Úvod, základné pojmy, testovanie bieleho šumu

Beáta Stehlíková

Časové rady

Fakulta matematiky, fyziky a informatiky, UK v Bratislave

Analýza časových radov: úvod

Klasický vzorový príklad

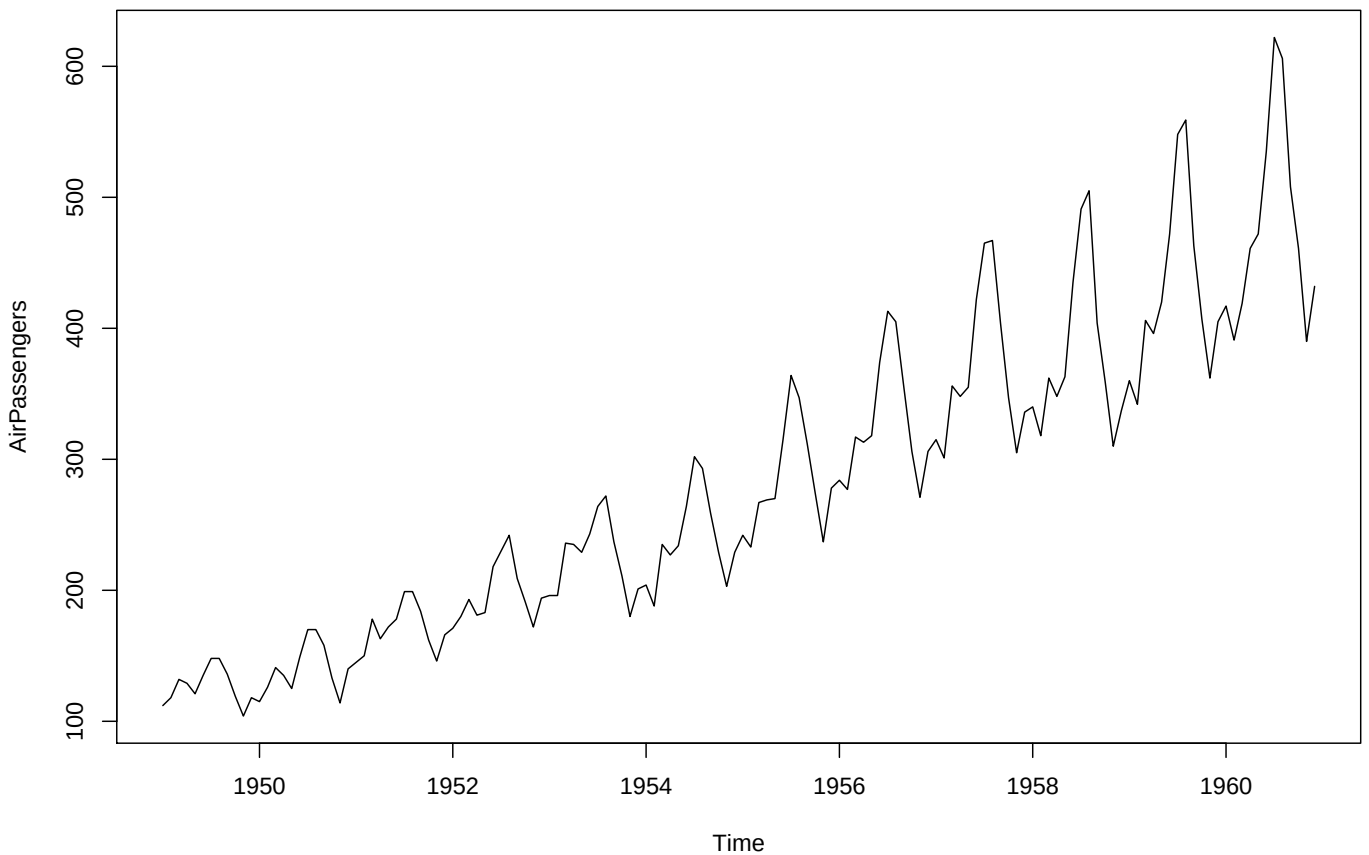
- Pozrieme sa na dáta - počty cestujúcich aerolinkami
- Dáta `AirPassengers` z balíka `datasets`
- Popis dát v dokumentácii (pomocou `?AirPassengers`):
  - The classic Box & Jenkins airline data. Monthly totals of international airline passengers, 1949 to 1960.
  - A monthly time series, in thousands.

```
library(datasets)
AirPassengers
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949 112 118 132 129 121 135 148 148 136 119 104 118
## 1950 115 126 141 135 125 149 170 170 158 133 114 140
## 1951 145 150 178 163 172 178 199 199 184 162 146 166
## 1952 171 180 193 181 183 218 230 242 209 191 172 194
## 1953 196 196 236 235 229 243 264 272 237 211 180 201
```

