Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
**56** t., 2015, 29–34

ISSN 0132-2818
www.mii.lt/LMR/

# Algebriniai laiko eilučių prognozės bei segmentavimo modeliai

**Kristina Lukoševičiūtė, Rita Palivonaitė**

*Kauno technologijos unversitetas, Matematikos ir gamtos mokslų fakultetas*
Studentų 50, LT-51368 Kaunas
E. paštas: kristina.lukoseviciute@ktu.lt, rita.palivonaite@ktu.lt

**Santrauka.** Straipsnyje pateiktas laiko eilučių segmentavimo modelis, kuris pagrįstas trumpų laiko eilučių prognozės modelio su vidiniu glotninimu eilės nustatymu. Segmentuojant daroma prielaida, kad seka sudaryta iš segmentų, kuriuose slypi atitinkamos eilės algebrinės sekos fragmentas. Modelis pritaikytas realios finansinės laiko eilutės segmentavimui.
**Raktiniai žodžiai:** Hankelio matrica, laiko eilučių prognozė, laiko eilučių segmentavimas, algebrinė seka.

## Įvadas

Laiko eilučių segmentavimo uždavinys yra aktualus daugelyje mokslo, inžinerijos ir finansų sričių. Paprastai pagrindinis visų segmentavimo modelių tikslas – detektuoti stacionarius ar nestacionarius laiko eilučių rėžimus, aproksimuoti atitinkamus segmentus paprastesniais modeliais. Pagrindiniai plačiai žinomi algoritmai yra slenkančių langų (angl. *sliding windows*) ir einantis nuo smulkmenų prie bendrųjų principų (angl. *bottom-up*) segmentavimo algoritmai [1].

Hankelio matricos laiko eilučių prognozės uždaviniuose panaudotos [6] straipsnyje, o algebrinis sekos segmentavimo modelis, panaudojant Hankelio matricą, pateiktas [4] straipsnyje. Pagrindinis šio straipsnio tikslas – segmentuoti seką, panaudojant [5] pasiūlytą trumpų laiko eilučių prognozavimo metodiką ir naujai konstruojamą segmentavimo metodiką bei rezultatus palyginti su gerai žinomais metodais.

## 1 Sekos rango sąvoka

Tiesinė rekurentinė seka, kurios eilė $n$, aprašoma lygtimi:

$$x_k = \alpha_{n-1}x_{k-1} + \alpha_{n-2}x_{k-2} + \cdots + \alpha_0 x_{k-n}, \quad k = 0, 1, \ldots, \tag{1}$$

čia koeficientai $\alpha_j$, $j = 0, 1, \ldots, n-1$. Pradinės sąlygos $x_k$, $k = 0, 1, \ldots, n-1$ vienareikšmiškai apibrėžia šios sekos evoliuciją [2]. Pagalbinis tiesinės rekurentinės sekos (1) daugianaris konstruojamas iš lygties:

$$P(\rho) = \rho^n - \alpha_{n-1}\rho^{n-1} - \alpha_{n-2}\rho^{n-2} - \cdots - \alpha_0. \tag{2}$$

Tiesinė rekurentinė seka (2) generuojama lygtimi:

$$x_j = \sum_{k=1}^{r} \sum_{l=0}^{n_k-1} \mu_{kl} \binom{j}{l} \rho_k^{j-l}, \tag{3}$$

čia koeficientai $\mu_1, \mu_2, \ldots, \mu_n$ apibrėžti taip, kad atitiktų pradines rekurentinės sekos sąlygas; $r$ – daugianario šaknų skaičius; $n_k$ – $k$-tosios šaknies kartotinumo indeksas; $n_1 + n_2 + \cdots + n_r = n$.

Jei nežinoma sekos $(x_j)_{j=0}^{+\infty}$ eilė, tiesinės rekurentinės sekos modelio rekonstravimas iš šių duomenų gali būti sudėtingas uždavinys. Tačiau, iš sekos $(x_j)_{j=0}^{+\infty}$, Hankelio transformacijos pagalba, galima sukonstruoti determinantų seką $(h_j)_{j=0}^{+\infty}$; čia $h_j = \det H_j$ ir $H_j = (x_{k+l-2})_{1\leqslant k,l\leqslant(j+1)}$ yra Hankelio matrica(matricos $H_j$ eilė $(j+1)\times(j+1)$). Jei egzistuoja toks $n \geqslant 1$ su kuriuo $h_n \neq 0$, bet $h_k = 0$ visiems $k > n$, tuomet $(x_j)_{j=0}^{+\infty}$ yra tiesinė rekurentinė seka, kurios eilė yra $n$ ir pagalbinė lygtis:

$$\begin{vmatrix} x_0 & x_1 & \cdots & x_n \\ x_1 & x_2 & \cdots & x_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n-1} & x_n & \cdots & x_{2n-1} \\ 1 & \rho & \cdots & \rho^n \end{vmatrix} = 0. \tag{4}$$

Ši tiesinių algebrinių lygčių sistema turi vienintelį sprendinį, nes $h_n \neq 0$ [3].

