LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTISKĀS ANALĪZES NODAĻA

PROGNOZĒŠANA AR KLĀSTERIZĀCIJAS METODI

DIPLOMDARBS

Autors: **Anete Rubine**

Stud. apl. ar08242

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2013

Anotācija

Laikrindu nākotnes vērtību prognozēšana ir viena no nozīmīgākajām problēmām, ar kuru saskaras datu analītiķi daudzās nozarēs. Darbā tiek analizēta nesen ieviestā klāsterizācijas metode [1] datu nākotnes vērtību prognozēšanai. Mērķis ir parādīt, ka šī metode sniedz labus rezultātus, salīdzināt to ar populāro ARIMA modeli. Darbā aplūkotā metode spēj prognozēt arī datu izlecējus, kas ir svarīgi praktiskās datu problēmās.

Atslēgas vārdi: ARIMA, Prognozēšana, Silhouette indekss, Klāsteru analīze, Laikrindu analīze.

Abstract

Forecasting future values of time sries is on of the most significant problems which data analysts face in many industries. In this work we analyze the Pattern Sequence - based Forecasting method [1]. Our aim is to show that this method gives good results and to compare it with the popular ARIMA model. The analyzed method is able to predict sudden changes in data, as outliers which is important in practical applications.

Keywords: ARIMA, Forecasting, Silhouette index, Cluster analysis, Time series.

APZĪMĒJUMU SARAKSTS

Darbā izmantotie apzīmējumi:

- PSF algoritms - prognozēšana ar klāsterizācijas metodi.
- Silh - Silhouette indekss.
- ARIMA - autoregresīvais integrētais slīdoša vidēja process.
- MAPE - vidējā absolūtā procentuālā kļūda.
- MSE - prognozes vidējā kvadrātiskā kļūda.

Saturs

Ievads

Laikrindu analizēšana un nākotnes vērtību prognozēšana ir viena no nozīmīgākajām problēmām, ar kuru saskaras datu analītiķi daudzās nozarēs, sākot no finansēm un ekonomikas līdz ražošanas darbības vadīšanai vai telekomunikācijām. *Prognozēšana* ir kāda nākotnes rezultāta paredzēšana, un tā problēmas tiek klasificētas kā:

• īslaicīgās prognozēšanas problēmas ietver īsa laika perioda nākotnes vērtību prognozēšanu (dienas, nedēļas, mēneši);

• vidēji ilgās prognozēšanas problēmas ietver laika periodu prognozēšanu no viena līdz diviem gadiem;

• ilglaicīgās prognozēšanas problēmas ietver laika periodu prognozēšanu, kas var pārsniegt divus gadus.

Laikrindu dati var tikt definēti pēc interesējošā mainīgā novērojumu hronoloģiskas secības. Prognozēšana ir svarīga ne tikai valsts iestādēm un lieliem uzņēmumiem, bet arī sabiedrībai, jo precīzākas prognozes tiek veiktas, jo cilvēkam ir lielāka iespēja saplānot savu personīgo budžetu. Piemēram, ir svarīgi laicīgi zināt par gāzes vai elektroenerģijas cenu pieaugumu. Savukārt uzņēmumiem ir svarīgi zināt pieprasījumu, jo no tā būs atkarīga arī cena, tāpec tiek risināta problēma par prognožu uzlabošanu, kas spētu precīzāk prognozēt visas izmaiņas - negaidīi augstu vai zemu pieprasījumu.

Diplomdarba mērķis ir veikt elektroenerģijas pieprasījuma un dabas gāzes cenas prognozēšanu. Mūsdienās ir daudzas metodes, kā to veikt, populārākā no tādām ir prognozēšana, izmantojot ARIMA modeļus. Šajā darbā tiks aplūkots nesen ieviestais PSF algoritmu (*Pattern Sequence - based Forecasting*) [1], kas ir interesants ar to, ka spēj prognozēt arī datu "izlecēju" vērtības. PSF algoritmam ir sava īpatnība, tas nestrādā ar reāliem datiem, bet gan ar to apzīmētajiem datiem, kurus mēs iegūstam klāsterizācijas ceļā. PSF algoritms ir turpinājums prognozēšanas tehnikai, kurā agrāk tikusi izmantota "tuvāko kaimiņu" metode [2]. Nobeigumā tiks salīdzinātas prognozes balstītas uz PSF algoritmu un ARIMA modeļiem.

Vispirms iepazīsimies ar darba teorētisko pusi:

• Pirmā nodaļa - klāsterizācija - pirmais solis ir silhouette indeksa aprēķināšana, kas norādīs cik klāsteros dati jāsadala, otrais solis - datu normalizācija, trešais solis - datu klāsterizācija, pēc kuras dati tiks aizstādi ar apzīmēti ar attiecīgo indeksu.

• Otrā nodaļa - PSF algoritms - tiek definēts algoritms, kuru izmantojot tiks veikta datu

nākotnes vērtību prognozēšana.

• Trešā nodaļa - ARIMA modelis - tiek aprakstīta automātiska funkcija, kas automātiski nosaka datu ARIMA modeli un veic automātisku datu nākotnes vērtību prognozēšanu.

• Ceturtā nodaļa - labākā modeļa noteikšana - prognozes vidējās absolūtās kļūdas un prognozes vidējās kvadrātiskās kļūdas aprēķināšana.

• Piektātā nodaļa - praktiskā daļa - tiek apskatīti divi datu piemēri par elektroenerģijas pieprasījumu Austrālijas Nacionālajā elektroenerģijas uzņēmumā [3] un dabas gāzes cenu kompānijai "DOWJONES" Ziemeļamerikā [4], kuros praktiski tiek pielietots PSF algoritms un ARIMA modelis datu nākotnes vērtību prognozēšanai.

1. Klāsterizācija

Definīcija 1. Klāsterizācija ir process, kura laikā laikrindas vērtības tiek aizstātas ar klāsteriem- apzīmētām vērtībām.

Definīcija 2. Klāsteris ir datu kopa, kas sevī apvieno līdzīgos datus no laikrindas, visi klāsteri kopā ir sākotnējā laikrinda.
Sākumā definēsim *Silhouette indeksu* pēc formulas:

$$silh(i) = \frac{a(i) - b(i)}{\max\{a(i), b(i)\}} = \begin{cases} 1 - \frac{a_i}{b_i}, a(i) < b(i) \\ 0, a(i) = b(i) \\ \frac{a_i}{b_i} - 1, a(i) > b(i) \end{cases},$$

kur vidējais attālums objektam i, $(i \in A)$, starp pārējiem objektiem, kas atrodas kopā A, apzīmē $a(i)$, un vidējais attālums objektam i starp visiem pārējiem objektiem, kas atrodas klāsterī $C \neq A$ tiek apzīmēts ar $d(i, C)$. Visiem klāsteriem, kam $C \neq A$, $d(i, C)$ tiek aprēķinātas vērtības un mazākā tiek izvēlēta sekojošā veidā:

$$b(i) = \min_{C \neq A} d(i, C), i \in A.$$

Vērtība $b(i)$ parāda atšķirību objektam i no tā tuvākā kaimiņa klāstera. $Silh(i)$ vērtība var būt no -1 līdz $+1$, kur $+1$ un -1 parāda vai objekts i pieder vai nepieder attiecīgajam klāsterim. Ja silhouette indeksa i vērtība pieder klāsterim A un tā vērtība ir tuva nullei, tas nozīmē, ka objekts i var piederēt tuvākajam A kaimiņa klāsterim. Ja klāsters A ir kopa tikai ar vienu elementu, tad silhouette indekss objektam i nav definēta, šādā gadījumā tā vērtība ir vienāda ar nulli. Mērķa funkcija ir vidējā $silh(i)$ vērtība un vislabākā klāsterizācija ir sasniegta, kad $silh(i)$ tiek maksimizēta. Programmā R, lai šo indeksu aprēkinātu, tiek lietota *manhattan* metode, kas nodrošina to, ka distance starp klāstera centru un datu punktiem tiek aprēķināta kā koordinašu attālumu absolūto vērtību summa.

$$d_{jk} = sum(abs(x_{ij} - x_{ik})).$$

Tālāk tiek veikta datu normalizācija pēc formulas:

$$x_j \leftarrow \frac{x_j}{\frac{1}{N}\sum_{i=1}^{N} x_i},$$

kur x_j piemēram, ir pieprasījums j - tajā dienas stundā un N ir vienāds ar 24, jo katra vērtība parāda vienas stundas izmaiņas vai x_j ir pieprasījums j - tajā gada mēnesī un N ir vienāds ar 12, jo katra vērtība parāda viena mēneša izmaiņas. Tagad varam veikt klāsterizāciju. Šajā darbā tiek izmantots K - vērtības algoritms, kas ir optimāls pieprasījuma datu kopu klasificēšanai [5]. K-vērtības algoritma procedūra ir vienkāršs un ērts veids, lai klasificētu attiecīgo datu kopumu, izmantojot noteiktu skaitu klāsteru (pieņemsim k klāsterus), kas ir fiksēts lielums. Galvenā ideja ir definēt k lielumu, kas ir minimālais attālums līdz klāstera vidum (katram klāsterim tie ir atšķirīgi). Šī attāluma atrašanai tika izmantota silhouette funkcija. K - vērtības algoritma mērķis ir minimizēt mērķa funkciju, kas šajā gadījumā ir kvadrātisko kļūdu funkcija. Mērķa funkcija ir:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2,$$

kur $||x_i^{(j)} - c_j||^2$ ir izvēlēto attālumu mērs starp datu punktu $x_j^{(i)}$ un klāstera centru c_j, n - attālumu rādītājs datu punktiem līdz attiecīgajiem klāsteru centriem. K - vērtību algoritms sadala visus datus k grupās, katrā grupā ir atlasīti tie dati, kas atbilst attālumam no klāstera centra līdz datu punktam. Tādā veidā mēs apzīmējam datus ar attiecīgo k indeksu [6].