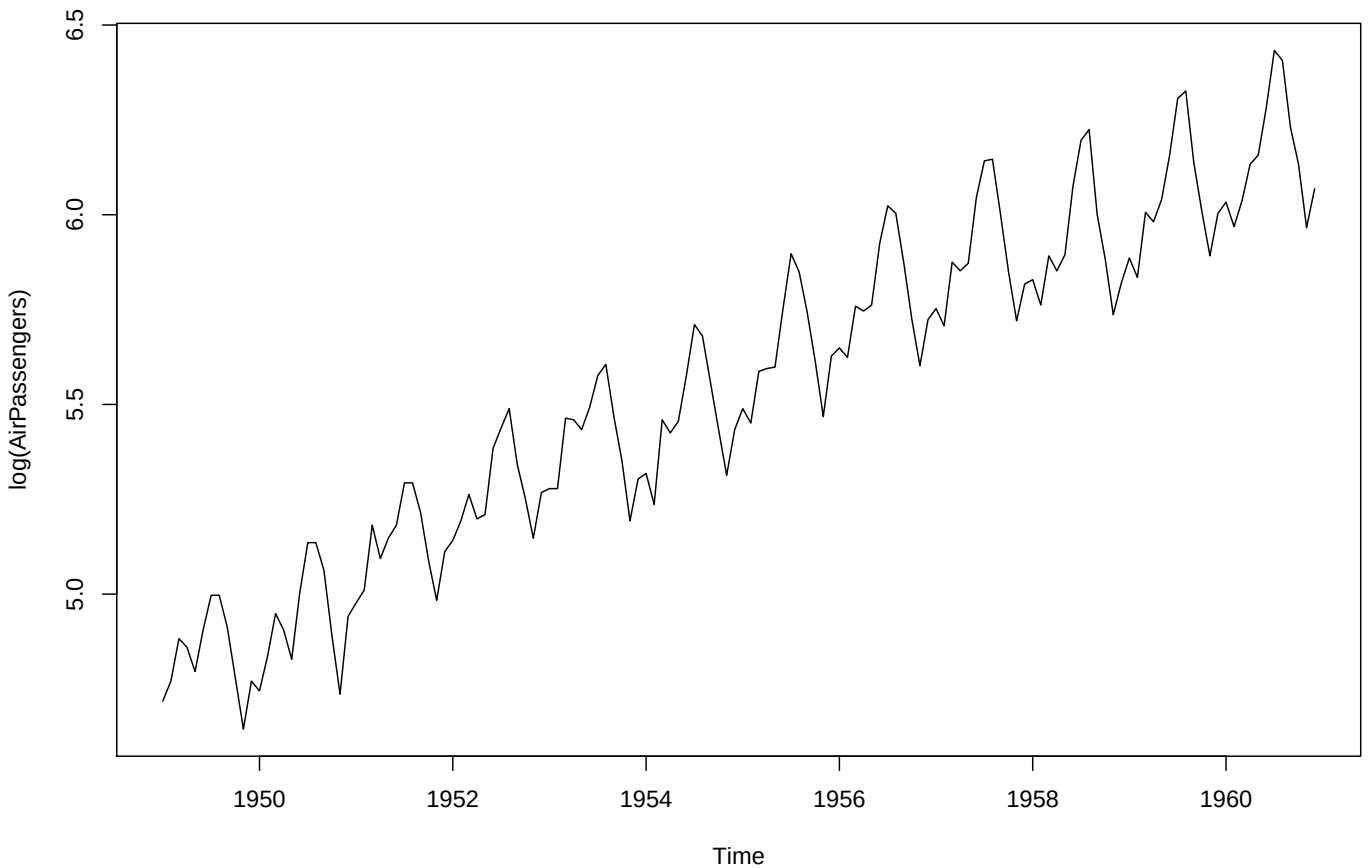
Klasický vzorový príklad - priebeh dát

Stačí `plot(AirPassengers)` a vďaka časovej štruktúre dát máme:



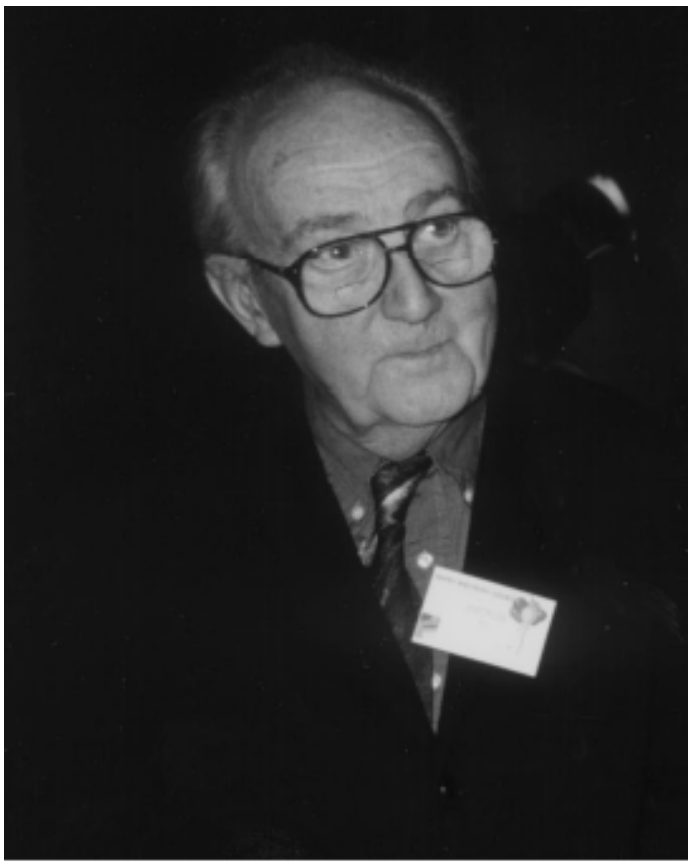
Klasický vzorový príklad - priebeh dát

Po zlogaritmovaní sa stabilizuje volatilita:



Box a Jenkins

Budeme sa zaoberať **prístupom od Boxa a Jenkinsa**.



Rozhovor s G. E. P. Boxom po oslave jeho 80. narodenín (1999), ako sa začal zaujímať o štatistiku a ďalšie otázky.

> *The first paper you wrote with Jenkins has been considered as a breakthrough in statistics.*

Peña, D. (2001). George Box: An interview with the International Journal of Forecasting. International Journal of Forecasting, 17(1), 1-9.

Informačný list predmetu

Výsledky vzdelávania:

Študenti bude vedieť modelovať jednorozmerné časové rady Box-Jenkinsovou metodológiou a bude poznať jej teoretické pozadie.

Základné pojmy

Obsah

- Časový rad, momenty
- Stacionarita, ergodicita
- Biely šum
- Autokorelačná funkcia
- Woldova reprezentácia
- Testy o autokorelačnej funkcii

Momenty časového radu

- Náhodný proces $x_1, x_2, \ldots, x_T$ je úplne charakterizovaný svojou T-rozmernou distribučnou funkciou
- Obvykle sa zameriavame na **prvé dva momenty**:
  - stredná hodnota $E(x_t)$
  - variancia $D(x_t)$
  - kovariancie $Cov(x_t, x_s)$, tzv. autokovariancie

Stacionarita a ergodicita

- Väčšinou máme len jeden časový rad - jednu realizáciu náhodného procesu $\rightarrow$ aby sa dala robiť štatistická inferencia, potrebujeme dodatočné predpoklady
- Napríklad: na to, aby sme odhadli strednú hodnotu, ... potrebujeme viac ako jednu realizáciu tejto náhodnej premennej
- Ergodický proces - výberové momenty počítané z časového radu s T pozorovaniami konvergujú pre $T \rightarrow \infty$ k zodpovedajúcim momentom
- Tento koncept má zmysel iba ak predpokladáme, že $E(x_t) = \mu$, $D(x_t) = \sigma^2, \ldots$ pre každé t

