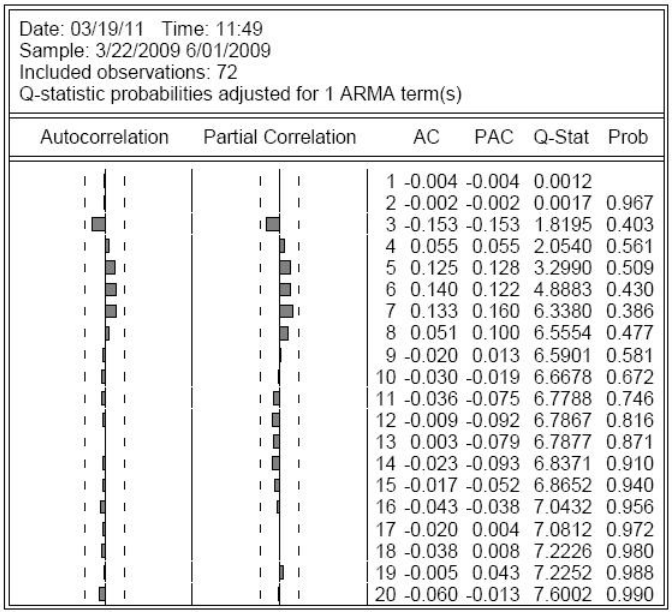
Konkrétne sa jedná o autoregresný proces do rádu dva, tj. AR(2). Model máme odhadnutý a môžeme začať overovať jeho správnosť.

Testovanie modelu

Najskôr otestujeme, či sa v modeli nachádza biely šum. Overíme to pomocou Q-štatistiky do rádu jedna aj dva.

Correlogram of Residuals

Date: 03/19/11 Time: 11:49
Sample: 3/22/2009 6/01/2009
Included observations: 72
Q-statistic probabilities adjusted for 1 ARMA term(s)

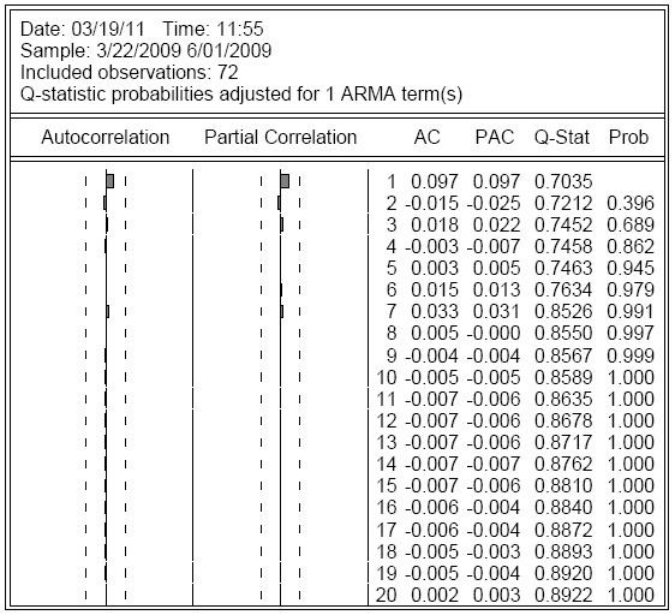


Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.004	-0.004	0.0012	
		2	-0.002	-0.002	0.0017	0.967
		3	-0.153	-0.153	1.8195	0.403
		4	0.055	0.055	2.0540	0.561
		5	0.125	0.128	3.2990	0.509
		6	0.140	0.122	4.8883	0.430
		7	0.133	0.160	6.3380	0.386
		8	0.051	0.100	6.5554	0.477
		9	-0.020	0.013	6.5901	0.581
		10	-0.030	-0.019	6.6678	0.672
		11	-0.036	-0.075	6.7788	0.746
		12	-0.009	-0.092	6.7867	0.816
		13	0.003	-0.079	6.7877	0.871
		14	-0.023	-0.093	6.8371	0.910
		15	-0.017	-0.052	6.8652	0.940
		16	-0.043	-0.038	7.0432	0.956
		17	-0.020	0.004	7.0812	0.972
		18	-0.038	0.008	7.2226	0.980
		19	-0.005	0.043	7.2252	0.988
		20	-0.060	-0.013	7.6002	0.990

Pri Q-štatistike je všade p-hodnota vyššia ako 5% a teda sa v modeli nachádza biely šum do rádu jedna a pozrieme sa na druhé mocniny reziduí.

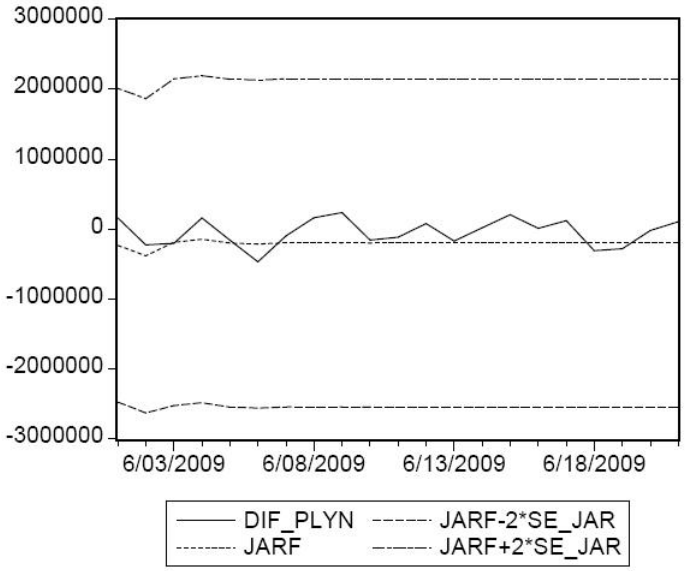
Correlogram of Residuals Squared

Date: 03/19/11 Time: 11:55
Sample: 3/22/2009 6/01/2009
Included observations: 72
Q-statistic probabilities adjusted for 1 ARMA term(s)