2. PSF algoritms

PSF algoritms tiek uzdots šādi:

PSF()

$ES_d \leftarrow \{\}$

$\widehat{X}(d) \leftarrow 0$

katrai dienai $d \in T$

$S_W^{d-1} \leftarrow [L_{d-W}, L_{d-W+1}, ..., L_{d-2}, L_{d-1}]$

katram j**, kurš** $X(j) \in D$

$S_W^j \leftarrow [L_{j-W+1}, L_{j-W+2}, ..., L_{j-1}, L_j]$

ja $(S_W^j = S_W^{d-1})$

$ES_d \leftarrow ES_d \cup j$

katram $j \in ES_d$

$\widehat{X}(d) \leftarrow \widehat{X}(d) + X(j+1)$

$\widetilde{X}(d) \leftarrow \widetilde{X}(d)/size(ES_d)$

$D \leftarrow D \triangleright \widehat{X}(d)$

$[L_1, L_2, ..., L_{d-1}, L_d] \leftarrow$klāsterizācija$(D, K)$

$d \leftarrow d + 1$

atgriež $\widehat{X}(d)$ visām T dienām.

Algoritmā **D** ir datu kopa, **K** klāsteru skaits, $[L_1, L_2, ..., L_{d-1}, L_{d-1}]$ apzīmētu datu kopa, **W** loga garums un **T** izmēģinājuma datu kopa (manā gadījumā $D = T$). Lai algoritms tiktu veiksmīgi pielietots, jāzin laikrindas vēsturiskās vērtības līdz dienai $d-1$ vai laikrindas vēsturiskās vērtības līdz mēnesim $d-1$. Prognozēšanas mērķis ir prognozēt nākamās 30 stundu elektroenerģijas pieprasījuma vērtības un nākamo 20 mēnešu dabas gāzes cenas. Pieņemsim, ka $X(i) \in R^t$ ir vektors, kur t sastāv no 24 stundu elektroenerģijas pieprasījuma vērtībām kādai konkrētai dienai i vai sastāv no 12 mēnešu dabas gāzes cenas vērtībām kādam konkrētam gadam i

$$X(i) = [x_1, x_2, ..., x_t].$$

Pieņemsim, ka $L_i \in \{1, ..., K\}$ apzīmējumi pieprasījumam dienai i vai cenai gadam i, kas tiek iegūti ar klāsterizācijas metodi, kur K ir klāsteru skaits. Pieņemsim, ka S_W^i ir apzīmēto pieprasījumu rezultāts W secīgām dienām vai gadiem, kas ir atpakaļejošs, sākot

ar dienu vai gadu i. Pieņemsim, ka $X(i) \in R^t$ ir vektors, kur t sastāv no 24 stundu elektroenerģijas pieprasījuma vērtībām kādai konkrētai dienai i vai sastāv no 12 mēnešu dabas gāzes cenas vērtībām kādam konkrētam gadam i

$$S_W^i = [L_{i-W+1}, L_{i-W+2}, ..., L_{i-1}, L_i],$$

kur loga garums W ir parametrs, kurš tiek aprēķināts sekojoši, tiek izmantoti klāsterizācijas ceļā apzīmētie dati. Atrast vērtību W var minimizējot funkciju

$$\sum_{d \in TS} ||\widehat{X}(d) - X(d)||,$$

kur $\widehat{X}(d)$ ir prognozētās vērtības dienai vai gadam d, atsaucoties uz PSF algoritmu $X(d)$ ir reālās vērtības un TS ir datu kopa. PSF algoritms pieprasījuma vai cenas prognozēšanai dienai vai gadam d vispirms meklē apzīmētos datus datu kopā, kas pilnīgi vienādi ar S_W^{d-1}, ja vienādā apakškopa ES_d ir definēta kā

$$ES_d = \{j, \text{ tādu ka} S_W^j = S_W^{d-1}\}.$$

Gadījumā, kad šāda vienādība netiek atrasta, algoritms meklē apakškopu, kas ir pilnīgi vienāda ar S_{W-1}^{d-1}. Tādā veidā loga garums, kas sastāv no apzīmētajiem datiem, samazinās par vienu vienību. Šī stratēģija garantē, ka tiks atrasta vismaz viena vienādība, kad W būs pilnīgi vienāds ar viens.

Pēc PSF algoritma, 24 stundu prognoze pieprasījuma laika rindu vērtībām dienai d un 12 mēnešu prognoze cenu laika rindu vērtībām gadam d tiek prognozētas ar vidējo vērtību dienām/gadiem, kas seko pēc ES_d,

$$\widehat{X}(d) = \frac{\sum_{j \in ES_d} X(j+1)}{size(ES_d)},$$

kur $size(ES_d)$ ir elementu skaits, kas pieder kopai ES_d.

Gadījumā, kad jāveic vidēja vai gara laika perioda prognoze, visas darbības tiek attiecinātas uz visu datu kopu, un klāsterizācijas process tiek atkārtots ar paplašinātām datu kopām, līdz prognozēšana ir izpildīta [1].

3. ARIMA modelis

Šajā nodaļā definēsim ARIMA (autoregresīvais integrētais slīdoša vidēja process) modeli, kā arī tiks apskatīti tā galvenie raksturlielumi. ARIMA ir statistiskās analīzes modelis, ko izmanto laikrindas datu labākai izprašanai vai laikrindas datu nākotnes tendendenču prognozēšanai.

Definīcija 3. Par laikrindu sauc novērojumu virkni $x_{t_1}, x_{t_2}, ..., x_{t_N}$, kuru iegūst, novērojot gadījuma lielumu x_t secīgos laika momentos $t_1, t_2, ..., t_N$.

Definīcija 4. Gadījuma procesu $\{x_t\}$, $t = 0, 1, \ldots$ sauc par jaukto (p, q) kārtas autoregresīvo slīdošā vidējā (ARMA(p, q)) proscesu, ja tas apmierina vienādojumu

$$x_t = a_0 + a_1 x_{t-1} + \cdots + a_p x_{t-p} + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q} + \epsilon_t, \tag{3.1}$$

kur ϵ_t ir baltais troknis ar $E\epsilon_t = 0$, $D\epsilon_t = \sigma^2$, a un b ir reāli skaitļi.

Definīcija 5. Saka, ka gadījuma process $\{x_t\}$ ir(p, d, q) - kārtas autoregresīvais integrētais slīdoša vidēja process ARIMA(p, d, q), ja $w_t = \Delta^d x_t = (1 - L)^d x_t$ ir ARMA (p, q) process.

Definīcija 6. Gadījuma procesu $\{x_t\}_{t=-\infty}^{\infty}$ sauc par q - tās kārtas slīdošā vidējā procesu (MA(q)), ja tas apmierina vienādojumu

$$x_t = \mu + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + \cdots + \epsilon_t, \tag{3.2}$$

kur ϵ_t baltais troksnis ar $E\epsilon_t = 0$ un $D\epsilon_t = \sigma^2$.

Definīcija 7. Par *p-tās kārtas autoregrasīvo procesu* AR(p) sauc gadījuma procesu $\{x_t\}_{t=1}^{\infty}$, kas apmierina vienādojumu

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \epsilon_t, \tag{3.3}$$

kur ϵ_t ir baltais troksnis, $E\epsilon_t = 0$, $D\epsilon_t = \sigma^2$, $a_p \neq 0$.

Definīcija 8. Gadījuma procesu $\{X_t\}_{t=-\infty}^{\infty}$ sauc par stacionāru stingrā nozīmē, ja tā galīgdimensionālie sadalījumi paliek nemainīgi pie patvaļīgas laika nobīdes, t.i., ja katram $n \geq 1$, katriem $t_1, t_2, ..., t_n$ un katram h n - dimensionālu gadījumu vektoru $(x_{t_1}, x_{t_2}, ..., x_{t_n})$ un $(x_{t_1+h}, x_{t_2+h}, ..., x_{t_n+h})$ sadalījuma funkcijas sakrīt

$$F_{t_1+h,t_2+h,...,t_n+h}(x_1, x_2, ..., x_n) = F_{t_1,t_2,...,t_n}(x_1, x_2, ..., x_n). \tag{3.4}$$

Definīcija 9. Gaījuma procesus, kuri apmierina nosacījumus

1. visiem t vidējās vērtības ir vienādas;
2. $cov(x_t, x_{t+\tau})$ ir atkaīgi tikai no laika atstarpes $\tau : cov(x_t, x_{t+\tau}) = \gamma(\tau)$;

sauc par "vājā" (jeb "plašā") nozīmē stacionāriem gadījuma procesiem.

Definīcija 10. Stacionāru procesu, kuram $\gamma_0 = \sigma^2 < \infty$ un $\gamma_\tau = 0(\tau \neq 0)$, sauc par balto troksni (plašā (vājā) nozīme). Lai gadījuma lielumu virkne $\{x_t\}_{t=-\infty}^{\infty}$ būtu baltais troksnis ir pietiekami, ka visiem virknes gadījuma lielumiem ir vienādas vidējās vērtības un dispersijas un ka tie ir pa pāriem nekorelēti. Bieži tiek prasīts, lai proces vidējā vērtība būtu nulle.

Definīcija 11. Aprakstot laikrindu, tajā parasti izdala četras komponentes:

$$x_t = f(T_t, S_t, C_t, I_t), \tag{3.5}$$

kur Tr_t - trends, S_t - sezonalitāte, C_t - cikliskā komponente, I_t - neregulāra komponente.

Definīcija 12. Par laikrindas trendu T_t pie $t = 1, ..., n$ sauc datu sistemātiskās izmaiņas, parasti to palielināšanās vai samazināšanās laikā. Šī komponente atspoguļo vispārīgo ilglaicīgo (ilglaicīgajā perspektīvā) tendenci analizējamas pazīmes x_t izmaiņās. Parasti šī tendence tiek aprakstīta ar monotonas negadījuma funkcijas $T_t = f(t, \Delta)$ palīdzību, kur Δ - parametru vektors. Šo funkciju sauc par trenda funkciju jeb vienkārši trendu.