Stacionarita a ergodicita

- Silná stacionarita: združená distribučná funkcia sa nemení pri posune v čase
- Obvykle sa pracuje so slabším predpokladom → slabá stacionarita:

$$E(x_t) = \mu \ \forall t \quad (1)$$

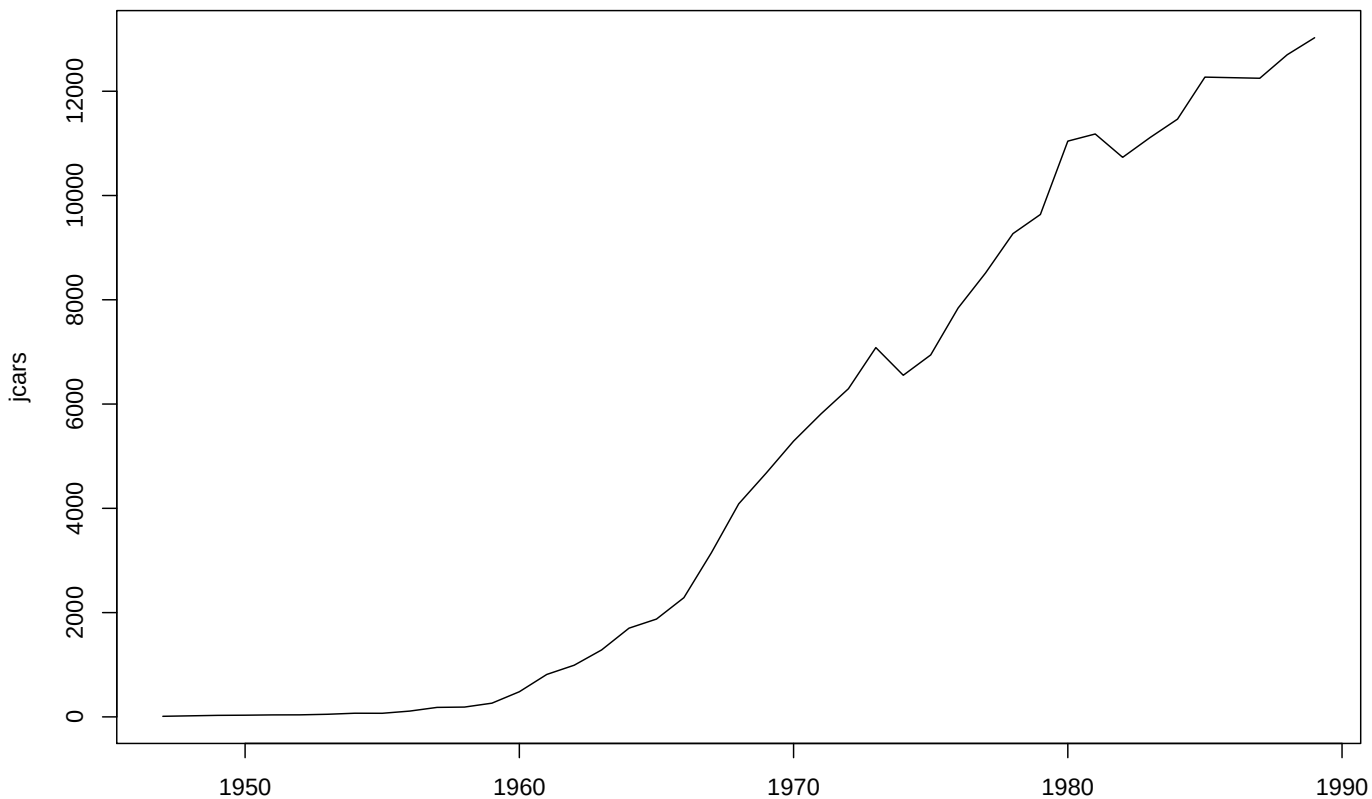
$$Cov(x_t, x_s) = \gamma(|t-s|) \ \forall t, s \quad (2)$$

- Z (2) vyplýva, že $D(x_t) = const.$ pre všetky t.
- Ďalej budeme pod stacionaritou rozumieť slabú stacionaritu.