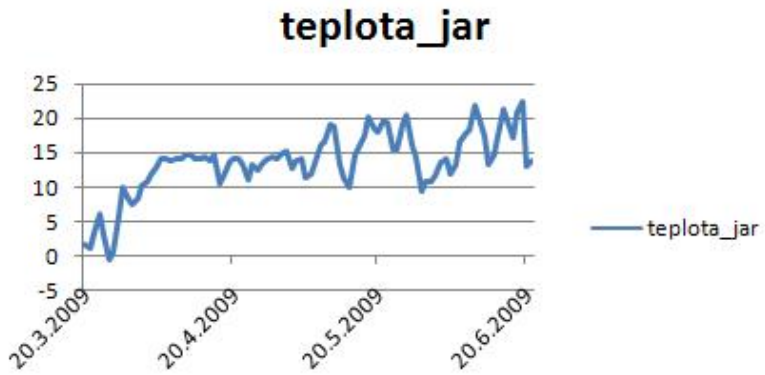
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.097	0.097	0.7035	
		2	-0.015	-0.025	0.7212	0.396
		3	0.018	0.022	0.7452	0.689
		4	-0.003	-0.007	0.7458	0.862
		5	0.003	0.005	0.7463	0.945
		6	0.015	0.013	0.7634	0.979
		7	0.033	0.031	0.8526	0.991
		8	0.005	-0.000	0.8550	0.997
		9	-0.004	-0.004	0.8567	0.999
		10	-0.005	-0.005	0.8589	1.000
		11	-0.007	-0.006	0.8635	1.000
		12	-0.007	-0.006	0.8678	1.000
		13	-0.007	-0.006	0.8717	1.000
		14	-0.007	-0.007	0.8762	1.000
		15	-0.007	-0.006	0.8810	1.000
		16	-0.006	-0.004	0.8840	1.000
		17	-0.006	-0.004	0.8872	1.000
		18	-0.005	-0.003	0.8893	1.000
		19	-0.005	-0.004	0.8920	1.000
		20	0.002	0.003	0.8922	1.000

Aj pri druhých mocninách reziduí sa nám potvrdilo, že biely šum je prítomný. Ako v predchádzajúcom období, aj teraz sme po overení normality, heteroskedasticity a Ramseyho teste určili model za správny, čo si môžeme potvrdiť ukážkou predikcií pre obdobie od 1/6/2009, zatiaľ bez pridaného vplyvu teploty ovzdušia.

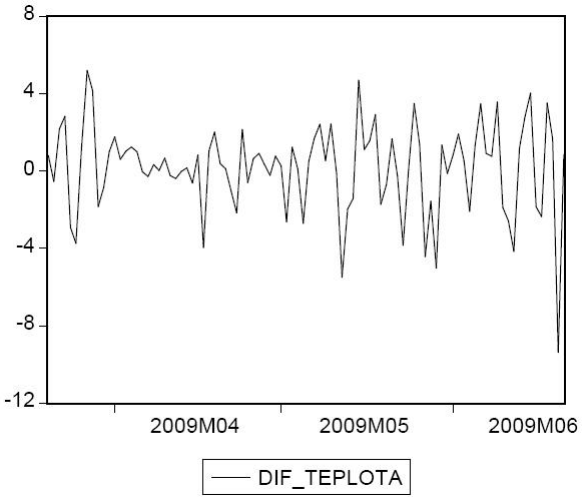


Modifikácia modelu

Ako už bolo spomínané, teraz sa pozrieme na to, či pridanie teploty ovplyvní náš model. Graf teploty pre jar:



Ani tento rad nie je stacionárny. Znova sa jedná o stochastickú nestacionaritu, ktorú odstránime prechodom na prvé diferencie. Zdiferencovaný rad vyzerá nasledovne.



Po pridaní takéhoto radu do predtým vytvoreného modelu sa nám zvýši hodnota Koeficientu determinácie, čo sa dá považovať za jedno z overení zlepšenia modelu. Q-štatistika aj po modifikácii modelu potvrdila prítomnosť bieleho šumu. Hodnoty koeficientov a niektorých štatistík sú znázornené v tabuľke.

Dependent Variable: DIF_PLYN
Method: Least Squares
Date: 04/10/11 Time: 01:25
Sample (adjusted): 3/22/2009 6/01/2009
Included observations: 72 after adjustments
Convergence achieved after 7 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DIF_TEPLOTA	-242021.5	59937.05	-4.037927	0.0001
AR(2)	-0.272919	0.113754	-2.399201	0.0191

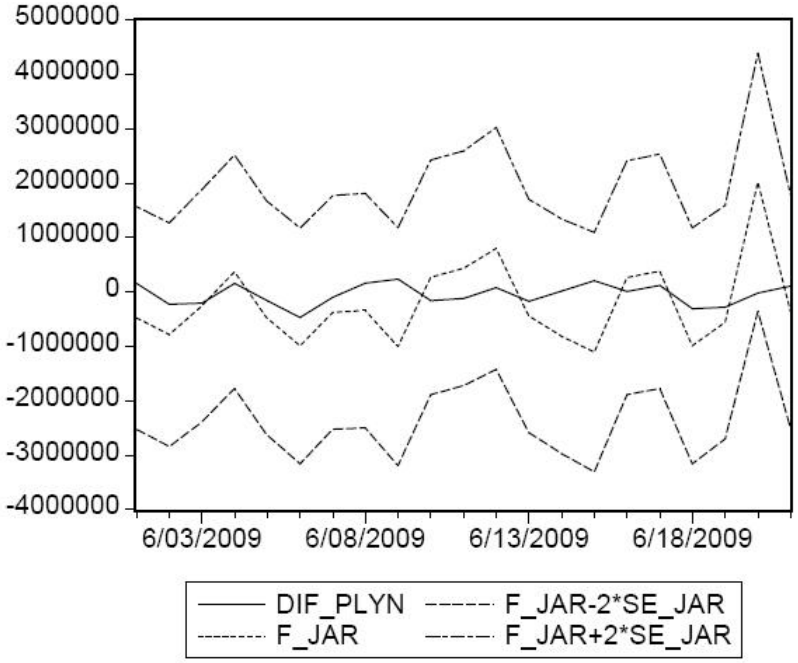
R-squared	0.220443	Mean dependent var	-191986.4
Adjusted R-squared	0.209307	S.D. dependent var	1155662.
S.E. of regression	1027625.	Akaike info criterion	30.55078
Sum squared resid	7.39E+13	Schwarz criterion	30.61403
Log likelihood	-1097.828	Durbin-Watson stat	2.148068

Rovnicový tvar

$$\Delta y_t = -242021.5\Delta x_t - 0.272919\Delta y_{t-2} + \epsilon$$

kde Δy_t je zdiferencovaná spotreba plynu a Δx_t zdiferencovaný vlyv teploty.