## 2 Prognozės modelis

Tarkime, turime $2n$ stebėjimų: $x_0, x_1, x_2, \ldots, x_{2n-1}$, čia $x_{2n-1}$ – dabarties momento reikšmė. Laikykime, kad sekos eilė $n$. Tuomet kitas elementas $x_{2n}$ tiesiogiai ir vienareikšmiškai suskaičiuojamas iš lygties (jei sekos eilė yra $n$):

$$\det H^{(n+1)} = \det \begin{bmatrix} x_0 & x_1 & \ldots & x_n \\ x_1 & x_2 & \ldots & x_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_{n+1} & \ldots & x_{2n} \end{bmatrix} = 0. \tag{5}$$

Tačiau, šis tiesioginis $x_{2n}$ sekos nario skaičiavimas tinka tik tiesinėms rekurentinėms sekoms. Tačiau, jei duotoji seka nėra algebrinė seka, triukšmo pašalinimas ir bazinio algebrinės sekos fragmento ir eilės $n$ identifikavimas tampa aktualiu uždaviniu.

Priimkime prielaidą, kad turime $2n$ stebėjimą ir duotoji seka sudaryta iš nežinomos „užtriukšmintos“ tiesinės rekurentinės sekos:

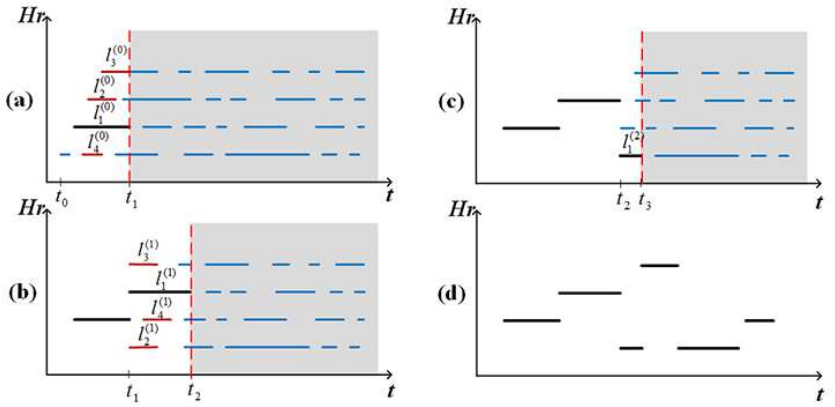
$$x_k = \tilde{x} + \varepsilon_k, \quad k = 0, 1, 2, \ldots, \tag{6}$$

čia $\varepsilon_k$; $k = 0, 1, 2, \ldots$ – triukšmas.

Tarkime, kad $\tilde{x}_k$; $k = 0, 1, 2, \ldots$ yra algebrinė išraiška, apibrėžianti globalią eilutės dinamiką. Tuomet pagal (5) determinantas

$$\det \tilde{H}^{(n+1)} = 0. \tag{7}$$

Kyla klausimas, kaip identifikuoti triukšmą. Toks uždavinys turi be galo daug sprendinių. Mūsų tikslas – minimizuoti bet kokius nuokrypius nuo tariamos algebrinės išraiškos. Kad sušvelnintume pagal (7) lygtį suskaičiuotą prognozę $\tilde{x}_{2n}$, įvedame nuokrypio nuo slenkančio vidurkio prognozės komponentę. Slenkančio vidurkio prognozė




**1 pav.** Algebrinio segmentavimo metodo segmento parinkimo schema.

skaičiuojama pagal formulę $\bar{x}_k = \frac{1}{s}\sum_{i=0}^{s-1} x_{k-i-1}$, čia $s$ – slenkančio vidurkio prognozės langas. Todėl triukšmų sekai $\{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{2n-1}\}$ siūlome tokią tikslo funkciją, kurią reikia maksimizuoti:

$$F(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{2n-1}) = \frac{1}{a\sum_{k=0}^{2n-1} |\varepsilon_k| + |\tilde{x}_{2n} - \bar{x}_{2n}|}, \tag{8}$$

