



Vilniaus Universitetas

2. Laiko eilučių koreliacinės funkcijos

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

vainius.gataveckas@mif.stud.vu.lt

matas.gaulia@mif.stud.vu.lt

dovydas.martinkus@mif.stud.vu.lt

Vilnius, 2022

Turinys

Jvadas	3
Duomenys	4
ACF	7
PACF	8
ACF ir PACF	9
CCF	10
EACF	11
Šaltiniai.....	12

Ivadas

Laiko eilučių teorijoje autoregresinių (AR), slenkančio vidurkio (MA) ir autoregresinio slenkančio vidurkio (ARMA) procesų eilės nustatymui pasitelkiami koreliacijų skaiciavimo metodai. MA q eilės nustatymui pasitelkiamas autokoreliacijos skaiciavimas. AR p eilės nustatymui dalinė autokoreliacija. ARMA p ir q eilę galima įvertinti su išplėstine autokoreliacijos funkcija (angl. extended autocorrelation function). Šis darbas aptaria ir kryžminės koreliacijos metodą (angl. cross correlation function).

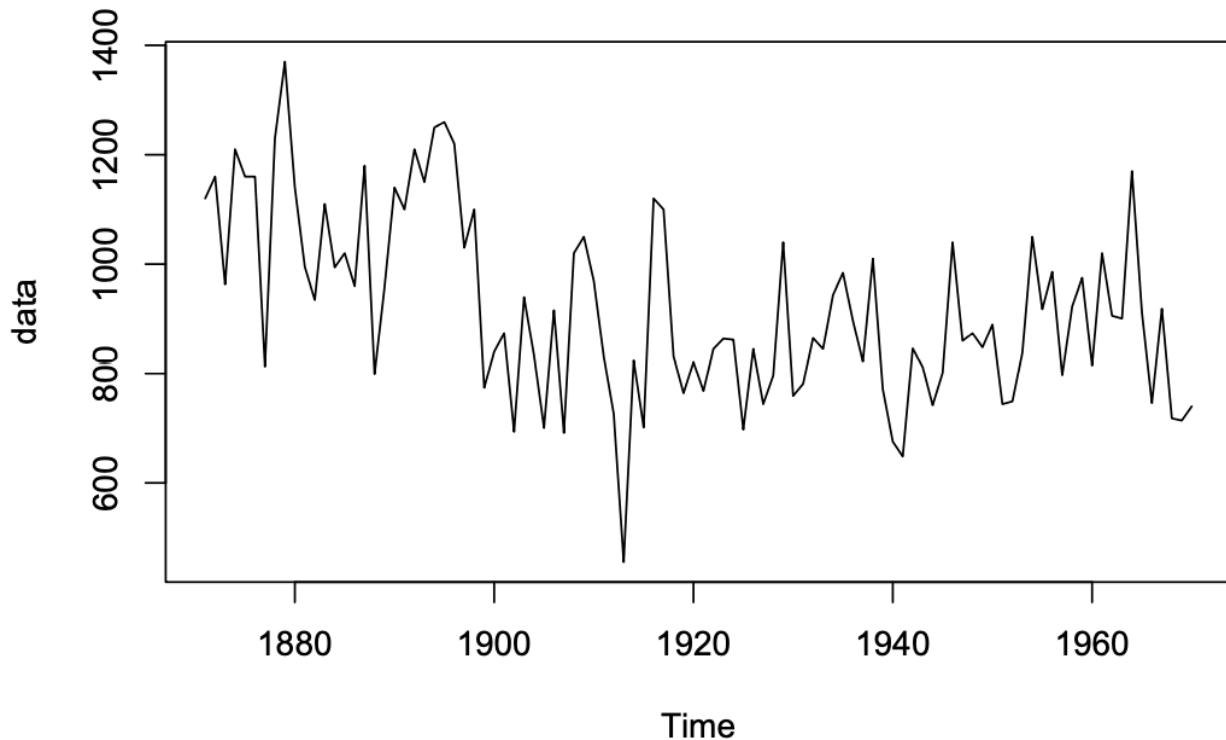
Metodai realizuoti R programine įranga. Naudojantis šios įrangos nupieštais grafikai galima lengviau identifikuoti tendencijas reikalingas įvertinti p ir q eiles, taip pat sezono periodo įtaką. Pateiktuose pavyzdžiuose koncentruojamės į pačius metodus ir jų skaiciavimą, t.y. netaikysime pilnos Box-Jenkins metodologijos.