Overenie pomocou predikcií:

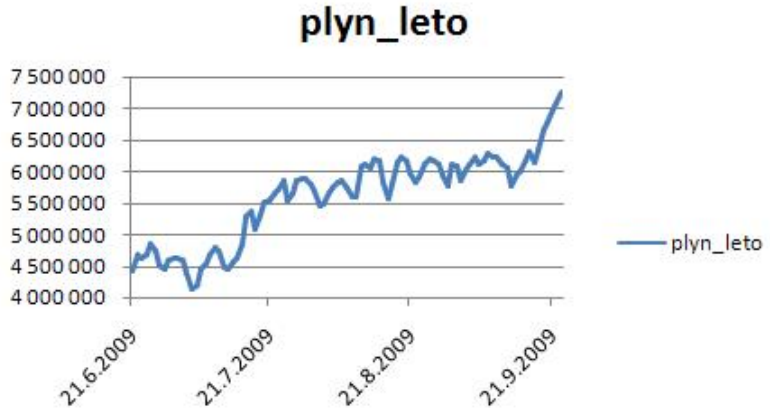


V porovnaní s predchádzajúcim grafom (bez vplyvu teploty) pre predikcie, tento nový graf vytvorený pomocou modifikovaného modelu lepšie opisuje skutočný priebeh dát. Znamená to, že vplyv teploty ovzdušia na jar výrazne ovplyvňuje spotrebu plynu na Slovenskom území a preto je potrebné do modelu tento časový rad zahrnúť.

4.3 Model pre plyn - LETO

Budeme pokračovať rovnako ako v predchádzajúcich prípadoch, len s dátami pre letné obdobie. Údaje berieme za obdobie od 21/6/2009 do 23/9/2009. S tým, že model určíme pomocou dát do 1/9/2009 a správnosť overíme na zvyšnom období. Postup pri identifikácii a verifikácii modelu bude totožný.

Dáta pre spotrebu plynu v lete:



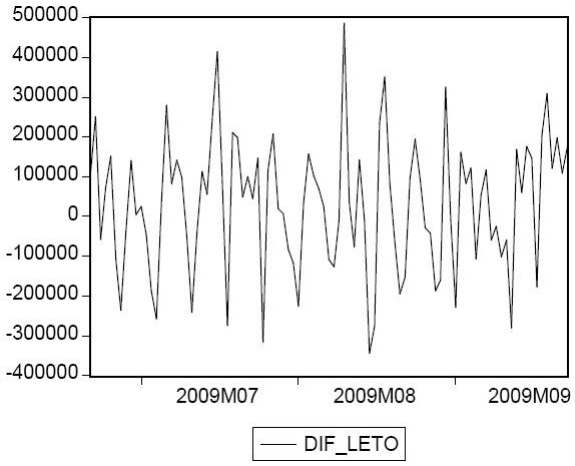
V tomto období sa spotreba plynu pohybuje v rozmedzí od 4 miliónov do 7,5 milióna narozdiel od jari, keď hodnoty dosahovali až 20 miliónov. Korelogram pre dané dáta:

Correlogram of PLYN

Date: 02/26/11 Time: 09:28
Sample: 6/21/2009 9/23/2009
Included observations: 95

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.928	0.928	84.341	0.000
		2	0.856	-0.034	156.88	0.000
		3	0.800	0.075	220.90	0.000
		4	0.758	0.071	279.03	0.000
		5	0.727	0.073	333.22	0.000
		6	0.711	0.102	385.53	0.000
		7	0.692	0.005	435.68	0.000
		8	0.641	-0.206	479.22	0.000
		9	0.592	0.008	516.80	0.000
		10	0.561	0.083	550.98	0.000
		11	0.541	0.039	583.11	0.000
		12	0.531	0.059	614.43	0.000
		13	0.514	-0.069	644.07	0.000
		14	0.484	-0.061	670.77	0.000
		15	0.435	-0.101	692.53	0.000
		16	0.386	-0.029	709.92	0.000
		17	0.342	-0.064	723.77	0.000
		18	0.308	-0.016	735.11	0.000
		19	0.288	0.046	745.17	0.000
		20	0.272	0.038	754.29	0.000

Podobne ako v predchádzajúcom prípade, aj tu hodnoty klesajú pomaly, čo znamená, že rad je stochasticky nestacionárny, preto ho stacionarizujeme pomocou diferencií. Použitím 1. diferencií sa dáta zmenia a dostaneme nasledovný graf:



Takýto rad je už stacionárny a môžeme odhadnúť model.