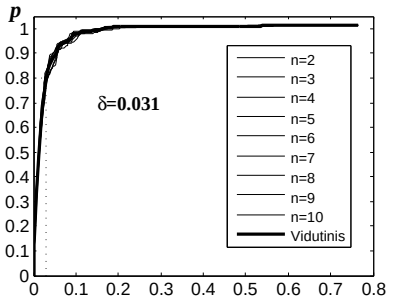
čia $\tilde{x}_{2n}$ (7) lygties sprendinys, $\bar{x}_{2n}$ – slenkančio vidurkio prognozė, $a > 0$. Siekiame, kad $\varepsilon_k$; $k = 0, 1, 2, \ldots, 2n-1$ būtų kuo mažesni ir algebrinė prognozė nenutoltų nuo slenkančio vidurkio prognozės. Pusiausvyra tarp šių dydžių koreguojama parenkant atitinkamą parametro $a$ reikšmę.

## 3 Segmentavimo algoritmas

Toliau pristatome segmentavimo algoritmą pagrįstą trumpų laiko eilučių prognozės paklaidų įvertinimu. Segmentuojant daroma prielaida, kad seka sudaryta iš segmentų, kuriuose slypi algebrinės sekos, kurių eilė gali būti iš intervalo $2 \leqslant n \leqslant 10$. Tuomet visai sekai atlikus prognozę su visais $n$ įvertinamos šios prognozės paklaidos. Kitas žingsnis – pasirinkti priimtiną atliktos algebrinės prognozės paklaidų lygį $\delta$. Pasiūlytos segmentavimo metodikos idėja paprasta – eilės $n$ algebrinis modelis yra pakankamai geras, jei prognozės paklaidos neviršija pasirinkto paklaidos lygio $\delta$.

Segmentavimo algoritmas, kurio schema pateikta 1 pav., vykdomas pagal šiuos žingsnius:

A. Pasirinktoms sekos eilėms $n$ atliekama algebrinė sekos prognozė.
B. Parenkamas paklaidos lygis $\delta$ ($\delta > 0$). Horizontaliomis linijomis žymimi skirtingų eilių $n$ sekos segmentai, kuriems algebrinės prognozės paklaidos neviršija $\delta$ reikšmės (1 pav.).
C. Identifikavimas pradedamas nuo pradinio laiko momento $t_0$, kai randamas pirmasis segmentas. Jeigu pirmasis aptiktas segmentas yra pakankamai ilgas (trumpiausias segmentas yra keturių laiko intervalų ilgio) ir ilgesnis už kitus segmentus tame laiko intervale, tuomet jis parenkamas kaip $n$ eilės segmentas $l_1^{(0)}$ (paryškintas storesne linija (1(a) pav.)).




**2 pav.** Segmentavimo paklaidų lygio $\delta$ nustatymas sekos eilėms $2 \leqslant n \leqslant 10$.

D. Išrinkus ilgiausią segmentą, susijusį su atitinkama eile, ištrinama visa kita informacija tame intervale.

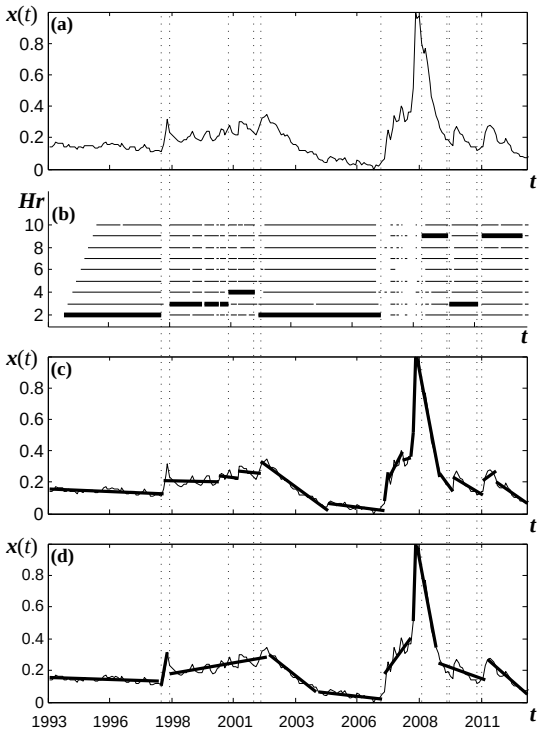
E. Kartojame B ir C algoritmo etapus, nuo laiko momento, kur baigėsi pirmasis segmentas ir toliau renkamas kitas ilgiausias segmentas $l_1^{(1)}$ (paryškintas storesne linija (1(b) pav.)). Procesas tęsiamas tol, kol randami visi segmentai, atitinkantys tam tikrą sekos eilę $n$.