Metodų panaudojimui reikalinga silpno stacionarumo prielaida. Tam pasiekti bus atliekamas duomenų diferencijavimas (skirtumų ēmimas).

Duomenys

Šiam darbui kaip pavyzdinius duomenis naudosime 2 duomenų rinkinius, abu yra jau iš karto instaliuoti su R kalba. Pirmasis duomenų rinkinys susideda iš kasmetinių Nilo upės matavimų, buvo matuojama kiek kubinių metrų vandens praplaukia pro Asuaną, stebėta tarp 1871 - 1970 metų. Antrajį duomenų rinkinį sudaro stebėjimai kiek per metus ant saulės paviršiaus atsiranda tamsių dėmių, stebėta tarp 1700 ir 1988. Duomenis galima prieiti tiesiog per R parašius duomenų šaltinių pavadinimus.

```
> data <- Nile  
> plot(data, type = "l")
```



1 pav. Nilo upės duomenys.

Iš grafiko (1 pav.) atrodo, kad laiko eilutė nestacionari, patikrinsime tą su Dickey-Fuller testu.

```
> adf.test(data)

Augmented Dickey-Fuller Test

data: data
Dickey-Fuller = -3.3657, Lag order = 4, p-value = 0.0642
alternative hypothesis: stationary
```

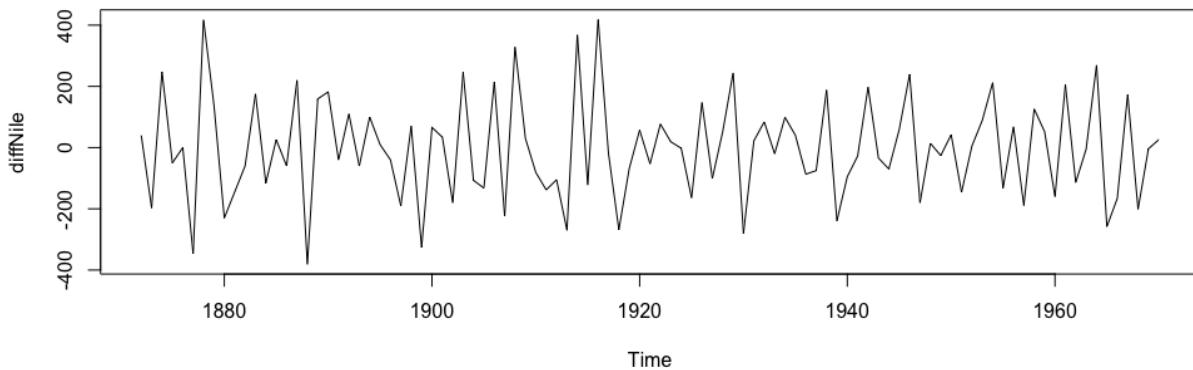
Kadangi p reikšmė yra daugiau nei 0.05, neatmetame nulinės hipotezės, tai reiškia kad mūsų seka nėra stacionari, reikia ją diferencijuoti ir pabandyti vėl.

```
> diffNile <- diff(data, differences=1)
> adf.test(diffNile)

Augmented Dickey-Fuller Test

data: diffNile
Dickey-Fuller = -6.5924, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Dabar p reikšmė yra mažiau nei 0.05, tad galime sakyti kad diferencijuota seka yra stacionari.



2 pav. Diferencijuoti Nilo upės duomenys

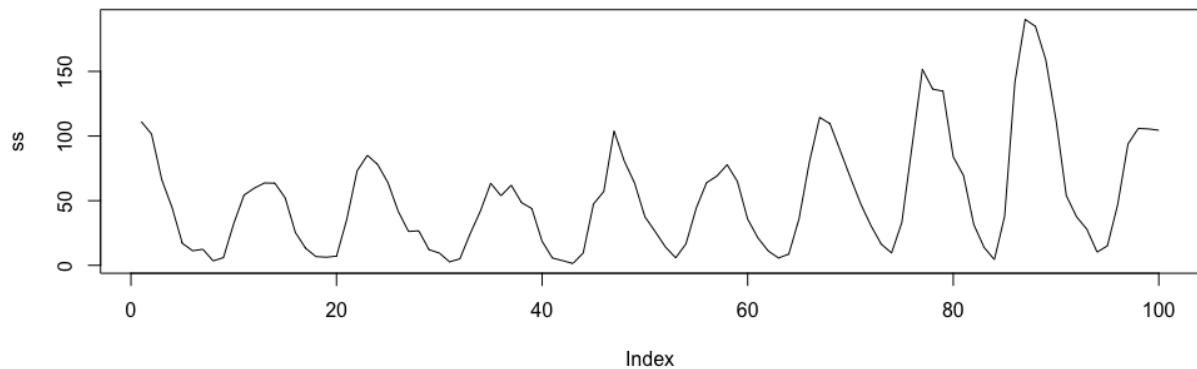
Antrieji duomenys yra saulės dėmių skaičius jos paviršiuje.