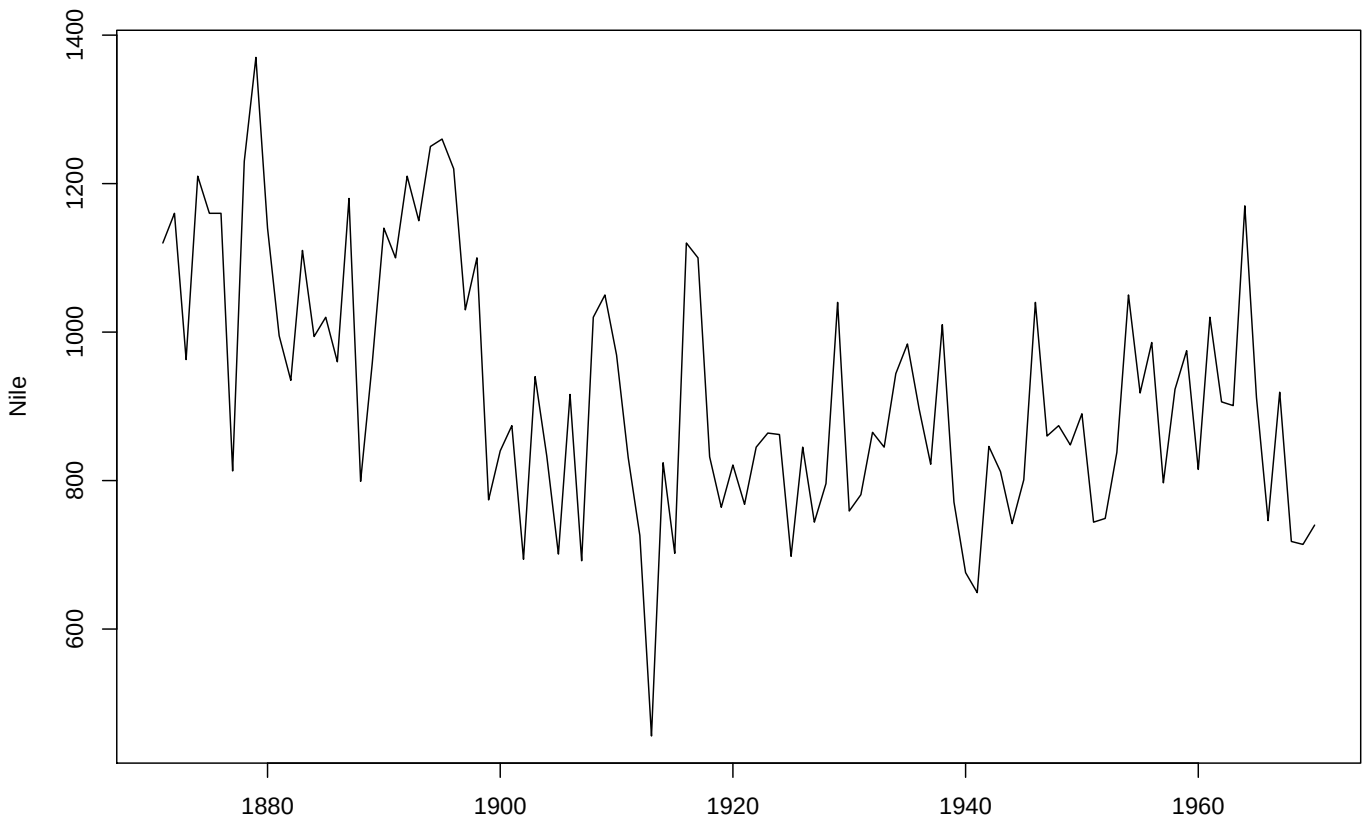
Stacionarita - dáta

- *Stacionárny časový rad*:
  - dáta sú priťahované k určitej rovnovážnej hodnote, okolo ktorej oscilujú
- *Nestacionárny časový rad*:
  - napríklad trend: rastúci trend $\rightarrow$ stredná hodnota nie je konštantná $\rightarrow$ proces nie je stacionárny
  - neskôr budeme vidieť aj iné druhy nestacionarity (napr. zatiaľ nejasne znejúci pojem *jednotkový koreň* v sylabe predmetu)

```
# PRIKLAD 1
# Japanese motor vehicle production in thousand (1947-1989)
library(fma)
plot(jcars)
```



```
# PRIKLAD 2
# Rocny prietok Nilu v Aswane v 10^8 m^3 (stavba priehrad)
library(datasets)
plot(Nile)
```

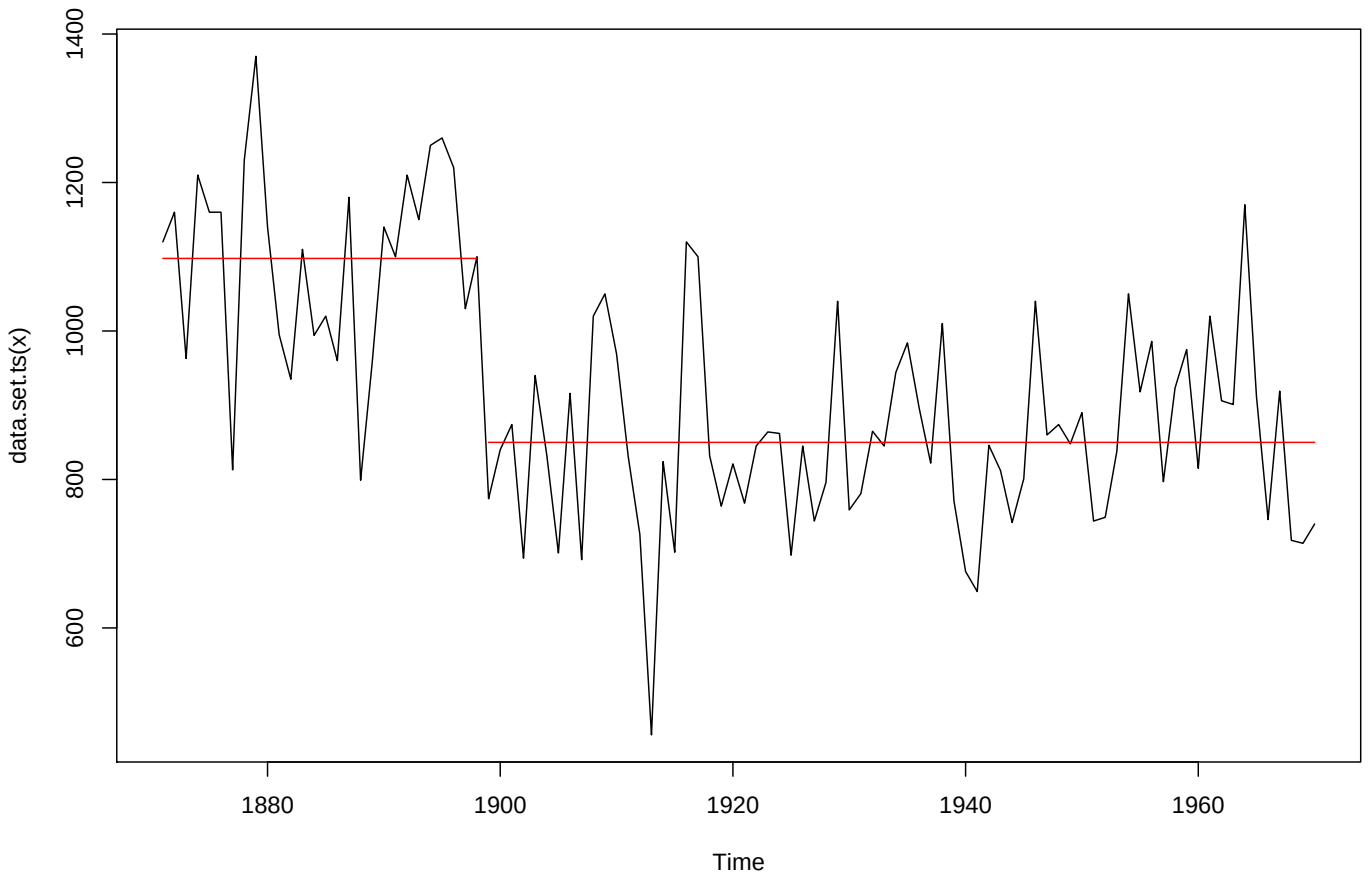


Nepovinné poznámky: Hľadanie bodov zmeny

- Tutoriál (useR! International R User 2017 Conference): https://www.youtube.com/watch?v=I7jUBro78RM
- Diplomovka Lucie Macháčkovej, zmeny v prietokoch slovenských riek (mEMM 2020, školiteľ: doc. Pekár)
- Článok s prehľadom rôznych metód: *Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. Knowledge and information systems, 51(2), 339-367.* https://link.springer.com/article/10.1007/s10115-016-0987-z

```
# pozrime sa na prietoky Nilu:
library(changepoint)
zmena <- cpt.mean(Nile)
plot(zmena)
```

Výstup z predchádzajúceho slajdu:



Biely šum

Definícia bieleho šumu

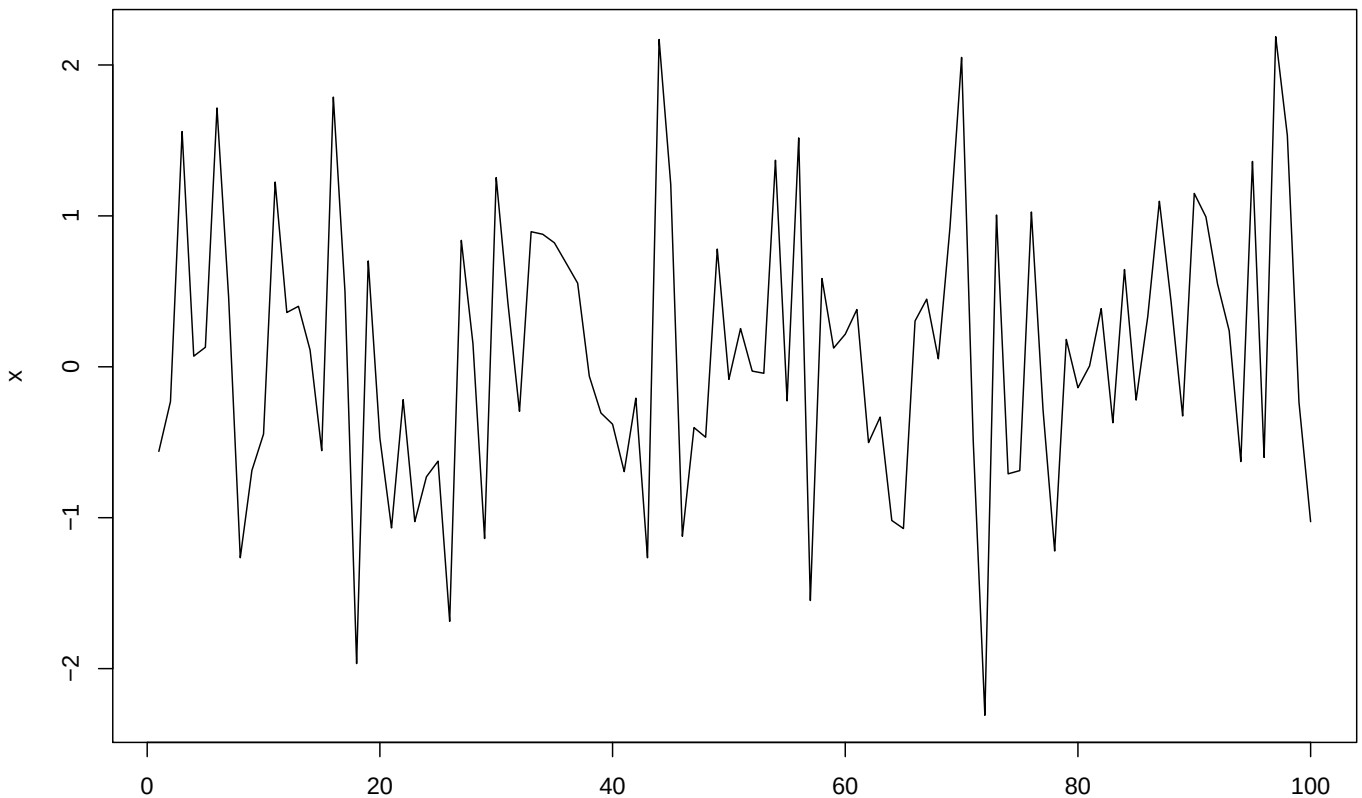
- Dôležitý príklad stacionárneho procesu, pomocou ktorého budeme definovať aj rôzne modely pre dáta
- Biely šum u_t je náhodný proces s nasledujúcimi vlastnosťami

$$\begin{aligned} \mathbb{E}(u_t) &= 0 \ \forall t \\ \mathbb{D}(u_t) &= \sigma^2 \ \forall t \\ \mathrm{Cov}(u_t, u_s) &= 0 \ \forall t \neq s \end{aligned}$$

- Napríklad postupnosť nezávislých náhodných premenných s rovnakým rozdelením (a konečnou strednou hodnotou a disperziou), ale nie je to jediná možnosť

Príklad

```
x <- rnorm(100)
plot(x, type = "l")
```



Príklady: zisťovanie stacionarity procesu

Príklad 1

- Nech u_t je biely šum, definujme

$$x_t = u_t + u_{t-1}$$

- Platí:

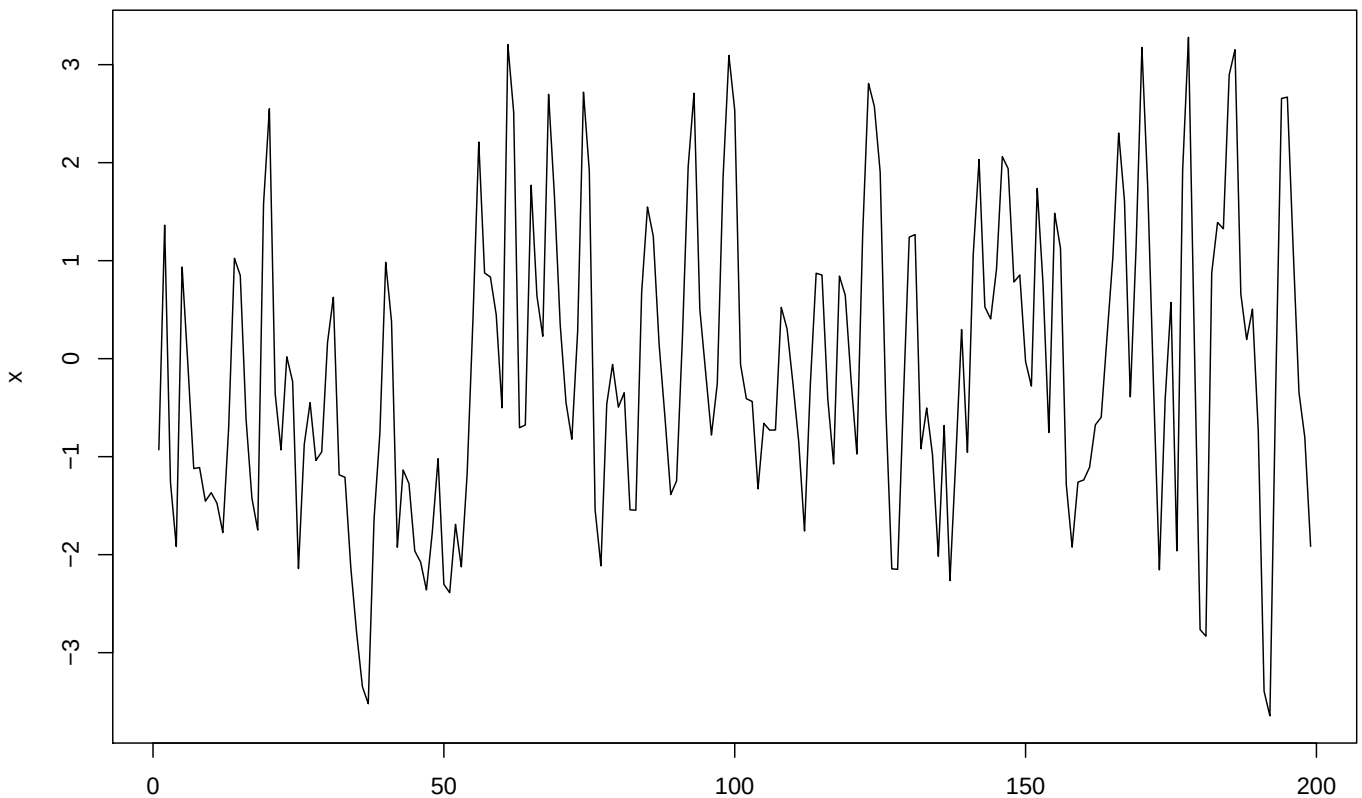
$$\mathbb{E}(x_t) = 0, \mathbb{D}(x_t) = 2\sigma^2$$

$$\mathrm{Cov}(x_t, x_{t+k}) = \begin{cases} \sigma^2 & \text{pre } k = 1, \\ 0 & \text{pre } k = 2, 3, \ldots \end{cases}$$