Dependent Variable: DIF_PLYN
Method: Least Squares
Date: 04/10/11 Time: 01:09
Sample (adjusted): 6/26/2009 9/01/2009
Included observations: 68 after adjustments
Convergence achieved after 16 iterations
Backcast: 6/24/2009 6/25/2009

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.235012	0.084263	2.789013	0.0070
AR(2)	-0.868314	0.172422	-5.035976	0.0000
AR(4)	-0.305727	0.110498	-2.766806	0.0074
AR(5)	-0.326182	0.083655	-3.899142	0.0002
MA(2)	0.472165	0.210912	2.238683	0.0287

R-squared	0.370195	Mean dependent var	15039.53
Adjusted R-squared	0.330208	S.D. dependent var	176697.3
S.E. of regression	144610.6	Akaike info criterion	26.67216
Sum squared resid	1.32E+12	Schwarz criterion	26.83536
Log likelihood	-901.8535	Durbin-Watson stat	1.859275

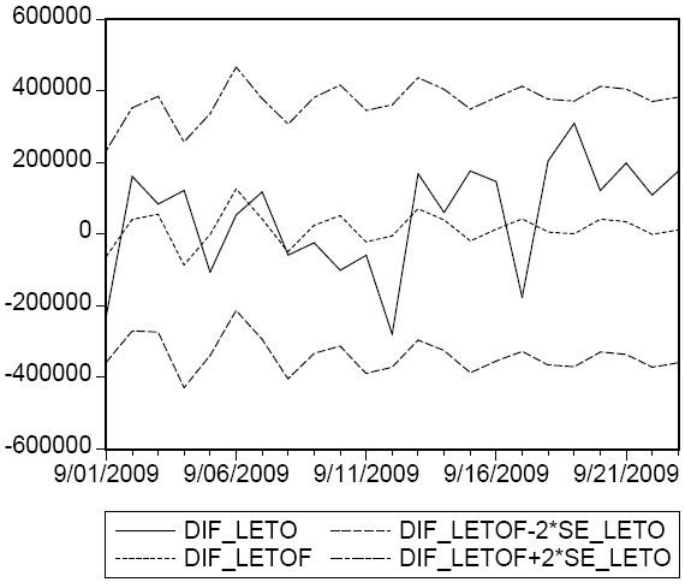
Inverted AR Roots	.53+.69i	.53-.69i	-.16+.90i	-.16-.90i
	-.51			

V tomto prípade sa jedná o autoregresný model s kĺzavými priemermi, tj. ARIMA(5,1,2) model očistený o nesignifikantné koeficienty. Model je stacionárny a invertovateľný, čo je znázornené v spodnej časti tabuľky. Pomocou Q-štatistiky sme overili, že sa tam nachádza biely šum pre reziduá aj druhé mocniny reziduí. Napíšeme si rovnicový tvar a či je model dobrý, overíme na predikciách urobených pre obdobie od 1/9/2009 do 23/9/2009.

Rovnicový tvar

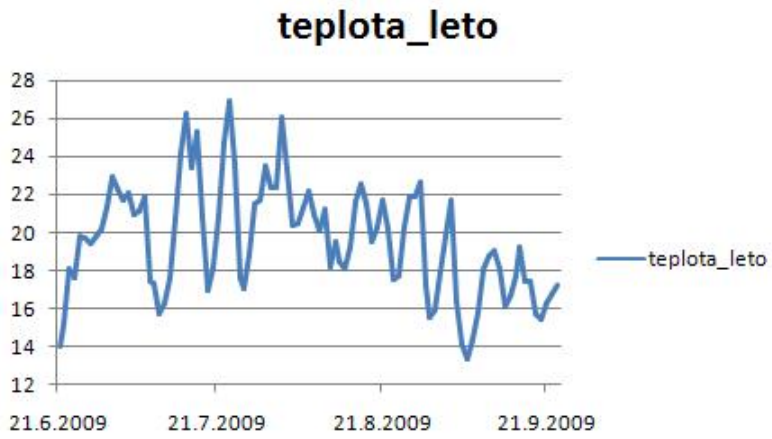
$\Delta y_t = 0.235012\Delta y_{t-1} - 0.868314\Delta y_{t-2} - 0.305727\Delta y_{t-4} - 0.326182\Delta y_{t-5} + \epsilon + 0.472165\epsilon_{t-2}$

Overenie predikcií:



Vplyv teploty

Graf pre teplotu v lete:



Tento rad je stacionárny, a preto nie je potrebné ho stacionarizovať a môžeme ho priamo pridať do modelu.

Dependent Variable: DIF_LETO
Method: Least Squares
Date: 02/26/11 Time: 10:19
Sample (adjusted): 6/26/2009 9/01/2009
Included observations: 68 after adjustments
Convergence achieved after 27 iterations
Backcast: 6/24/2009 6/25/2009

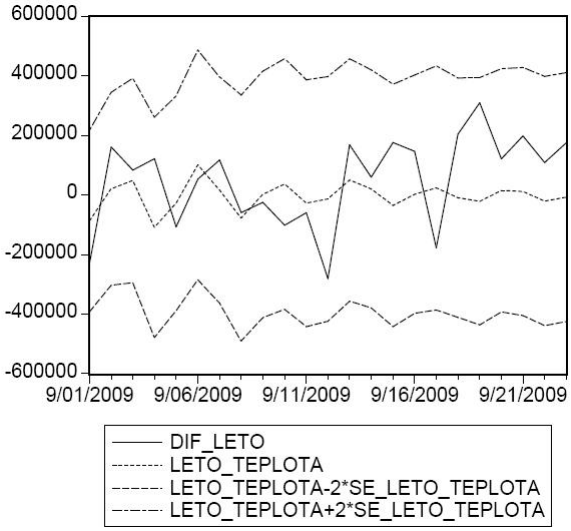
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-85414.62	142116.3	-0.601019	0.5501
TEPLOTA	5056.520	6884.681	0.734460	0.4655
AR(1)	0.228052	0.083839	2.720106	0.0085
AR(2)	-0.890875	0.172143	-5.175188	0.0000
AR(4)	-0.334519	0.111787	-2.992464	0.0040
AR(5)	-0.322869	0.086000	-3.754303	0.0004
MA(2)	0.450666	0.214283	2.103130	0.0396

R-squared	0.403037	Mean dependent var	15039.53
Adjusted R-squared	0.344320	S.D. dependent var	176697.3
S.E. of regression	143079.0	Akaike info criterion	26.67743
Sum squared resid	1.25E+12	Schwarz criterion	26.90591
Log likelihood	-900.0326	F-statistic	6.863995
Durbin-Watson stat	1.944728	Prob(F-statistic)	0.000013