```
> ss <- sunspot.year  
> ss <- ss[time(ss) >= 1871 & time(ss) <= 1970]  
> plot(ss, type = "l")  
> adf.test(ss)
```

Augmented Dickey-Fuller Test

```
data: ss  
Dickey-Fuller = -6.5954, Lag order = 4, p-value = 0.01  
alternative hypothesis: stationary
```

Matome kad antroji laiko eilutė yra stacionari ir jos nereikia diferencijuoti.



3 pav. Saulės dėmių skaičiaus duomenys

ACF

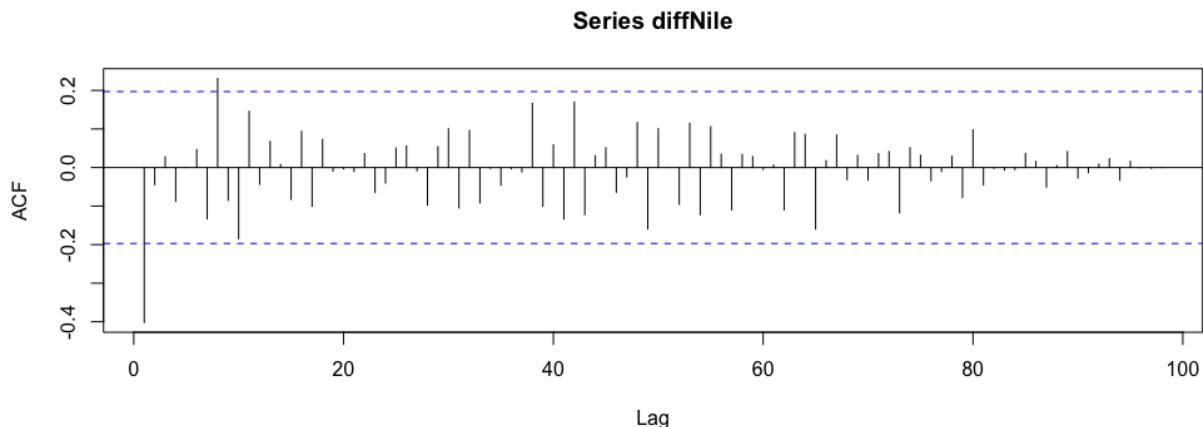
ACF - autokoreliacijos funkcija, kuri matuoja kaip koreliuoja dabartinio momento reikšmė su praetomis. Funkcija apibrėžta taip:

$$\rho_X = \frac{r_X(h)}{r_X(0)}$$

Kur r – kovariacinė funkcija, h – vėlavimo operatoriaus laipsnis.

Naudojant R galima labai paprastai gauti ACF grafiką. Funkcijos `acf()` parametras „`lag.max`“ nusako maksimalų h kuris bus vaizduojamas grafike.