- Proces teda **je stacionárny**

Príklad 1: simulácia priebehu procesu, N = 200 pozorovaní

```
u <- rnorm(N + 1)
x <- u[2:N] + u[1:(N - 1)]
```



Príklad 2

- Nech u_t je biely šum, definujme

$$x_t = \begin{cases} u_1 & \text{pre } t = 1, \\ x_{t-1} + u_t & \text{pre } t = 2, 3, \ldots \end{cases}$$

- x_t sa dá zapísať v tvare $x_t = \sum_{i=1}^{t} u_i$
- Platí:

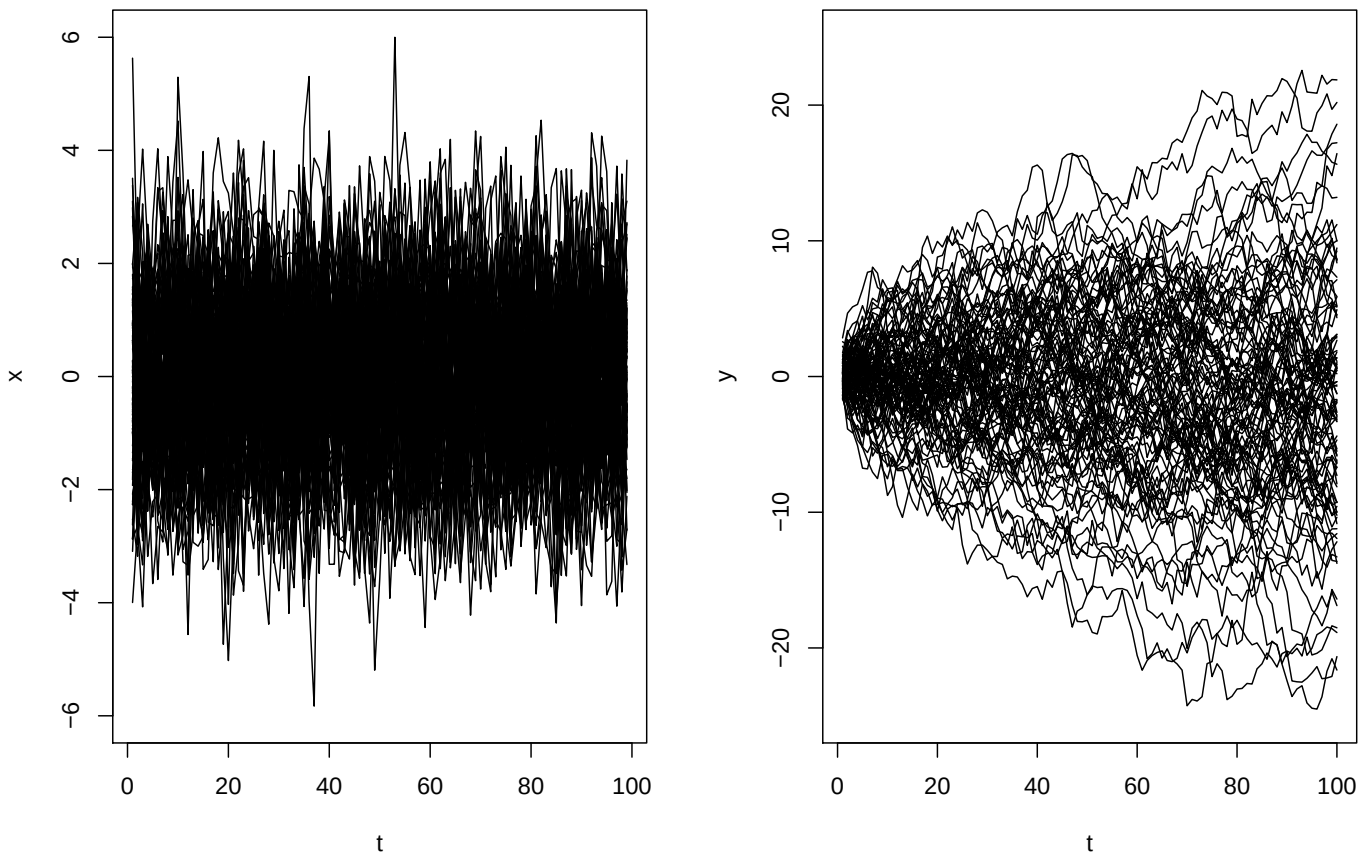
$$\mathbb{E}(x_t) = 0, \mathbb{D}(x_t) = t\sigma^2$$

$$\mathrm{Cov}(x_t, x_{t+k}) = t\sigma^2 \quad \text{pre } k > 0$$

- Proces teda **nie je stacionárny** (to vieme povedať už po výpočte disperzie)

Príklady 1, 2: porovnanie simulácií procesov

- Vľavo: stacionárny proces z príkladu 1, vpravo: proces s rastúcou disperziou z príkladu 2