## 4 Skaitiniai eksperimentai

Segmentavimo algoritmas pritaikytas segmentuojant finansinę laiko eilutę. Segmentuojama eilutė yra vienas iš labiausiai žinomų indikatorių, kuris matuoja JAV finansinės sistemos būklę/stresą – STLFSI (St. Louis Fed Financial Stress Index). Tyrime naudojama laiko eilutė nuo 1993-12-31 iki 2013-08-23 metų duomenys rinkti kas mėnesį. Segmentuojant laiko eilutės paklaidų lygio parinkimui vertinamas vidutinis procentas $p$ prognozės reikšmių, kurių paklaidos neviršija to lygio. Pagal Palivonaite ir kt. [4] rekomendacijas, laikome, kad geriausias paklaidų lygis gaunamas, kai $p = 0.8$. Tuomet STLFSI laiko eilutei paklaidų lygis turi neviršyti reikšmės $\delta = 0.031$ (2 pav.). Algebrinio segmentavimo rezultatai pateikti 3 pav. Galima pastebėti, kad pasiūlyta algebrinio segmentavimo metodika identifikuoja daugiau nei du skirtingas eiles $n$ atitinkančius segmentus (3 pav.). Naujo segmentavimo algoritmo rezultatus galima palyginti su gerai žinomo slenkančių langų (3(c) pav.) ir einantis nuo smulkmenų prie bendrųjų principų (3(d) pav.) segmentavimo algoritmų rezultatais. 3 paveiksle pastebime, kad STLFSI eilutės segmentai yra dažnesni, nei algebrinio segmentavimo atveju. Mūsų pristatyto algoritmo atveju, segmentai labai aiškiai atskiria pokyčių fragmentus, kai finansinė situacija visiškai pasikeičia. Pasiūlytas segmentavimo metodas identifikuoja laiko eilutės segmentus – kiekvienam jų priskiria $n$ eilės algebrinį dėsnį.

## 5 Išvados

Pristatyti laiko eilučių prognozės bei segmentavimo modeliai patogūs naudoti, nes jie nereikalauja turėti ilgos praeities duomenų. Šie modeliai pagrįsti konkrečiu algebriniu sekų formavimo modeliu, todėl yra lengvai atkartojami, jei seka nėra triukšminga. Pristatyto segmentavimo algoritmo privalumas, kad identifikuodami segmento algeb-




**3 pav.** STLFSI eilutės algebrinio segmentavimo rezultatai (a) STLFSI eilutė; (b) segmentavimas priskiriant sekos algebrinį dėsnį, kai $2 \leqslant n \leqslant 10$; (c) slenkančių langų segmentavimo algoritmas; (d) einantis nuo smulkmenų prie bendrųjų principų segmentavimo algoritmas.

rinį modelį, mes tuo pačiu metu automatiškai nustatome modelio sudėtingumo eilę (kas nebūdinga kitiems segmentavimo algoritmams).

## Literatūra

[1] E. Keogh, S. Chu, D. Hart and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 289–296, 2001.

[2] V.L. Kurakin, A.S. Kuzmin and A.A. Nechavev. Linear complexity of polinear sequences. *J. Math. Sci.*, **76**:2793–2915, 1995.

[3] Z. Navickas and L. Bikulciene. Expressions of solutions of ordinary differential equations by standard functions. *Math. Model. Anal.*, **75**(1):399–412, 2006.

[4] R. Palivonaite, K. Lukoseviciute and M. Ragulskis. Algebraic segmentation of short non-stationary time series based on evolutionary prediction algorithms. *Neurocomputing*, **121**:354–364, 2013.

[5] R. Palivonaite and M. Ragulskis. Short-term time series algebraic forecasting with internal smoothing. *Neurocomputing*, **127**:161–171, 2014.

[6] M. Ragulskis, K. Lukoseviciute, Z. Navickas and R. Palivonaite. Short-term time series forecasting based on the identification on skeleton algebraic sequences. *Neurocomputing*, **64**:1735–1747, 2011.

SUMMARY

**Algebraic time series forecasting and segmentation models**

*K. Lukoševičiūtė, R. Palivonaite*

An algebraic segmentation method based on algebraic predictor with internal smoothing is proposed. The concept of the rank of the sequence is proposed for the detection of a base fragment of the sequence. Numerical experiments with a real-world financial time series illustrate the segmentation method.

*Keywords*: Hankel matrix, time series forecasting, time series segmentation, algebraic sequence.