Definīcija 13. Par laikrindas sezonalitāti sauc datu sistemātiskas, periodiskas izmaiņas laikā (gada ietvaros). Vispārīgi, ja laikrindai piemīt sezonalitāte ar periodu s, tas līdzīgas laikrindas izmaiņas atkārtojas ik pēc s bāzes laika intervāliem.

Definīcija 14. Laikrindas cikliskā komponente C_t apraksta gludus kvazi-periodiskas datu svārstības ap trendu, kuru periods ir lielāks par 1 gadu. Tā parasti saistīta ar biznesa vai ekonomikas stāvokļiem. Par cikla periodu sauc cikla garumu. To var mērīt no viena maksimuma punkta līdz nākamajam maksimumam, vai no viena minimuma punkta līdz nākamajam minimumam.

Definīcija 15. Neregulārā komponente I_t apraksta nedeterminētas svārstības, kuras var sadalīt divās grupās: pēkņas strukturālas datu izmaiņas, kuras var būt izraisītas ar karu vai ekoloģisko katastrofu un gadījumu svārstības, kuras rodas pateicoties daudzu relatīvi vāju maznozīmīgu faktoru iedarbībai[7].

Piemērotāko ARIMA modeli pielāgosim ar automātiski iebūvētu funkciju programmā R *auto.arima*:

auto.arima(x, d = NA, D = NA, max.p = 5, max.q = 5, max.P = 2, max.Q = 2, max.order = 5, start.p = 2, start.q = 2, start.P = 1, start.Q = 1, stationary = FALSE, ic = c(aic, aicc, bic), stepwise = TRUE, trace = FALSE, approximation = lenght(x)>100 | frequency(x)>12, xreg = NULL, test = c(kpss, adf, pp), allowdrift = TRUE)

x - laikrinda.

d - pirmā diference, gadījumā ja šādas vērtības nav, to izvēlas pēc KPSS testa.

D - sezonālā diference, gadījumā ja šādas vērtības nav, to izvēlas pēc CH testa.

max.p - maksimālā p vērtība.

max.q - maksimālā q vērtība.

max.P - maksimālā P vērtība.

max.Q - maksimālā Q vērtība.

max.order - maksimālā $p+q+P+Q$ vērtība, ja modeļa izvēle nav veikta pēc pakāpienveida procedūras.

start.p - sākuma p vērtība pakāpienveida procedūrā.

start.q - sākuma q vērtība pakāpienveida procedūrā.

start.P - sākuma P vērtība pakāpienveida procedūrā.

start.Q - sākuma Q vērtība pakāpienveida procedūrā.

stationary - ja patiess, tad ierobežo stacionāra modeļa meklēšanu.

ic - informācijas kritērijs, kas tiek izmantots modeļa izvēlē.

stepwise - ja patiess, ātrāk spēs veikt pakāpievneida procedūru, pretējā gadījumā tas meklē pa visiem modeļiem; sezonāliem modeļiem tā var būt gara procedūra.

trace - ja patiess, visu iespējamie ARIMA modeļi tiks apskatīti.

approximation - pielāgošana modelim.

xreg - tāds pats skaitlis, kā x rindu skaits.

test - unit root tests- pārbauda vai laikrinda ir stacionāra vai nav.

allowdrift - ja patiess, modeļi ar novirzītiem nosacījumiem tiek apskatīti.

Šī funkcija izvēlas labāko ARIMA modeli saskaņā ar kādu no AIC, AICc vai BIC vērtībām. Funkcija veic meklēšanu pār iespējamiem modeļiem noteikto ierobežojumu ietvaros [8].

Definīcija 16. AIC - Akaike informācijas kritērijas vispārīgi ir

$$AIC = 2k - 2ln(L),$$

kur k ir parametru skaits statistiskajā modelī, L ir statistikā modeļa parametru funkcijas maksimālā vērtība. No visiem modeļiem, kas tiek attiecināti uz konkrētu laikrindu, tiek izvēlēts tas modelis, kuram AIC vērtība ir vismazākā.

Definīcija 17. AICc ir AIC ar korekciju ierobežotajam parauga lielumam:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1},$$

kur n norāda parauga lielumu, tas ir, AICc ir AIC ar augstāku jūtīgumu pret papildus parametriem. AICc kritērijs tiek izmantots tad, kad n ir mazs vai k ir liels, bet AICc konverģē uz AIC, kad n kļūst liels.

Definīcija 18. BIC - Bejiesa informācijas kritērijs vai Švarca kritērijs (SBC, SBIC) ir statistisks kritērijs piemērotākā modeļa izvēlei kādai konkrētai laikrindai, tas ir radniecisks Akaike informācijas kritērijasm (AIC):

$$BIC = n \cdot ln(\widehat{\sigma_e^2}) + k \cdot ln(n),$$

kur $\widehat{\sigma_e^2}$ ir dispersijas kļūda. Šajā gadījumā dispersijas kļūda tiek definēta šādi:

$$\widehat{\sigma_e^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2,$$

kur x - novērotie dati, n - datu daudzums kopā x, k - parametru skaits statistiskajā modelī, L - statistikā modeļa parametru funkcijas maksimālā vērtība. No visiem modeļiem, kas tiek attiecināti uz konkrētu laikrindu, tiek izvēlēts tas modelis, kuram BIC vērtība ir vismazākā.

Definīcija 19. ARIMA modeļa standarta pieraksts ARIMA(p, d, q) kur

- p - autoregresīvā polinoma kārta;
- d - integrācijas kārta (diferenču operatora kārta);
- q - slīdošā vidējā polinoma kārta;

Definīcija 20. ARIMA modeļa standarta pieraksts ARIMA$(p, d, q)(P, D, Q)_s$ kur

- p - autoregresīvā polinoma kārta;
- d - integrācijas kārta (diferenču operatora kārta);
- q - slīdošā vidējā polinoma kārta;
- P - sezonālā autoregresīvā polinoma kārta;
- D - sezonālā diferenču operatora kārta;
- Q - sezonālā vidējā polinoma kārta;
- s - sezonalitātes komponente.

Automātiskā funkcija programmā R *auto.arima*, lai atrastu d un D lieto vienības saknes testu (*Unit root test*). Tomēr vairumam šo testu nulles hipotēze ir tāda, ka vienības sakne eksistē, tas rada vairāk lielas nekā mazas diferences rezultātā. Dickey - Fuller testā tiek pieņemts, ka vienības sakne ir pie laga 1. Tāpēc šajā funkcijā tiek izmantots paplašinātais Dickey - Fuller (*Augmented Dickey ˘ Fuller test − ADF*) tests, kas atšķiras ar to, ka tajā iekļauti papildus sakotnējā mainigā pirmās starpības lagi, lai izvairītos no autokorelācijas atlikumos.

Datiem bez sezonalitātes tiek apskatīts modelis ARIMA(p, d, q) kur d tiek izvēlēts balstoties uz Kwiatkowski-Phillips-Schmidt-Shin (KPSS) testu. Tas ir, dati tiek pārbaudīti vienības saknei; ja rezultāts ir nozīmīgs, tiek pārbaudīti diferencēti dati vienības saknei, tas tiek darīts līdz brīdim, kad tiek iegūts pirmais nenozīmīgais rezultāts.

Datiem ar sezonalitāti ARIMA$(p, d, q)(P, D, Q)_s$, kur s ir sezonālā komponente un $D = 0$ vai $D = 1$ tiek izmantots Canova-Hansen (CH) tests[9]. Šis tests nodrošina tikai kritiskās vērtības $2 < s < 13$. Šajā gadījumā testa īstenošanā tiek pieļauta jebkura vērtība $s > 1$.

Pieņemsim, ka C_s ir kritiskā vērtība sezonalitātes periodam s. Tiek zīmēts C_s pret s, kur s vērtības ir līdz 365 un tiek ievērots, ka tās gandrīz sakrīt ar līniju $C_s = 0.269_s^{0.928}$. Tātad priekš vērtībām $s > 12$ tiek izmantota šī vienādība, lai iegūtu kritisko vērtību.
Kad D ir izvēlēts, tiek meklēts d pielietojot KPSS testu diferencētiem sezonāliem datiem (ja $D = 1$) vai sākotnējiem datiem (ja $D = 0$). Kad d (un iespējamais D) ir izvēlēti, tiek izvēlētas arī vērtības p, q, P un Q minimizējot AIC vērtību. Tiek atļauts $c \neq 0$ modeļiem, kur $d + D < 2$.
Pakāpienveida procedūra:

1. solis: tiek apskatīti pirmie četri iespējamie modeļi.
• ARIMA$(2, d, 2)$, ja $s = 1$ un ARIMA$(2, d, 2)(1, D, 1)$, ja $s > 1$.
• ARIMA$(0, d, 0)$, ja $s = 1$ un ARIMA$(0, d, 0)(0, D, 0)$, ja $s > 1$.
• ARIMA$(1, d, 0)$, ja $s = 1$ un ARIMA$(1, d, 0)(1, D, 0)$, ja $s > 1$.
• ARIMA$(0, d, 1)$, ja $s = 1$ un ARIMA$(0, d, 1)(0, D, 1)$, ja $s > 1$.

Ja $d + D \leq 1$, tad modeļi tiek piemēroti ar $c \neq 0$. Pretējā gadījumā pieņemam, ka $c = 0$. No šiem modeļiem tiek izvēlēts tas, kuram ir vismazākā AIC vērtība. Esošais modelis tiek apzīmēts ar ARIMA(p, d, q), ja $s = 1$ vai ARIMA$(p, d, q)(P, D, Q)_s$, ja $s > 1$.