Inverted AR Roots	.53+.70i -.50	.53-.70i	-.17+.90i	-.17-.90i

Podľa veľkosti pravdepodobnosti znázornenej v poslednom stĺpci v tabuľke vieme určiť, či je ten, ktorý parameter signifikantný alebo nie. Vidíme, že hodnota pri teplote je vyššia ako 5%, a teda tento parameter je nesignifikantný. Znamená to, že v lete nie je potrebné do modelu zahrnúť vplyv teploty, čo sa dalo aj očakávať, pretože nezáleží na tom, či je 16 alebo 26 stupňov, vykurovanie už nie je potrebné a spotreba plynu je nezávislá od teploty ovzdušia. Keďže sa nám model nezlepšil, použijeme predchádzajúci model, ktorý už máme napísaný v rovnicovom tvare, a pre ktorý už máme aj ukážku predikcií.

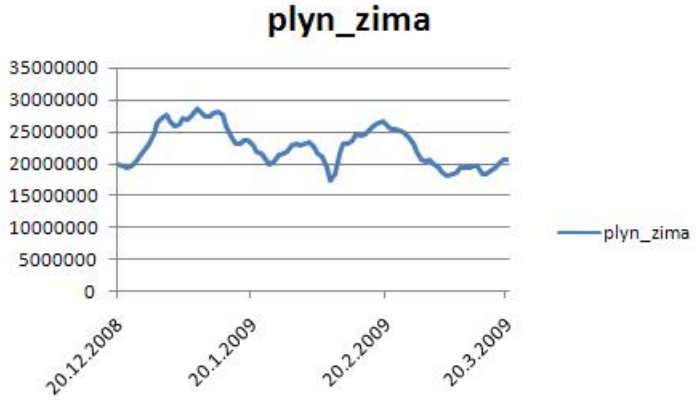
Ako dôkaz, že vplyv teploty je v tomto období nepodstatný si môžeme ukázať graf predikcií po pridaní teploty.



Graf nám potvrdil, že naše úvahy boli správne.

4.4 Model pre plyn - ZIMA

Ako posledné obdobie si zoberieme zimu. Z dôvodu dostupnosti dát, nebudeme brať toto obdobie od 20/12/2009 ale o rok skôr, tj. od 20/12/2008 po 20/3/2009. Rovnako ako v predchádzajúcich prípadoch aj tu odhadneme model z dát pre 2 mesiace a overíme od 1/3/2009 do konca obdobia. Spotreba plynu v zime vyzerá nasledovne:



V tomto prípade je spotreba plynu najvyššia spomedzi všetkých období a dosahuje hodnoty až 30 miliónov. Korelogram pre tieto dáta:

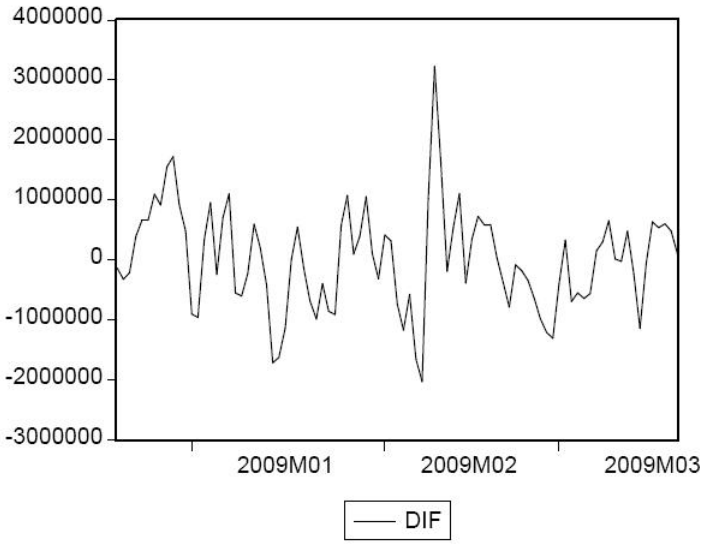
Correlogram of PLYN

Date: 02/06/11 Time: 11:36
Sample: 12/20/2008 3/20/2009
Included observations: 91

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.953	0.953	85.442	0.000
		2	0.861	-0.516	156.02	0.000
		3	0.763	0.144	211.97	0.000
		4	0.661	-0.176	254.45	0.000
		5	0.558	-0.004	285.06	0.000
		6	0.463	0.035	306.39	0.000
		7	0.373	-0.120	320.42	0.000
		8	0.277	-0.148	328.27	0.000
		9	0.184	0.064	331.76	0.000
		10	0.105	0.023	332.91	0.000
		11	0.039	-0.032	333.06	0.000
		12	-0.020	-0.025	333.11	0.000
		13	-0.073	-0.088	333.69	0.000
		14	-0.128	-0.132	335.50	0.000
		15	-0.194	-0.129	339.68	0.000
		16	-0.256	0.082	347.08	0.000
		17	-0.301	0.046	357.42	0.000
		18	-0.321	0.108	369.36	0.000
		19	-0.321	0.021	381.47	0.000
		20	-0.311	-0.088	393.00	0.000

Znova sa jedná o stochastickú nestacionaritu, kvôli pomalému poklesu hodnôt. Potrebujeme z tohto radu spraviť rad stacionárny. Použijeme 1. diferencie

a dostaneme graf s upravenými hodnotami.



Takýto rad už je stacionárny a môžeme pomocou neho odhadnúť model.