```
> acf(diffNile, lag.max = 100)
```



4 pav. Nilo duomenų ACF grafikas

Grafike matoma punktyrinė linija yra statistiškai reikšmingų koreliacijų riba su hipoteze $\rho = 0$. Taip pat matoma silpstanti sinusoidė susidaranti iš koreliacijų reikšmių. Toks koreliacijų elgesys nėra išskirtinis.

Šis grafikas yra naudingas tame, kad galime surasti slenkančio vidurkio komponentės reikšmę, reikia ieškoti pirmos vėlavimo operatoriaus reikšmės, kuri statistiškai reikšmingai nekoreliuoja su dabartiniu laiko momentu.

Taigi iš grafiko galime matyti, kad vėlavimo operatorius 2 yra pirmasis, kuris statistiškai reikšmingai nebepriskluso nuo tuometinio momento, iš to galime daryti išvadą, kad mūsu seką galima modeliuoti su slenkančio vidurkio eile 1. Gauname MA(1).

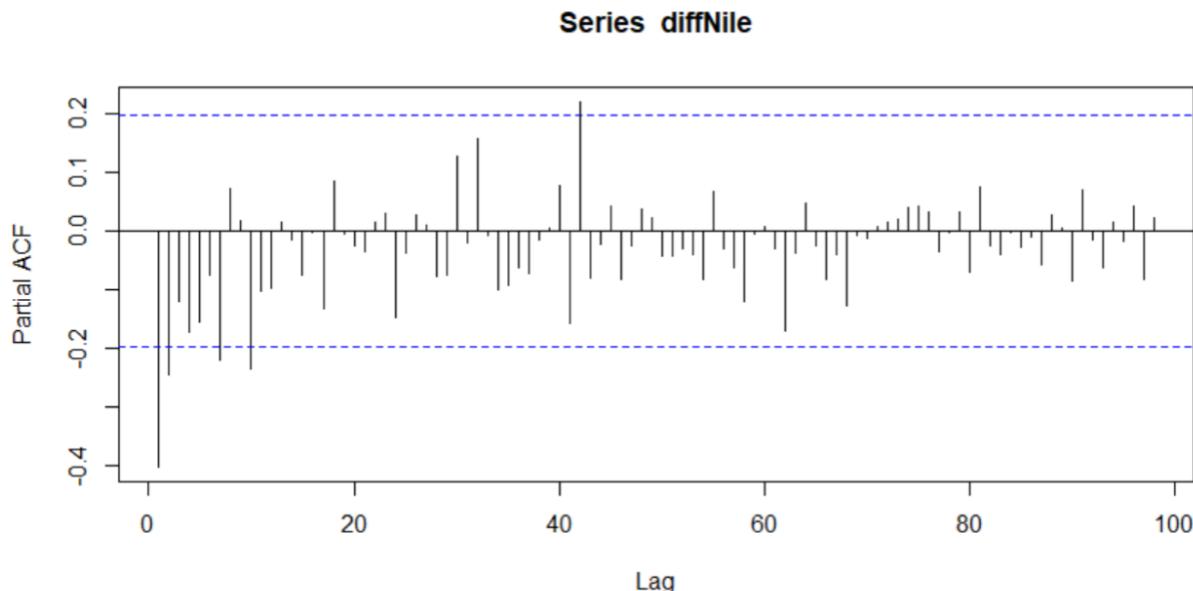
PACF

PACF - dalinė autokoreliacijos funkcija (angl. partial Autocorrelation function). Ši funkcija yra dalinė, nes priešingai negu ACF, ji matuoja tik koreliaciją tarp 2 laiko momentų, eliminuodama įtaką iš preities laiko momentų koreliacijai.

$$PACF(T_i, k) = \frac{\text{Covariance}([T_i | T_{(i-1)}, T_{(i-2)} \dots T_{(i-k+1)}], [T_{(i-k)} | T_{(i-1)}, T_{(i-2)} \dots T_{(i-k+1)}])}{\sigma_{[T_i | T_{(i-1)}, T_{(i-2)} \dots T_{(i-k+1)}]} \times \sigma_{[T_{(i-k)} | T_{(i-1)}, T_{(i-2)} \dots T_{(i-k+1)}]}}$$

Kur $T(i)$ yra laiko eilutė, o σ – standartinis nuokrypis