2. solis: Tiek apskatītas trīspadsmit variācijas esošajam modelim:

• kur vienai no p, q, P un Q vērtībām tiek atļauts variēt no esošā modeļa par ± 1 vienību;
• kur abām p un q vērtībām tiek atļauts variēt no esošā modeļa par ± 1 vienību;
• kur abām P un Q vērtībām tiek atļauts variēt no esošā modeļa par ± 1 vienību;
• kur ir iekļauta konstante c, ja esošajam modelim $c = 0$, vai konstante netiek iekļauta, ja esošajam modelim $c \neq 0$.

Kad tiek atrasts modelis ar mazāko AIC vērtību, tas kļūst par jauno esošo modeli un procedūra tiek atkārtota. Process beidzas, kad vairs nevar atrast modeli, kas tuvs esošajam modelim, ar mazāku AIC[10].

4. Labākā modeļa noteikšana

Prognozēšanas precīzākā modeļa noteikšanai tiks izmantoti divi lielumi *MSE* un *MAPE*, kas balstītas uz prognozes kļūdu e_t, kuru izsaka šādi:

$$e_t = y_t - \widehat{y_t},$$

kur y_t - reālā vērtība no dotās datu kopas un $\widehat{y_t}$ - prognozētā vērtība laika momentā t. Prognozes vidējā kvadrātiskā kļūda (MSE) izmanto atlikumu kvadrātu summu,

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n},$$

kur n prognozēto vērtibu skaits. Vidējā absolūtā procentuālā kļūda (MAPE) apskata katras prognozes relatīvo absolūto kļūdu,

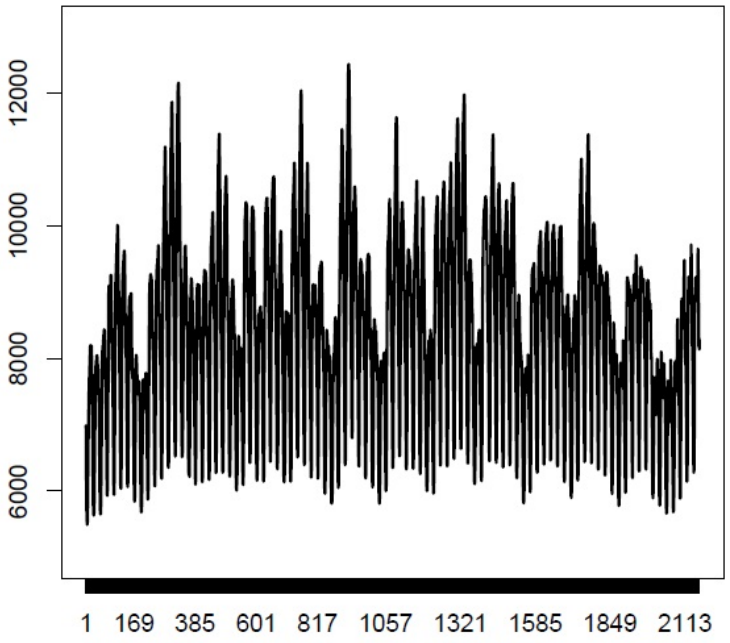
$$MAPE = \frac{\sum_{t=1}^{n} |\frac{e_t}{y_t}|}{n}.$$

Prognozēšanas modelis, kuram abas kļūdas ir vismazākās, ir piemērotākais reālajai datu kopai[11].

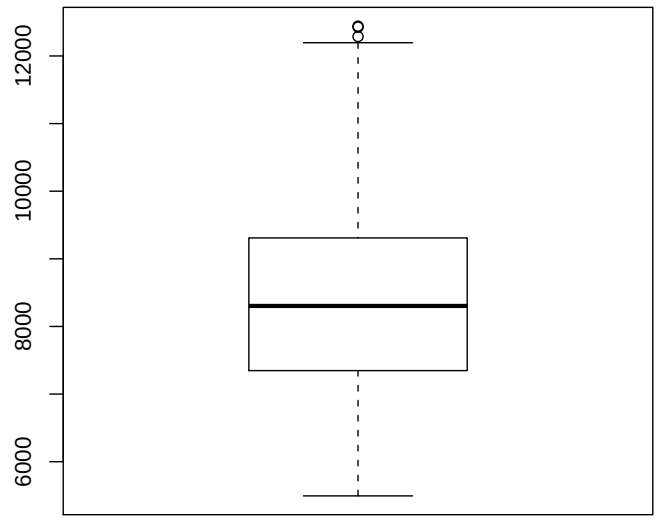
5. Praktiskā daļa

5.1. Elektoenerģijas pieprasījuma prognozēšana

Savam darbam izmantoju 2005. gada pirmo trīs mēnešu (janvāris, februāris, marts) elektroenerģijas pieprasījuma datus, kas uzdoti pa vienai stundai (kopā 2165 stundas), no Austrālijas Nacionālā elektroenerģijas pārstāvja (ANEM). Šie dati izvēlēti no publikācijas [1], lai labāk varētu salīdzināt prognozi ar ARIMA modeļa prognozi. Dati ir pieejami ikvienam Austrālijas Nacionālā elektroenerģijas pārstāvja mājas lapā [3].
Elektroenerģijas pieprasījuma laikrindas ir interesantas ar to, ka tajās ir novērojami dati "izlecēji". Tie var rasties, piemēram, neparadzētu laika apstākļu un dabas stihiju dēļ, kas ietekmē pieprasījuma pieaugumu vai samazināšanos. Darbā apskatīto datu vizuālā interpretācija.



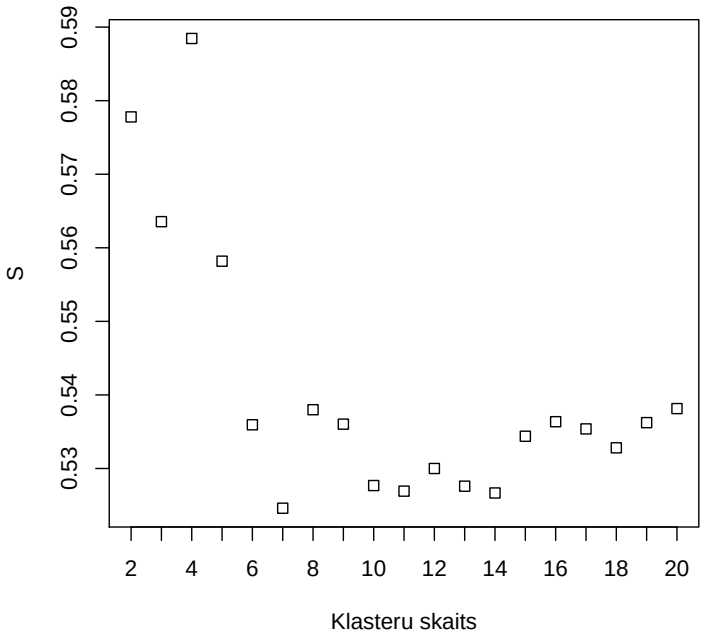
1. att.: Austrālijas elektroenerģijas pieprasījuma dati pa stundām, laika posmā no 2005.gada janvāra līdz 2005.gada martam.



2. att.: Kastu grafiks Austrālijas elektroenerģijas pieprasījuma dati pa stundām, laika posmā no 2005.gada janvāra līdz 2005.gada martam.

5.1.1. Silhouette indekss

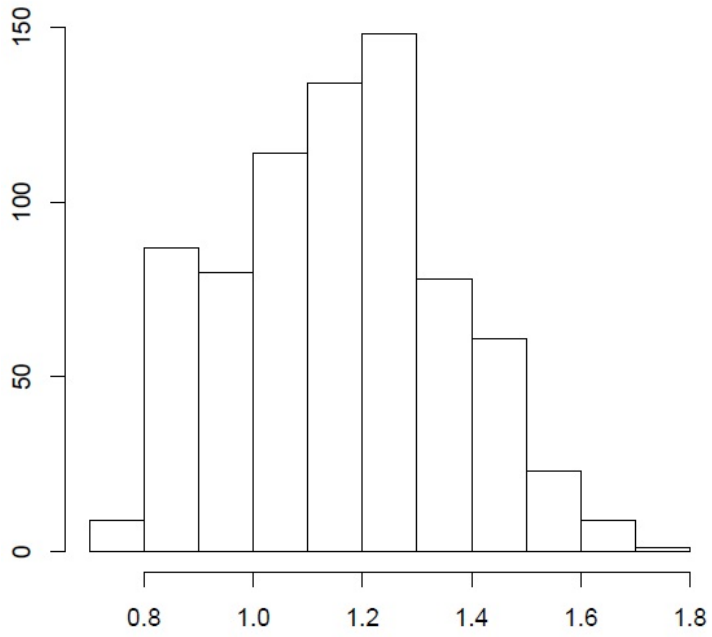
Silhouette indekss norāda, cik klāsteros dati būs jādala, lai veiksmīgāk varētu veikt datu apzīmēšanu. Jāizvēlas tas klāsteru skaits, kuram silhouette indekss ir vismazākais. No 3. attēla redzams, ka elektroenerģijas pieprasījuma klāsteru skaitu pēc Silhouette indeksa izvēlamies vienādu ar 7.



3. att. Silhouette indekss no 1 līdz 20 elektroenerģijas pieprasījuma datiem.

5.1.2. Datu normalizācija

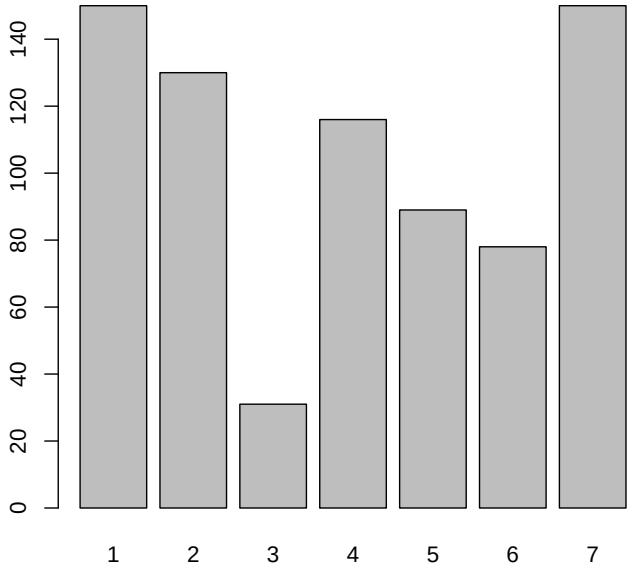
Pirms klāsterizācijas dati vēl ir jānormalizē. Varam pieņemt, ka tendence pieprasījumam izmaiņām gada garumā ir tāda pati kā iepriekšējo gadu izmaiņām, tas ir, oriģinālā tendence tiek nogludināta ar sākotnējiem datiem.