Dependent Variable: DIF
Method: Least Squares
Date: 02/06/11 Time: 11:28
Sample (adjusted): 12/24/2008 3/01/2009
Included observations: 68 after adjustments
Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	1.002761	0.119510	8.390602	0.0000
AR(2)	-0.815658	0.157503	-5.178669	0.0000
AR(3)	0.594226	0.156968	3.785654	0.0003
AR(4)	-0.295752	0.121991	-2.424375	0.0182

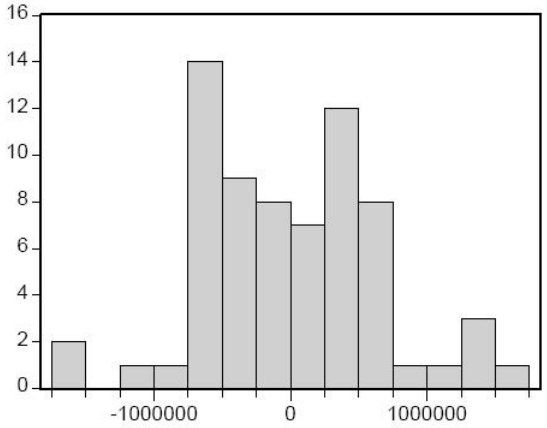
R-squared	0.527317	Mean dependent var	8375.647
Adjusted R-squared	0.505160	S.D. dependent var	951018.6
S.E. of regression	668992.7	Akaike info criterion	29.72196
Sum squared resid	2.86E+13	Schwarz criterion	29.85252
Log likelihood	-1006.547	Durbin-Watson stat	1.990783

Inverted AR Roots	.61-.37i	.61+.37i	-.11+.75i	-.11-.75i

Jedná sa o autoregresný model AR(4). Model je stacionárny a môžeme prejsť k verifikácii.

Testovanie modelu

Rovnakým postupom ako pri ostatných obdobiach aj tu overíme, či sa v modeli nachádza biely šum pomocou Q-štatistiky. Najskôr si ale ukážeme, či sú reziduá z normálneho rozdelenia.



Series: Residuals	
Sample 12/24/2008 3/01/2009	
Observations 68	
Mean	0.002277
Median	-24657.75
Maximum	1696299.
Minimum	-1559817.
Std. Dev.	653808.7
Skewness	0.199422
Kurtosis	3.135526
Jarque-Bera	0.502758
Probability	0.777728

P-hodnota je približne 0.78, tj. vyššia ako 5%, čo potvrdzuje normálne rozdelenie a prejdeme ku Q-testom.

Correlogram of Residuals

Date: 02/06/11 Time: 11:35
Sample: 12/24/2008 3/01/2009
Included observations: 68
Q-statistic probabilities adjusted for 4 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.002	0.002	0.0003	
		2	0.039	0.039	0.1099	
		3	-0.058	-0.058	0.3526	
		4	-0.002	-0.003	0.3528	
		5	-0.011	-0.007	0.3623	0.547
		6	-0.050	-0.054	0.5563	0.757
		7	0.257	0.260	5.7244	0.126
		8	-0.102	-0.113	6.5469	0.162
		9	0.070	0.055	6.9466	0.225
		10	-0.209	-0.193	10.536	0.104
		11	-0.010	-0.012	10.545	0.160
		12	-0.132	-0.126	12.015	0.151
		13	-0.009	0.013	12.022	0.212
		14	0.165	0.107	14.422	0.155
		15	0.017	0.061	14.447	0.209
		16	-0.089	-0.181	15.177	0.232
		17	-0.167	-0.047	17.779	0.166
		18	-0.052	-0.117	18.033	0.205
		19	-0.106	-0.012	19.125	0.208
		20	-0.032	-0.094	19.225	0.257
		21	0.065	0.037	19.653	0.292
		22	0.187	0.157	23.264	0.181
		23	0.007	0.037	23.269	0.226
		24	-0.072	-0.043	23.831	0.250
		25	-0.136	-0.124	25.867	0.212
		26	-0.166	-0.211	28.992	0.145
		27	-0.064	-0.086	29.467	0.165
		28	0.017	-0.082	29.504	0.202

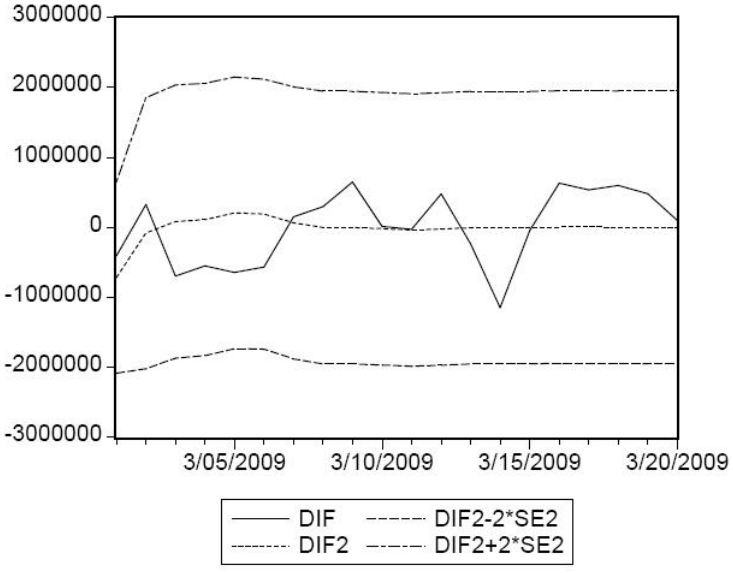
A ešte overíme Q-štatistiku pre druhé mocniny reziduí.