```
> pacf(diffNile, lag.max = 100)
```



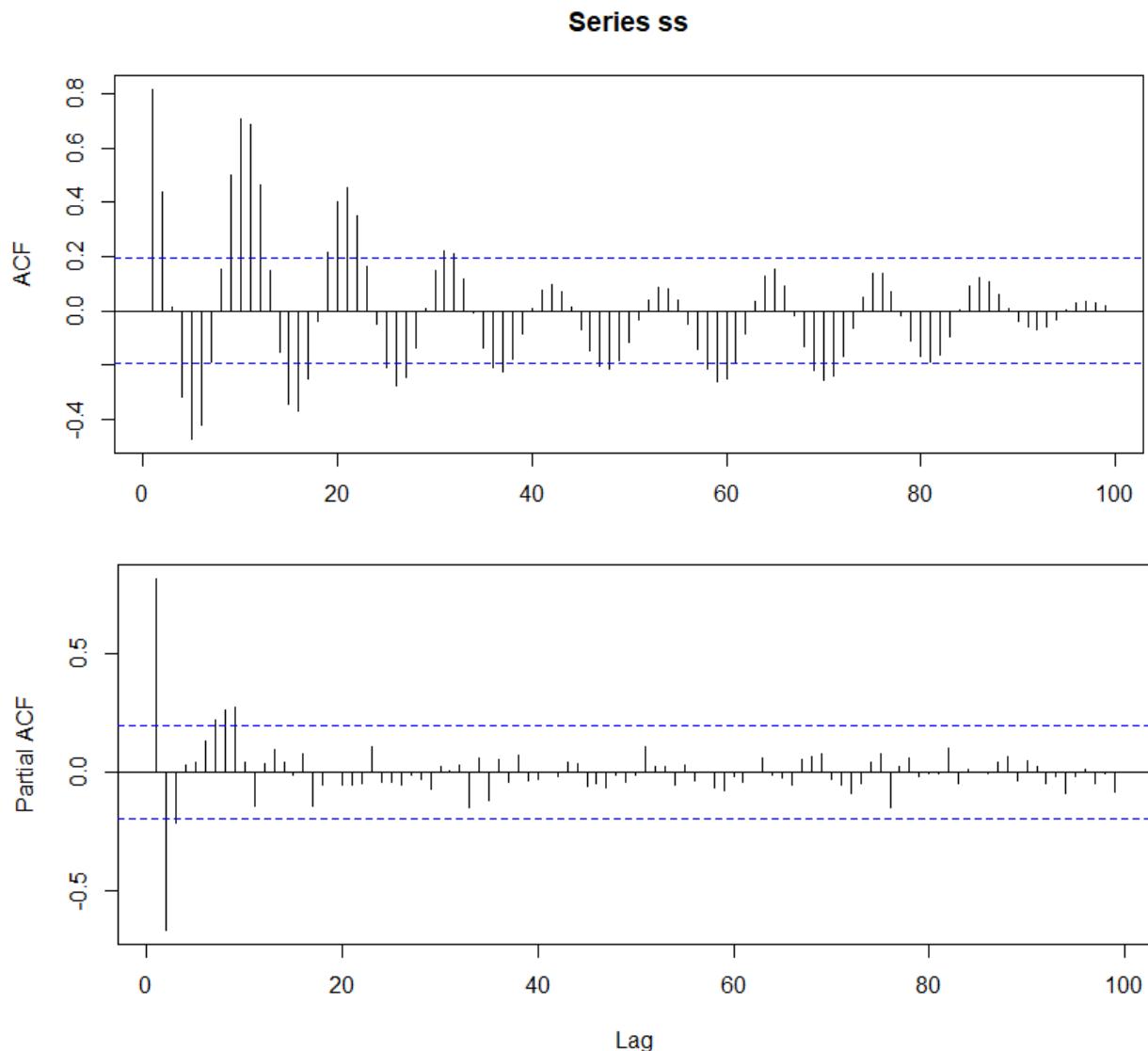
5 pav. PACF grafikas

PACF grafikas padeda nustatyti AR modelių eilę. Lygiai tokiu pačiu principu, kaip ACF, galime teigti, kad 3-asis vėlavimo operatorius yra lygus nuliui, todėl seką galime modeliuoti su AR(2).

ACF ir PACF

Bendruoju atveju reikėtų naudoti abu metodus. Pasinaudojus gautais rezultatais (4 ir 5 pav.) sekai modeliuoti turime kelis pasirinkimus: MA(1), AR(2) taip pat ir ARMA(1,1). Priimtina modeliuoti paprasčiausią įmanomą modelį. Šiuo atveju MA(1) turi vieną parametrumą, kaip kiti pasirinkimai turi po 2.

Atlikus tuos pačius veiksmus su duomenimis, kuriuose yra sezonas gaunama ryškesnė sinusoidė.



6 pav. Saulės dėmių duomenys. ACF ir PACF

Egzistuojant sezoniškumo komponentei ACF grafike koreliacijos nėra lygios nuliui i^* s vėlavimams, kai $i = 1, 2, 3, \dots$, o s yra sezonų periodas. Atsižvelgus į sezonų įtaką - prieš tai aptartos eilės įvertinimo taisyklės galioja lygiai taip pat.

CCF

CCF – kryžminė koreliacijos funkcija (angl. cross correlation function). **CORRELATION BETWEEN $x(t)$ and $y(t + h)$, kur h is in $[-\text{maxlag}; +\text{maxlag}]$**

Formulė:

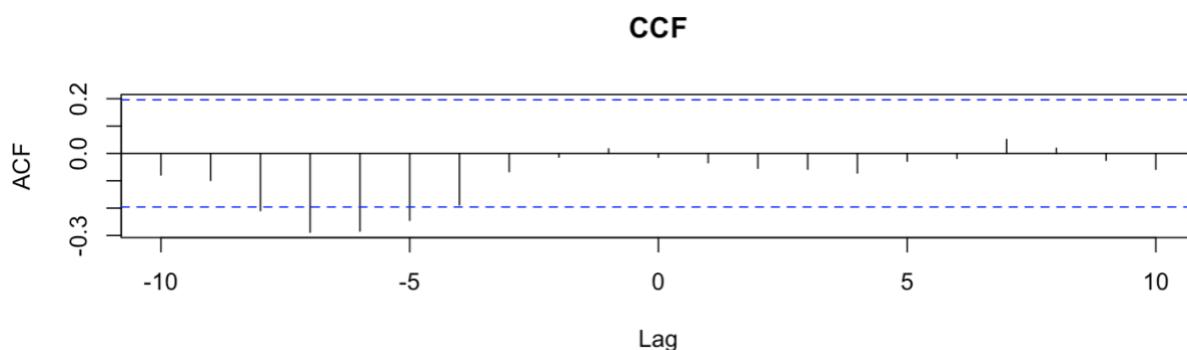
$$r = \frac{\sum_i [(x(i) - mx) * (y(i-d) - my)]}{\sqrt{\sum_i (x(i) - mx)^2} \sqrt{\sum_i (y(i-d) - my)^2}}$$

Kur $x(i)$ ir $y(i)$ laiko eilutės, mx ir my yra x ir y laiko eilučių vidurkiai, o d – vėlavimo operatorius.