4. att. Histogramma elektroenerģijas pieprasījuma datiem pēc normalizācijas.

5.1.3. Klāsterizācija

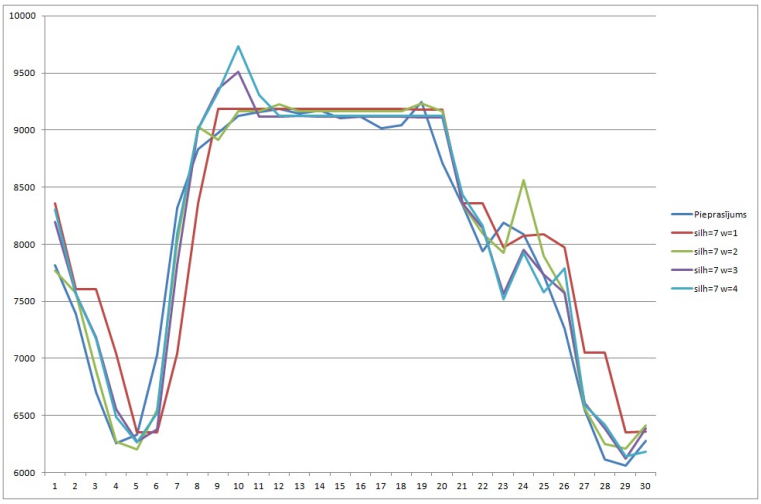
Tagad esam nonākuši pie vienas no svarīgākajām daļām - klāsterizācijas. Datu kopa, kas sastāv no ik stundas pieprasījumu datiem, klāsterizācijas problēma ir sadalīt datus K klāsteros tā, lai pieprasījuma dati sastāvētu no K klāsteriem. Kā rezultātā, datubāzes izmēri tiek krasi samazināti no 24 dimensijām uz vienu dimensiju. Izmantojot K-vērtības algoritma mērķa funkciju mēs apzīmējam datus ar attiecīgo k indeksu - klāsterizācija.



5. att.: Histogramma elektroenerģijas pieprasījuma datiem, kas sakārtoti tiem atbilstošajos klāsteros.

5.1.4. Loga garuma atrašana un prognozēšana

Loga garuma atrašanai tiek veikta prognoze datu kopai ar dažādiem logu garumiem un modelis, kuram būs vismazākās MSE un MAPE kļūdas, parādīs labāko loga garumu. Tagad veikšu prognozi elektroenerģijas pieprasījumam, izmantojot PSF algoritmu pirmajām 30 stundām, un atradīšu labāko loga garumu prognozēšanas modelim.



6. att.: Reālais pieprasījums un prognozētie dati 30 stundām ar dažādiem logu garumiem.

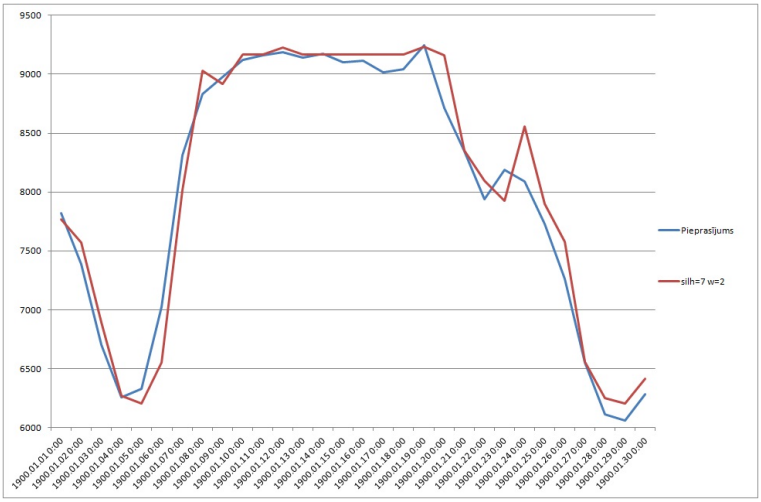
Grafikā redzams reālais elektroenerģijas pieprasījums un prognozētie dati ar dažādiem logu garumiem, kuri mainās no 1 līdz 4. Visi modeļi prognozē izmaiņas, kuru rašanās iemeslu var izskaidrot ar nakts iestāšanos, kad tas pieaug, vai dienas iestāšanos, kad tas samazinās. Pieprasījuma izmaiņas var būt skaidrojamas arī ar citiem iemesliem.

1. tabula Prognozes vidējā kvadrātiskā kļūda un vidējā absolūtā kļūda.

MSE			
silh=7 w=1	silh=7 w=2	silh=7 w=3	silh=7 w=4
215803.9469	40052.03358	77803.91658	85518.52224

MAPE			
silh=7 w=1	silh=7 w=2	silh=7 w=3	silh=7 w=4
0.044367284	0.018964833	0.026793525	0.028781916

Redzams, ka elektroenerģijas pieprasījuma prognozēšanai optimālais loga garums ir $W = 2$, jo tam ir vismazākā MSE un MAPE kļūda. Labākais modelis elektroenerģijas pieprasījuma prognozēšanai ir PSF algoritms ar Silhouette indeksu 7 un loga garumu 2.

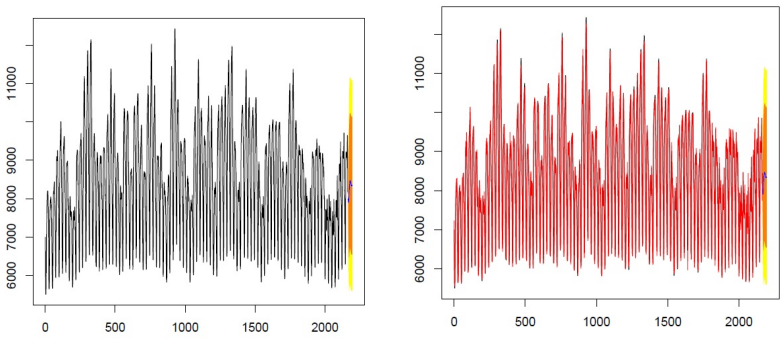


7. att.: Reālais pieprasījums (zilā krāsā) un PSF algoritma prognoze (sarkanā krāsā) 30 stundām.

Grafikā redzams, ka PSF algoritms labi reaģē uz pieprasījuma izmaiņām, kad tās samazinās vai palielinās.

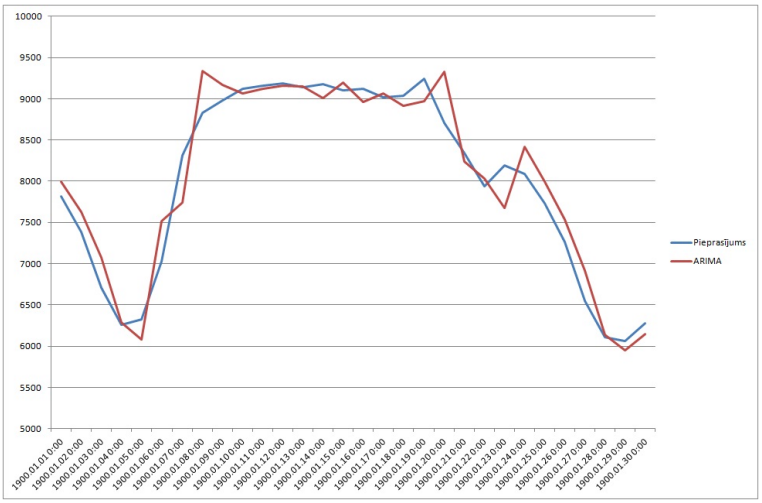
5.1.5. Prognozēšana ar ARIMA modeli

Piemērotāko ARIMA modeli $ARIMA(5,0,2)$ atrodu ar automātiski iebūvētu funkciju programmā R *auto.arima*:



8. att.: Kreisajā pusē elektroenerģijas pierasījuma grafiskais attēlojums ar pirmās stundas prognozi un tai atbilstošajām ticamības joslām . Labajā pusē modeļa $ARIMA(5,0,2)$ atbilstība reālajam elektroenerģijas pierasījumam.

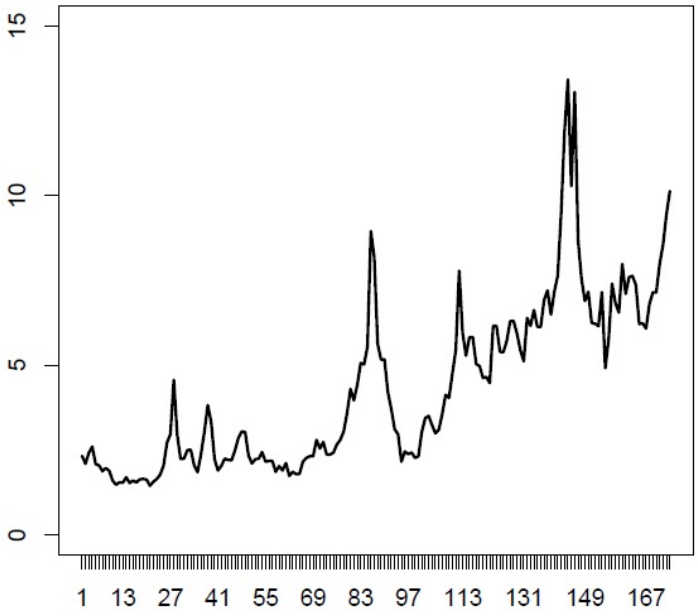
Grafikā redzams, ka ARIMA modelis arī labi atbilst pieprasījumam.