Príklad 3

- Nech u_t je biely šum, definujme

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j u_{t-j}, \tag{3}$$

kde koeficienty ψ_j spĺňajú $\psi_0 = 1$, $\sum \psi_j^2 < \infty$

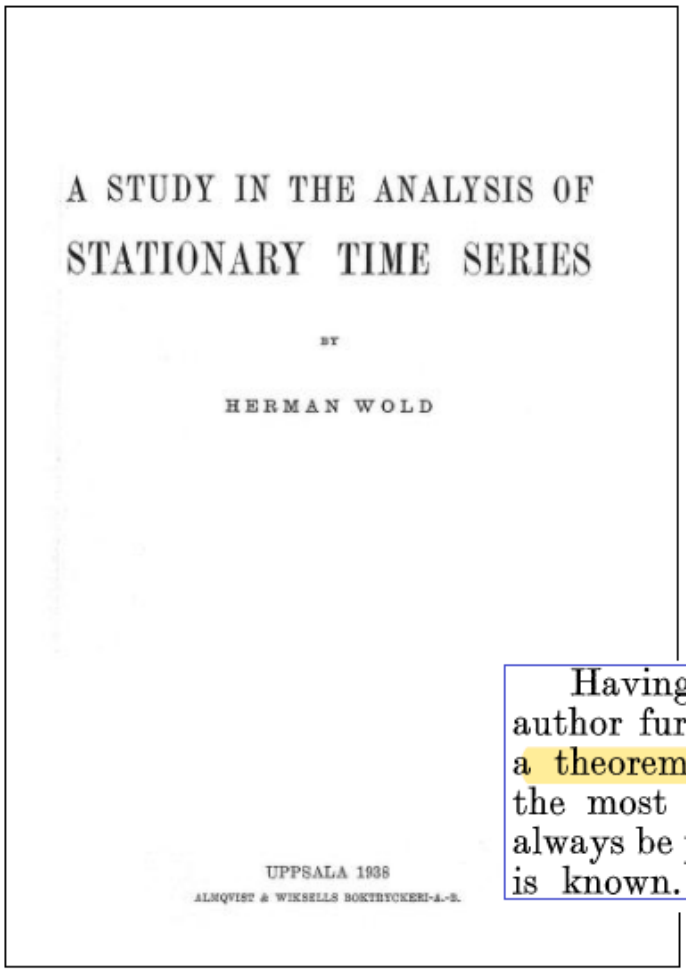
- Platí:

$$\mathbb{E}(x_t) = \mu, \mathbb{D}(x_t) = \sigma^2 \sum_{j=1}^{\infty} \psi_j^2$$

$$\mathrm{Cov}(x_t, x_{t+k}) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{k+j}$$

- Proces **je stacionárny**.

Woldova reprezentácia

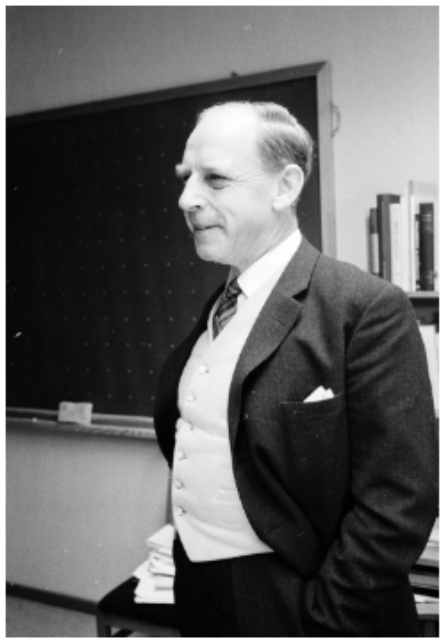


A STUDY IN THE ANALYSIS OF
STATIONARY TIME SERIES

BY

HERMAN WOLD

UPPSALA 1938
ALMQVIST & WIKSELLS BOKTRYCKERI-A.-B.



Having discussed a number of types of random processes, the author furnishes proofs of their various properties, and then gives a theorem of considerable interest, concerning the structure of the most general discrete stationary process. This, in fact, can always be presented as a sum of two components the nature of which is known.

J. N. (1939). Journal of the Royal Statistical Society, 102(2), 295-298.

https://archive.org/details/in.ernet.dli.2015.262214
https://www.jstor.org/stable/298000

https://digitaltmuseum.se/021016543711/professor-herman-wold-uppsala-1969

Woldova reprezentácia stacionárneho procesu

- V príklade 3: Proces tvaru (3) je stacionárny
- Dá sa dokázať: Každý stacionárny proces x_t sa dá zapísať v tvare

$$x_t = \mu_t + \sum_{j=0}^{\infty} \psi_j u_{t-j}$$

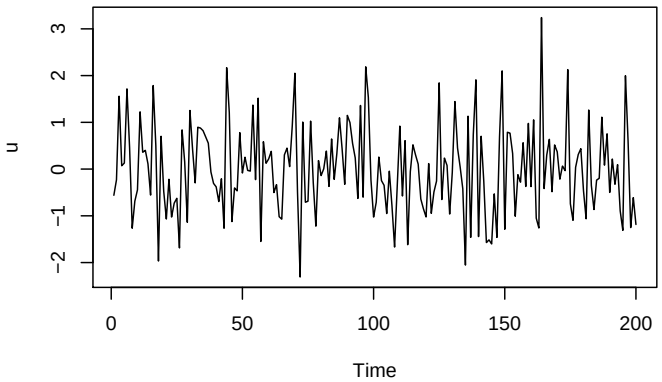
kde u je biely šum, $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ a μ_t sa dá presne predikovať z predchádzajúcich hodnôt procesu x (v našich aplikáciách to bude konštanta)

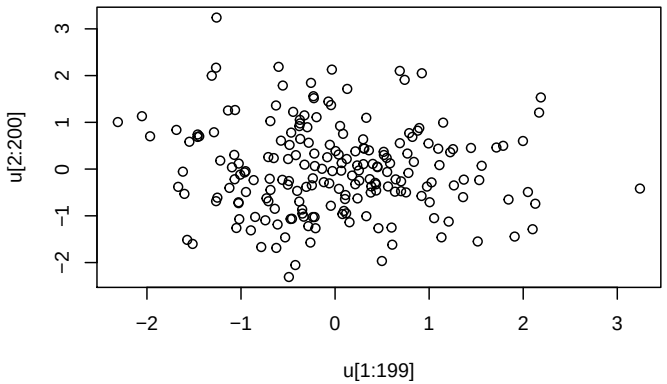
- Toto vyjadrenie sa nazýva Woldova reprezentácia

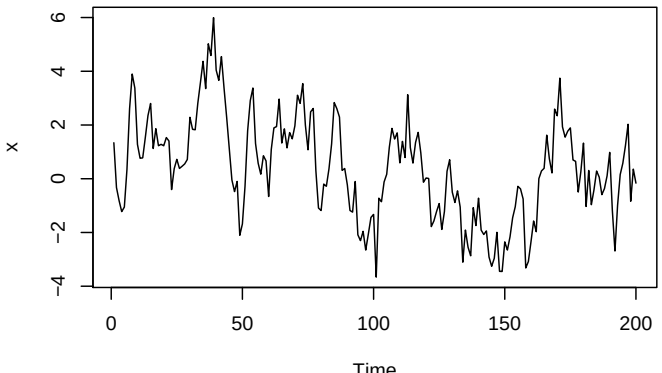
Autokorelačná funkcia (ACF)