```
> ccfval <- ccf(c(data), ss, lag.max = 10, main = "CCF")
> ccfval
```

Autocorrelations of series ‘X’, by lag

-10	-9	-8	-7	-6	-5	-4	-3	-2
-0.079	-0.099	-0.209	-0.288	-0.284	-0.245	-0.188	-0.067	-0.014
-1	0	1	2	3	4	5	6	7
0.017	-0.014	-0.034	-0.054	-0.058	-0.072	-0.029	-0.018	0.051
8	9	10						
0.019	-0.026	-0.058						



Matome kad su vėlavimo operatoriaus reikšmėmis -6 ir -7 gauname stipriausias koreliacijas.

EACF

EACF – išplėsta autokoreliacijos funkcija (angl. Extended AutoCorrelation Function). Tai funkcija padedanti surasti geriausias ARMA modelio p ir q reikšmės. Kadangi tikrinant ACF ir PACF funkcijas atskirai gaunami rezultatai apie AR ir MA yra irgi atskiri, tad EACF padeda modeliuoti juos kartu.

Vykdomas funkcijs reikia nurodyti laiko eilutę ir maksimalias AR ir MA modelių parametru reikšmės, tuomet kiekvienai kombinacijai bus parašoma ar autokoreliacija reikšminga su simboliumi „x“ ar nereikšminga su simboliu „o“.

```
> eacf(diffNile, 5, 5)
```

AR/MA

0 1 2 3 4 5

0 x o o o o o

1 x x o o o o

2 x o o o o o

3 x x x o o o

4 x x x o o o

5 x o x o x o

Gavus lentelę reikia ieškoti simboliu „o“ apsupty kuo daugiau „x“, taigi galimi variantai yra:

ARMA(0, 1)

ARMA(2, 1)

ARMA(5, 1)

Šaltiniai

http://web.vu.lt/mif/v.kvedaras/files/2013/09/Konspektas_2005.pdf

[Model Specification for Time Series \(sc.edu\)](#)

[R: The R Datasets Package \(ethz.ch\)](#)

[XploRe Help : eacf \(hu-berlin.de\)](#)

[Cross Correlation \(CCF\) Plots \(sigmaxl.com\)](#)

[Cross Correlation \(paulbourke.net\)](#)

[time series - Eacf table interpretation in R - Cross Validated \(stackexchange.com\)](#)