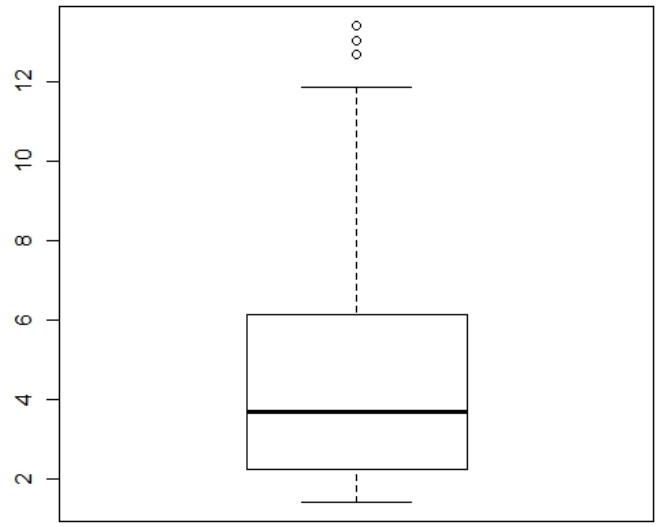
9. att.: Reālais pieprasījums (zilā krāsā) un prognozētie lielumi ar ARIMA modeli (sarkanā krāsā) 30 stundām.

5.2. Dabas gāzes cenas prognozēšana

Darbā izmantoju dabas gāzes cenu (dolāri par $28m^3$) datus no kompānijas "DOWJONES" datubāzes, kuri parāda katra mēneša cenu laika posmā no 1993.gada novembra līdz 2008.gada martam, kas uzdoti pa vienam mēnesim (kopā 174 mēneši). Dati ir pieejami ikvienam kompānijas "DOWJONES" mājas lapā [4]. Dabas gāzes cenas datus izvēlējos tāpēc, ka atradu to izmaiņas sākot ar 1993.gadu un tos bija viegli iegūt.
Dabas gāzes cenas laikrinda ir interesanta ar to, ka tajā ir novērojami dati "izlecēji", kas skaidrojami ar to, ka dabas gāzes cena ir piesaistīta mazuta cenai, kas parāda pasaules ekonomisko stāvokli un starpvalstu attiecības. Mazuts ir šķidrā kurināmā izejviela, kas līdzinās dīzeļdegvielai, tikai tā nav tik ļoti attīrīta. Darbā apskatīto datu vizuālā interpretācija.



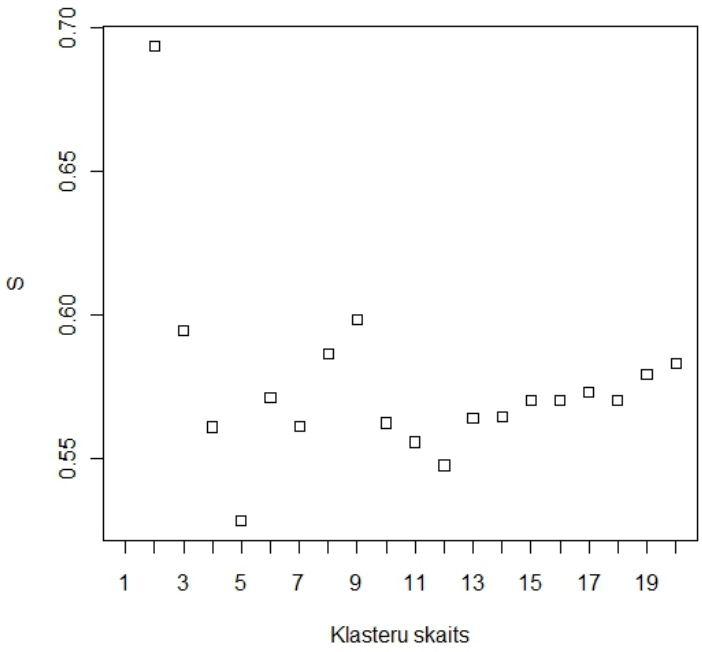
10. att.: Dabas gāzes cena dolāros par $28m^3$, laika posmā no 1993.gada novembra līdz 2008.gada martam.



11. att.: Kastu grafiks dabas gāzes cena dolāros par $28m^3$, laika posmā no 1993.gada novembra līdz 2008.gada martam.

5.2.1. Silhouette indekss

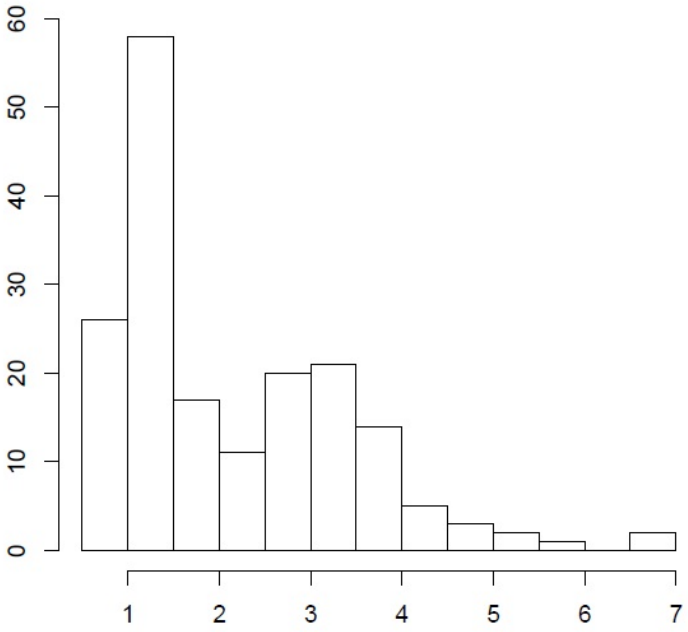
Silhouette indekss norāda, cik klāsteros dati būs jādala, lai veiksmīgāk varētu veikt datu apzīmēšanu. Jāizvēlas tas klāsteru skaits, kuram silhouette indekss ir vismazākais. No 12. attēla redzams, ka dabas gāzes cenas klāsteru skaitu pēc Silhouette indeksa izvēlamies vienādu ar 5.



12. att. Silhouette indekss no 1 līdz 20 dabas gāzes cenas datiem.

5.2.2. Datu normalizācija

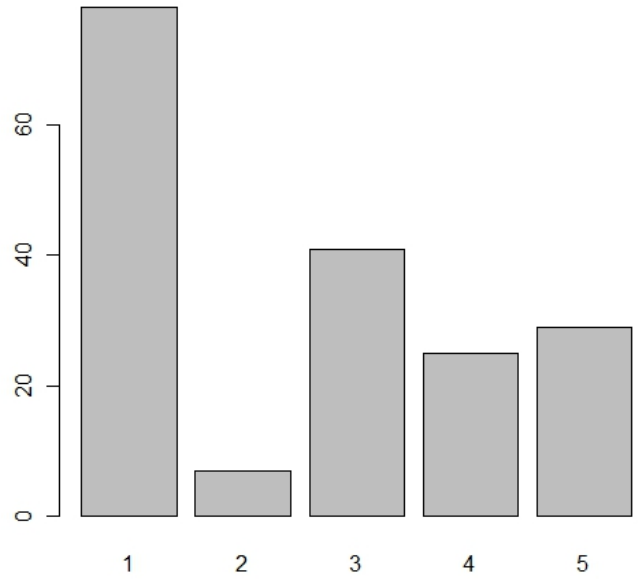
Pirms klāsterizācijas dati vēl ir jānormalizē. Varam pieņemt, ka tendence pieprasījumam izmaiņām gada garumā ir tāda pati kā iepriekšējo gadu izmaiņām, tas ir, oriģinālā tendence tiek nogludināta ar sākotnējiem datiem.



13. att. Histogramma dabas gāzes cenas datiem pēc normalizācijas.

5.2.3. Klāsterizācija

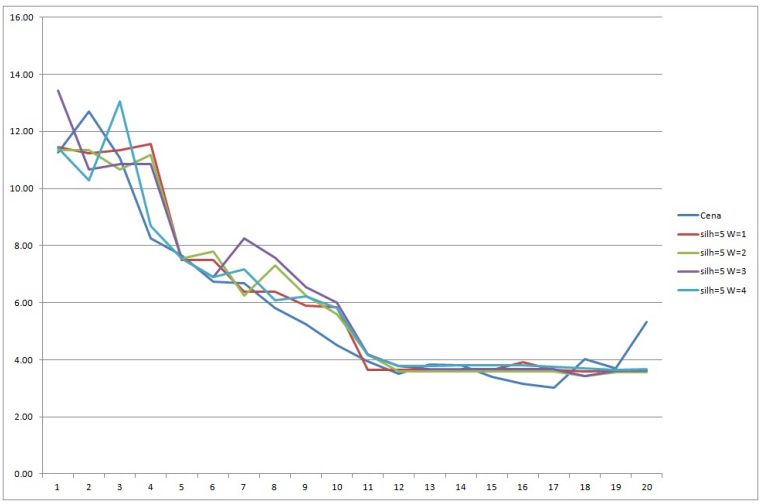
Tagad esam nonākuši pie vienas no svarīgākajām daļām - klāsterizācijas. Datu kopa, kas sastāv no ik stundas pieprasījumu datiem, klāsterizācijas problēma ir sadalīt datus K klāsteros tā, lai pieprasījuma dati sastāvētu no K klāsteriem. Kā rezultātā, datubāzes izmēri tiek krasi samazināti no 24 dimensijām uz vienu dimensiju. Izmantojot K-vērtības algoritma mērķa funkciju mēs apzīmējam datus ar attiecīgo k indeksu - klāsterizācija.