Correlogram of Residuals Squared

Date: 02/06/11 Time: 11:35
Sample: 12/24/2008 3/01/2009
Included observations: 68
Q-statistic probabilities adjusted for 4 ARMA term(s)

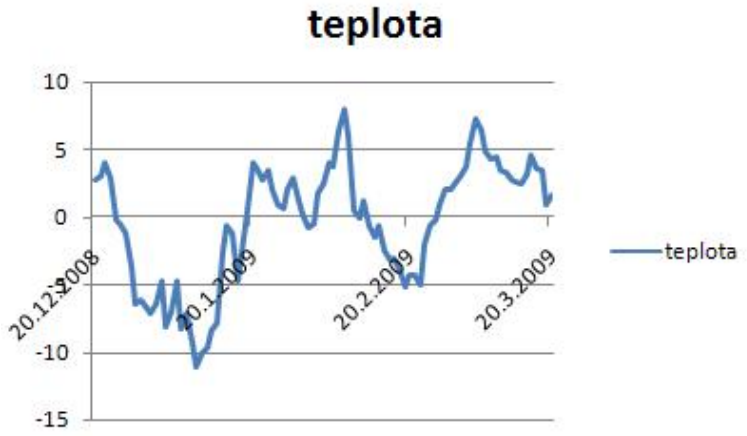
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.068	-0.068	0.3321	
		2	0.112	0.108	1.2358	
		3	0.061	0.076	1.5083	
		4	-0.119	-0.125	2.5689	
		5	-0.056	-0.090	2.8066	0.094
		6	-0.014	0.001	2.8227	0.244
		7	-0.005	0.030	2.8249	0.419
		8	-0.042	-0.046	2.9655	0.564
		9	0.141	0.121	4.5744	0.470
		10	-0.160	-0.148	6.6689	0.353
		11	-0.042	-0.094	6.8170	0.448
		12	-0.166	-0.178	9.1631	0.329
		13	-0.064	-0.022	9.5134	0.391
		14	-0.005	0.022	9.5154	0.484
		15	-0.133	-0.145	11.094	0.435
		16	-0.042	-0.134	11.256	0.507
		17	0.083	0.081	11.903	0.536
		18	-0.119	-0.132	13.256	0.506
		19	-0.037	-0.085	13.388	0.572
		20	-0.052	-0.123	13.651	0.625
		21	-0.132	-0.122	15.416	0.566
		22	-0.032	-0.139	15.522	0.626
		23	0.152	0.093	17.976	0.524
		24	-0.051	-0.072	18.252	0.571
		25	0.160	0.054	21.073	0.454
		26	0.150	0.029	23.625	0.367
		27	-0.098	-0.130	24.737	0.364
		28	0.019	-0.124	24.781	0.418

Všetky hodnoty p-value sú väčšie ako 5% a teda sa tam biely šum nachádza a môžeme prejsť k overovaniu správnosti daného modelu pomocou predikcií.

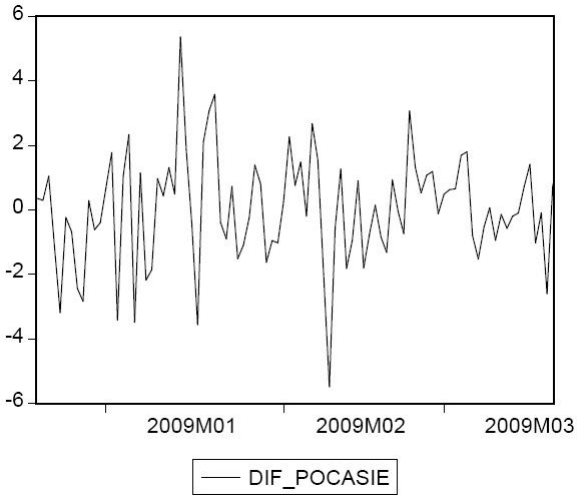


Vplyv teploty

Pri tomto období očakávame, že nám pridanie teploty pomôže zlepšiť model, keďže v zimných mesiacoch je teplota ovzdušia najnižšia. Graf pre teplotu v zime:



Teplota ovzdušia nadobúda hodnoty približne od -12 po 7 stupňov. Korelogram pre teplotu vyzerá podobne ako pre plyn, tzn. rad je nestacionárny a stacionarizujeme ho pomocou prvých diferencií. Zdifrencovaný rad vyzerá nasledovne:



Teraz do modelu zahrnieme vplyv teploty.

Dependent Variable: DIF
Method: Least Squares
Date: 04/10/11 Time: 01:03
Sample (adjusted): 12/24/2008 3/01/2009
Included observations: 68 after adjustments
Convergence achieved after 18 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DIF TEPLOTA	-74141.55	32739.76	-2.264572	0.0270
AR(1)	0.943658	0.121506	7.766321	0.0000
AR(2)	-0.740926	0.159221	-4.653449	0.0000
AR(3)	0.508100	0.158587	3.203916	0.0021
AR(4)	-0.275009	0.124968	-2.200634	0.0314

R-squared	0.560225	Mean dependent var	8375.647
Adjusted R-squared	0.532302	S.D. dependent var	951018.6
S.E. of regression	650386.6	Akaike info criterion	29.67921
Sum squared resid	2.66E+13	Schwarz criterion	29.84241
Log likelihood	-1004.093	Durbin-Watson stat	1.998379

Inverted AR Roots	.60-.39i	.60+.39i	-.13+.72i	-.13-.72i

Jedná sa o autoregresný model AR(4). Hodnota Prob. pri teplote je nižšia ako 5%, takže vplyv teploty je dôležitý. Ukážeme si rovnicový tvar a overenie pomocou grafu pre predikcie.

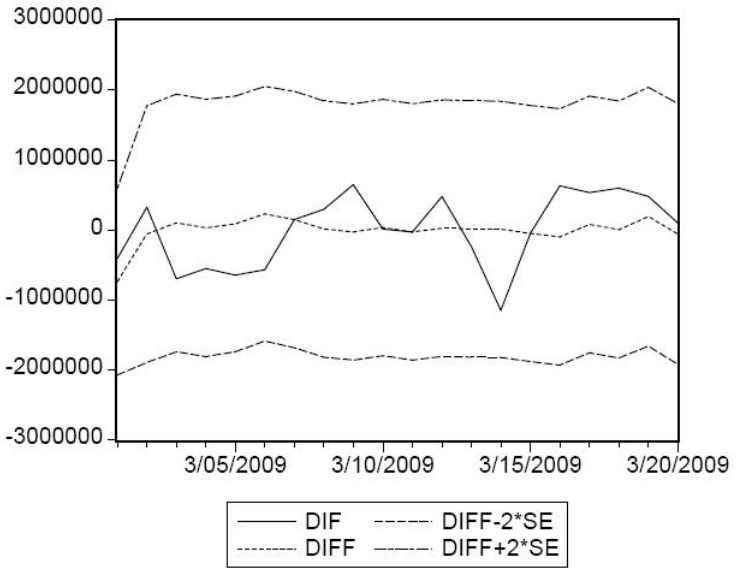
Rovnicový tvar

Rovnica:

$$C = \frac{74141.55}{1 - 0.943658 + 0.740926 - 0.508100 + 0.275009}$$

$$\Delta y_t = -74141.55\Delta x_t + C + 0.943658\Delta y_{t-1} - 0.740926\Delta y_{t-2} + 0.508100\Delta y_{t-3} - 0.275009\Delta y_{t-4} + \epsilon$$