14. att.: Histogramma dabas gāzes cenas datiem, kas sakārtoti tiem atbilstošajos klāsteros.

5.2.4. Loga garuma atrašana un prognozēšana

Loga garuma atrašanai tiek veikta prognoze datu kopai ar dažādiem logu garumiem un modelis, kuram būs vismazākās MSE un MAPE kļūdas, parādīs labāko loga garumu. Tagad veikšu prognozi dabas gāzes cenai, izmantojot PSF algoritmu nākamajiem 20 mēnešiem, un atradīšu labāko loga garumu prognozēšanas modelim. Grafikā redzams reālās dabas gāzes cenas un prognozētie lielumi ar dažādiem logu garumiem, kuri mainās no 1 līdz 4. Visi modeļi prognozē izmaiņas, kuru rašanās iemeslu var izskaidrot ar pasaules ekonomisko stāvokli.



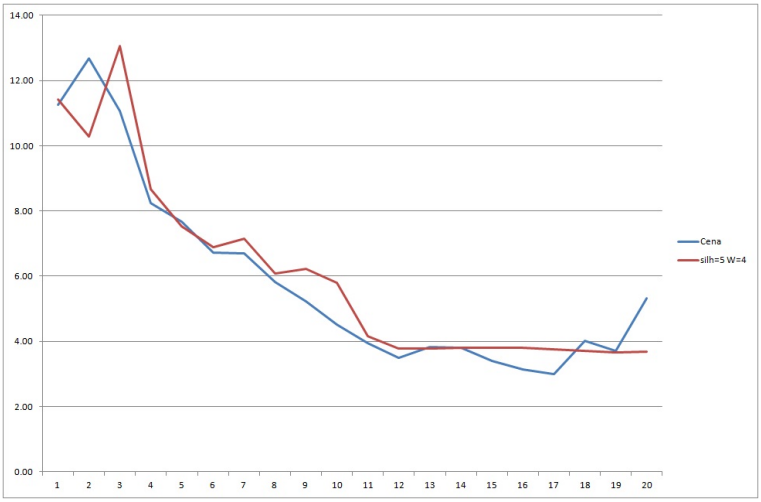
15. att. Reālā cena un prognozētie ldati 20 mēnešiem ar dažādiem logu garumiem.

2. tabula Prognozes vidējā kvadrātiskā kļūda un vidējā absolūtā kļūda.

MSE			
silh=5 w=1	silh=5 w=2	silh=5 w=3	silh=5 w=4
1.03704	1.02349	1.47027	0.85207

MAPE			
silh=5 w=1	silh=5 w=2	silh=5 w=3	silh=5 w=4
0.12230	0.12770	0.15154	0.10970

Redzams, ka dabas gāzes cenas prognozēšanai optimālais loga garums ir $W = 4$, jo tam ir vismazākā MSE un MAPE kļūda. Labākais modelis dabas gāzes cenas prognozēšanai ir PSF algoritms ar Silhouette indeksu 5 un loga garumu 4.

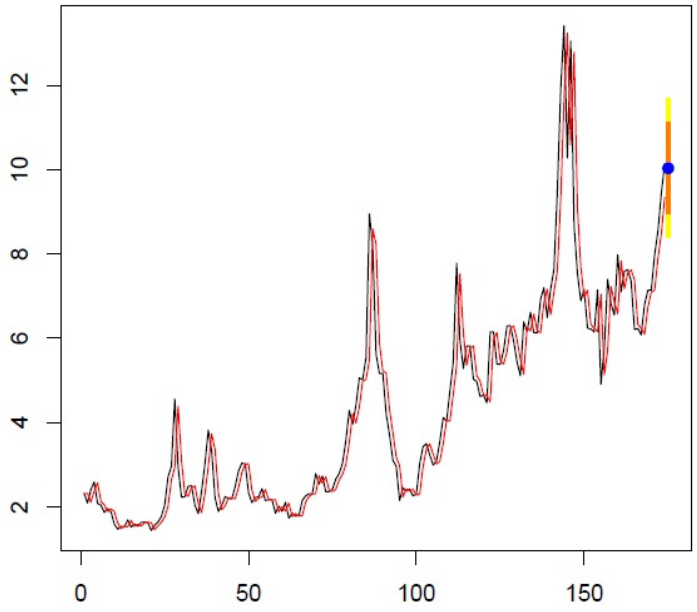


16. att.: Reālā cena (zilā krāsā) un PSF algoritma prognoze (sarkanā krāsā) 20 mēnešiem.

Grafikā redzams, ka PSF algoritms labi reaģē uz cenas izmaiņām, kad tā samazinās vai palielinās.

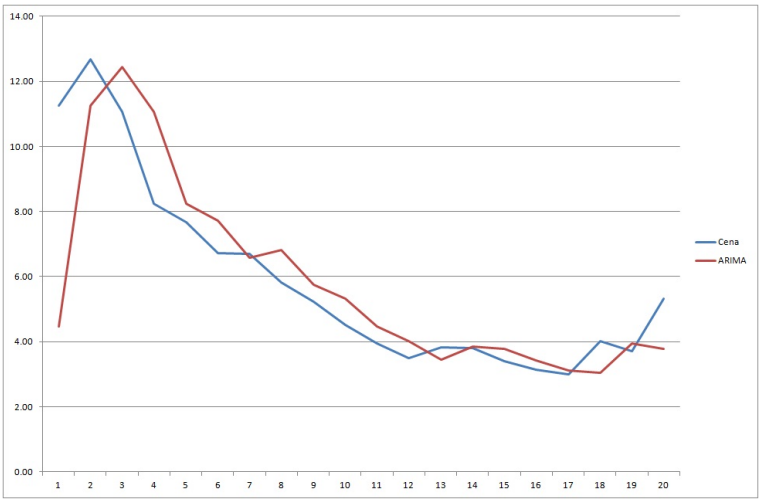
5.2.5. Prognozēšana ar ARIMA modeli

Piemērotāko ARIMA modeli $ARIMA(1,1,1)$ atrodu ar automātiski iebūvētu funkciju programmā R *auto.arima*:



17. att.: Dabas gāzes cenas grafiskais attēlojums ar pirmā mēneša prognozi un tai atbilstošajām ticamības joslām, modeļa $ARIMA(1,1,1)$ atbilstība dabas gāzes cenām (sarkanā krāsā).

Grafikā redzams, ka ARIMA modelis arī labi atbilst pieprasījumam, tikai pirmās prognozes tas veic ar nobīdi no reālās cenas.



18. att.: Reālā cena (zilā krāsā) un prognozētie lielumi ar ARIMA modeli (sarkanā krāsā) 20 mēnešiem.

6. Rezultāti un secinājumi

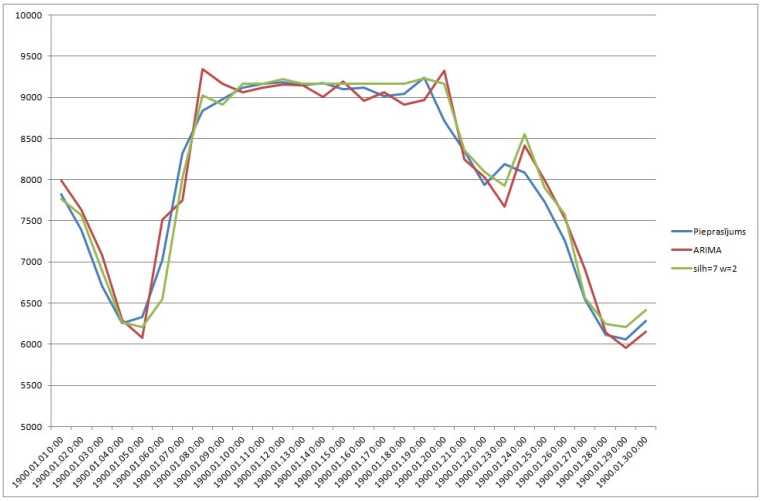
PSF algoritms prognozēšanai abos gadījumos ir precīzāks par ARIMA modeļa prognozēšanu, jo precīzāk spēja noteikt lielas atšķirības elektroenerģijas pieprasījuma un dabas gāzes cenas samazināšanās vai palielināšanās gadījumā. Tabulās redzams, ka kļūdas MSE un MAPE PSF algoritmam ir mazākas.

3. tabula: Elektroenerģijas pieprasījuma kļūdas prognozētajiem lielumiem ar ARIMA modeli un PSF algoritmu.

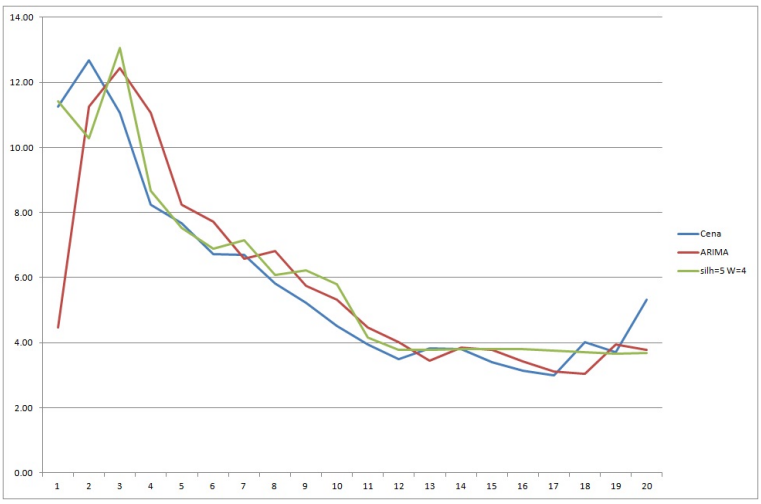
MSE		MAPE	
silh=7 w=2	ARIMA	silh=7 w=2	ARIMA
40 052	79 441	0.018964833	0.023705348

4. tabula: Dabas gāzes cenas kļūdas prognozētajiem lielumiem ar ARIMA modeli un PSF algoritmu.