Overíme správnosť modelu na dvadsiatich dátach od $1/3/2009$ do $20/3/2009$ podobne ako pri predchádzajúcich modeloch.



Graf sa zlepšil, avšak nie až tak výrazne ako to bolo po pridaní teploty v jarnom období, čo sme aj mohli očakávať, keďže v jarných resp. aj v jesenných mesiacoch sú najväčšie teplotné zmeny, čo sa odráža aj na spotrebe plynu na Slovenskom území.

Záver

Cieľom tejto práce bolo na začiatku priblížiť čitateľovi základné teoretické poznatky o časových radoch, ich niektorých vlastnostiach a špecializovať sa na ARMA modely, ktoré sa využívajú na tvorbu krátkodobých predikcií predovšetkým v ekonómii, a následne aplikovať teóriu na reálnych dátach. Časový rad zachytával spotrebu zemného plynu na území Slovenskej republiky za obdobie od 20/12/2008 do 20/12/2009. Dáta boli denného charakteru (vrátane víkendov).

Pri tvorbe prvého modelu sme sa pokúsili pracovať s celým časovým radom, ale kvôli nepriaznivým výsledkom niektorých testov sme sa nakoniec rozhodli tento rad rozdeliť na 4 časti podľa ročných období. Metódou "pokus, omyl" sme po vyskúšaní viacerých modelov nakoniec vybrali tie, ktoré najlepšie opisujú priebeh radu. Nakoniec sme teda mali štyri finálne modely pre spotrebu plynu.

V snahe vylepšiť dané modely sme si zobrali ďalší rad, ktorý zachytáva vývoj teploty ovzdušia v rovnakom období na rovnakom území. Tento rad sme tam pridali ako vysvetľujúcu premennú. Vo väčšine prípadov sa potvrdilo, že teplota ovzdušia výrazne ovplyvňuje spotrebu plynu. Jedinou výnimkou bolo obdobie keď teplota dosahovala stále hodnoty viac ako 16 stupňov.

V snahe ešte viac vylepšiť modely sme sa pokúsili z radov odstrániť víkendy, a pracovať len s pracovnými dňami, čo ale neviedlo k zlepšeniu výsledkov a preto sme ich do práce nezahrnuli. Ďalším riešením bolo rozdeliť časový rad o spotrebe plynu na veľkých odberateľov a domácnosti, čím by sme dostali dva samostatné rady, ale ani toto riešenie neviedlo k celkovému zlepšeniu.

Jedným z ďalších cieľov bolo navrhnúť nové druhy premenných o charaktere počasia, ktoré by mohli pomôcť spresniť prognózy SPP. Po dlhom premýšľaní sme usúdili, že žiadny iný vplyv, okrem teploty ovzdušia nie je dôležitejší pre spotrebu plynu, pretože všetky meteorologické podmienky, ako je dážď, sneh a iné sa v konečnom dôsledku odzrkadlia na teplote ovzdušia.

Zoznam použitej literatúry

[1] Tomáš Cipra *Finanční ekonometrie*, Ekopress,s.r.o., 2008, ISBN 978-80-86929-43-9

[2] Marček, D.,Pančíková, L.,Marček, M. *EKONOMETRIA A SOFT COMPUTING*, EDIS, 2008, ISBN 978-80-8070-746-0

[3] Tomáš Cipra *Analýza časových řad s aplikacemi v ekonomii*, 1986, ISBN 04-012-86

[4] Ing. Miroslav Kľúčik, Ing. Jana Juriová *Základy programovacieho jazyka EViews a ich aplikácia na analýzy, prognózy a rýchle odhady vývoja makroekonomických ukazovateľov, 2.časť*, Infostat, 2011, ISBN 978-80-89398-19-5

[5] Marius Ooms *Introduction Eviews for Orientation course Econometrics*, Amsterdam , 2004

[6] Jozef Arlt, Markéta Arltová, Eva Rublíková *Analýza ekonomických časových řad s příklady*, 2002

[7] Doc. Ing. Emil Pelikán, CSc. *Predikční metody*, SOFTECON, 2005

[8] Kozák, J.,Hindls, R., Artl, J.*Úvod do analýzy ekonomických časových řad*, VŠE v Praze , 1994

[9] Hušek, R.*Ekonometrická analýza*, Ekopress , 1999, ISBN 80-86119-19

[10] Box, G.E.P., Jenkins, G.M.*Time Series Analysis, Forecasting and Controls*, Holden Day, 1976

[11] Hanke, J.E.*Business Forecasting*, New Jersey , 2001

[12] Kozák, J.,Seger, J.*Jednoduché statistické metody v prognostice*, Praha, SNTL , 1975

[13] Chung-Ming Kuan *Concepts and Methods (2nd edition)*, Huatai Publisher , Taipei, 2004, http://idv.sinica.edu.tw/ckuan/pdf/et01/ch8.pdf

[14] Cromwell, J.B.,Labys, W.C.,Terraza, M.*Univariate tests for time series models*, SAGE , 1994

Internetové zdroje

http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09/cv1.html
http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09/cv2.html
http://www.iam.fmph.uniba.sk/institute/stehlikova/cr09/cv4.html
http://core.ecu.edu/econ/rothmanp/EViewsUG.pdf