MSE		MAPE	
silh=5 w=4	ARIMA	silh=5 w=4	ARIMA
0.85207	3.27206	0.10970	0.15537



19. att.: Reālais elektroenerģijas pieprasījums un prognozētie lielumi ar ARIMA modeli un PSF algoritmu.



20. att.: Reālā dabas gāzes cena un prognozētie lielumi ar ARIMA modeli un PSF algoritmu.

Prognozēšana ar PSF algoritmu, kas nesen ieviesta publikācijā [1], tika pielietota divos dažādos piemēros, elektroenerģijas pieprasījuma prognozēšanai un dabas gāzes cenas prognozēšanai. Abos gadījumos metode sniedza labus rezultātus, kas tika salīdzināti ar automātiski piedāvātu ARIMA modeļa funkcijas prognozēšanu, rezultātā pēc MSE un MAPE kļūdu salīdzināšanas, PSF algoritma prognoze bija precizāka.

Problēmas sagādā apjomīgs datu daudzums, kas palēnina darbu. To būtu iespējams novērst uzlabojot programmu- analizējot datus pa daļām.

Nākotnē varētu tik pētītas citas datu problēmas ar vairāk datu izlecējiem, kā arī apskatīti citi līdzīgi algoritmi [12].

7. Pateicība

Izsaku pateicību par iespēju diplomdarbu izstrādāt docenta Jāņa Valeiņa vadībā, par palīdzību materiālu iegūšanā un metodiskajiem ieteikumiem.

Izmantotā literatūra un avoti

[1] J. C. Riquelme F. Martinez-Alvarez, A. Troncoso and J. M. Riquelme. *Energy time series forecasting based on pattern sequence similarity.* IEEE Transactions on Knowledge and Data Engineering, 2010.

[2] J. M. Riquelme J. L. Martinez A. Troncoso, J. C. Riquelme and A. Gomez. Electricity market price forecasting based on weighted nearest neighbours techniques. In *IEEE Transactions on Power Systems*, pages 1294–1301, 2007.

[3] http://www.aemo.com.au/.

[4] http://research.stlouisfed.org/fred2/series/GASPRICE/downloaddata.

[5] R. Xu and D. C. Wunsch II. Survey of clustering algorithms. In *IEEE Transactions on Neural Networks*, page 645678, 2005.

[6] Morven Leese Daniel Stahl Brian S. Everitt, Sabine Landau. *Cluster Analysis 5th Edition.* Kings's College London, UK, 2011.

[7] J. D. Hamilton. *Time Series Analysis.* Princeton, 1984.

[8] http://cran.r project.org/web/packages/forecast/forecast.pdf.

[9] http://fabiobusetti.altervista.org/busetti$_h$*arvey*$_s$*easonality*$_t$*ests.pdf*.

[10] Rob J. Hyndmanr. *Automatic Time Series Forecasting: The forecast Package for R.* Journal of Statistical Software, July 2008, Volume 27, Issu 3.

[11] J. Valeinis N .Sinenko. *On comparison of univariate forecasting methods: the case of latvian residental property prices.* Proceedings of the International Conference APLIMAT 2010, 2010.

[12] J. C. Riquelme F. Martinez-Alvarez, A. Troncoso and J. M. Riquelme. Partitioning-clustering techniques applied to the electricity price time series. In *Lecture Notes in Computer Science*, pages 990–999, 2007.

Pielikums

1. Programmas kods

```
###########DATU iegūšana###########
library(clusterSim)
x<-c()
j=1:2165
x[j]<-scan(file="jandem.txt")

n=91
m=24
matrix(x[j], n , m)

###########DATU NORMALIZĒŠANA###########
i=1:24
x[j]<-x[j]/(sum(x[i])/24)

n=31
m=24
matrix(x[j],n,m)

M<-matrix(x[j],n,m)

c(t(M))
data<-c(t(M))
hist(data)
###########Silhouette index###########
library(clusterSim)
data<-scan(file="jandem.txt")
md <- dist(data, method="manhattan")
# nc - number_of_clusters
min_nc=2
```

```
max_nc=5
res <- array(0, c(max_nc-min_nc+1, 2))
res[,1] <- min_nc:max_nc
clusters <- NULL
for (nc in min_nc:max_nc)
{
cl2 <- pam(md, nc, diss=TRUE)
res[nc-min_nc+1, 2] <- S <- index.S(md,cl2$cluster)
clusters <- rbind(clusters, cl2$cluster)
}
print(paste("max S for",(min_nc:max_nc)[which.max(res[,2])],
"clusters=",max(res[,2])))
print("clustering for max S")
print(clusters[which.max(res[,2]),])
write.table(res,file="S_res.csv",sep=";",dec=",",
row.names=TRUE,col.names=FALSE)
plot(res,type="p",pch=0,xlab="Klasteru skaits",
ylab="S",xaxt="n")
axis(1, c(min_nc:max_nc))
############K-mean clustering############
data<-scan(file="jandem.txt")
kres<-kmeans(data,7)
plot(data)
kmeansRes<-factor(kres$cluster) #labeled data
plot(kmeansRes)
############Forecasting############
data<-scan(file="jandem.txt")
logi<- function(virkne, garums)
{
mekleta_virkne <- c(); pirmslogu_vertibas<-c(); j=1;
pirmslogu_indeksi<-c();
for (i in 1:garums)
```

```
{
 mekleta_virkne[garums+1-i] <- virkne[length(virkne)+1-i]
}
for (i in 1:(length(virkne)-garums-1))
{
 test <- identical(as.integer(virkne[i:(i+garums-1)]),
 mekleta_virkne)
 if (test == TRUE) #ja sakrit
 {
 pirmslogu_indeksi[j]<-(i+garums); j = j+1
 i = i+garums
 }
}
return (pirmslogu_indeksi)
}

indeksi<- logi(kmeansRes, 2)
iistie<-data[indeksi]
mean(iistie)
```

2. Loga garuma atrašana un prognozēšana ar PSF algoritmu - elektroenerģijas pieprasījuma datiem

Pieprasījums	silh=7 w=1	silh=7 w=2	silh=7 w=3	silh=7 w=4	silh=7 w=5
7819.36167	8357.246	7766.521	8193.757	8302.26	7982.334
7388.675	7604.79	7572.304	7564.896	7566.027	7754.137
6704.92167	7604.79	6891.71	7185.351	7169.566	7125.456
6259.23	7045.185	6268.874	6552.184	6488.435	6480.094
6328.51	6349.446	6205.209	6268.119	6264.326	6331.52
7027.825	6349.446	6549.928	6378.421	6519.227	6530.861
8315.72333	7041.84	8017.195	7823.948	8095.553	8180.91
8833.44667	8355.351	9026.259	9010.792	9010.792	9024.844
8976.06167	9183.187	8915.516	9363.842	9337.734	9337.734
9121.43333	9187.678	9167.898	9509.342	9732.924	9729.159
9160.34167	9182.752	9167.898	9115.7	9309.999	9061.273
9187.01667	9182.624	9223.925	9115.7	9124.01	9295.97
9142.215	9182.577	9167.748	9122.073	9124.01	9171.423
9172.39833	9187.037	9167.801	9116.142	9124.323	9124.25
9103.765	9186.943	9167.73	9116.241	9124.41	9117.14
9116.11667	9182.481	9167.744	9116.454	9124.646	9117.469
9014.07667	9182.318	9167.568	9116.406	9124.544	9117.388
9040.6	9182.18	9167.428	9116.405	9124.503	9117.38
9243.33333	9178.56	9230.156	9113.525	9123.97	9185.28
8712.43667	9178.274	9162.272	9113.252	9123.569	9123.362
8346.51667	8356.546	8353.622	8357.967	8431.349	8371.628
7936.33333	8356.546	8096.545	8143.28	8159.031	8170.326
8187.83167	7969.975	7923.308	7562.693	7516.8	7823.412
8087.82333	8074.636	8559.025	7951.165	7922.969	7983.115
7731.285	8090.045	7897.238	7731.597	7577.562	7576.303
7260.14833	7970.743	7576.539	7570.41	7789.81	8155.805
6551.53833	7047.238	6561.854	6611.36	6591.876	NaN
6112.85833	7047.238	6247.28	6386.344	6422.19	NaN
6059.21667	6351.713	6206.61	6122.039	6143.128	NaN
6280.21333	6356.249	6414.644	6386.061	6179.072	NaN

3. Loga garuma atrašana un prognozēšana ar PSF algoritmu - dabas gāzes cenas datiem

Cena	silh=5 W=1	silh=5 W=2	silh=5 W=3	silh=5 W=4
11.26	11.462	11.3575	13.42	11.42
12.69	11.24	11.3575	10.67	10.28
11.06	11.338	10.67	10.865	13.05
8.25	11.56333	11.175	10.865	8.68
7.67	7.492308	7.54	7.54	7.54
6.73	7.492308	7.78353	6.89	6.89
6.69	6.374118	6.251429	8.258	7.16
5.81	6.374118	7.315	7.57125	6.08
5.23	5.896098	6.24125	6.55	6.225
4.52	5.844	5.5875	6.002857	5.8
3.94	3.642759	4.18	4.18	4.165
3.50	3.642759	3.585	3.79	3.79
3.83	3.652667	3.585	3.66	3.79
3.81	3.647742	3.581304	3.66	3.794444
3.39	3.653437	3.591667	3.670625	3.794444
3.15	3.918	3.6004	3.678824	3.796
3.01	3.650294	3.592308	3.662778	3.759091
4.02	3.595143	3.421154	3.434737	3.708333
3.70	3.618611	3.555714	3.6045	3.654615
5.33	3.629459	3.571724	3.624286	3.680714

DIPLOMDARBS“PROGNOZĒŠANA AR KLĀSTERIZĀCIJAS METODI” izstrādāts Latvijas Universitāte Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Anete Rubine

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: doc. Dr.math. Nadežda Siņenko

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātiskās analīzes nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)