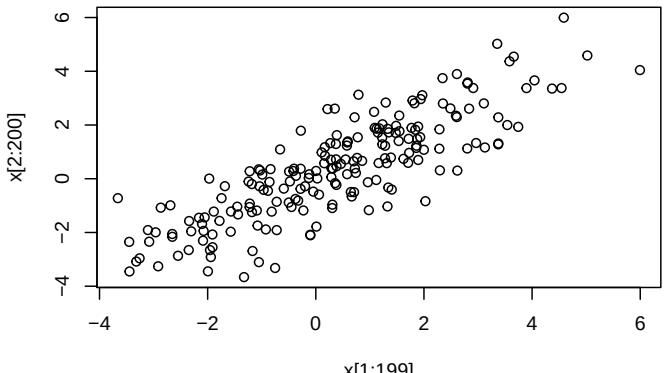
Motivácia

- Hore: u_t, dolu: $x_t = 0.9x_{t-1} + u_t$ - simulácia a závislosť x_t od x_{t-1}









Definícia základné vlastnosti

- Autokorelačná funkcia (ACF) stacionárneho procesu je definovaná ako

$$\rho(\tau) = \mathrm{cor}(x_t, x_{t+\tau}) = \frac{\mathrm{cov}(x_t, x_{t+\tau})}{\sqrt{\mathbb{D}(x_t)\mathbb{D}(x_{t+\tau})}}$$

- ACF sa teda dá vyjadriť pomocou autokovariančnej funkcie γ:

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)},$$

- Platí:

$$\rho(0) = 1, \rho(-\tau) = \rho(\tau),$$

stačí nám teda počítať $\rho(\tau)$ pre $\tau = 1, 2, \ldots$

Príklad

- Nech u_t je biely šum, definujme

$$x_t = u_t + u_{t-1}$$

- Pre tento proces sme odvodili stacionaritu a vlastnosti:

$$\mathbb{D}(x_t) = 2\sigma^2, \mathrm{Cov}(x_t, x_{t+k}) = \begin{cases} \sigma^2 & \text{pre } k = 1, \\ 0 & \text{pre } k = 2, 3, \ldots \end{cases}$$

- ACF teda je

$$\rho(k) = \mathrm{Cor}(x_t, x_{t+k}) = \begin{cases} 1/2 & \text{pre } k = 1, \\ 0 & \text{pre } k = 2, 3, \ldots \end{cases}$$

Odhadovanie ACF z dát

- Ergodický proces $\rightarrow$ stredná hodnota, disperzia a autokovariancie sa dajú konzistentne odhadnúť z dát $x_1, \ldots, x_T$:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \hat{\gamma}(0) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{\mu})^2$$

$$\hat{\gamma}(\tau) = \sum_{t=1}^{T-\tau}(x_t - \hat{\mu})(x_{t+\tau} - \hat{\mu})$$

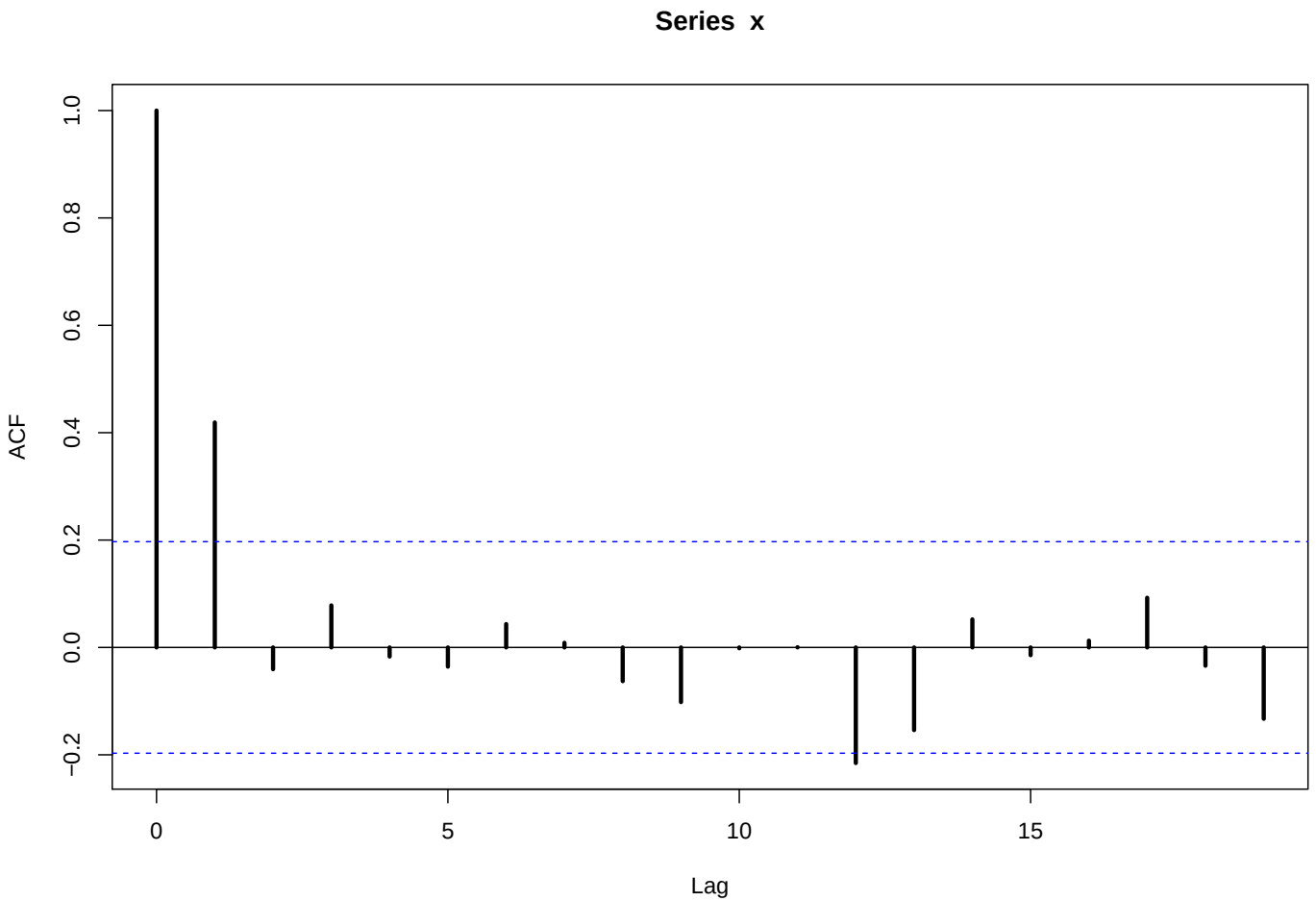
- Z toho - konzistentný odhad autokorelačnej funkcie:

$$\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)}$$

- je asymptoticky nevychýlený

Odhadovanie ACF z dát v R-ku: funkcia acf

```
acf(x) # pre data x <- u[2:N] + u[1:(N - 1)]
```



Využitie na kontrolu výpočtov

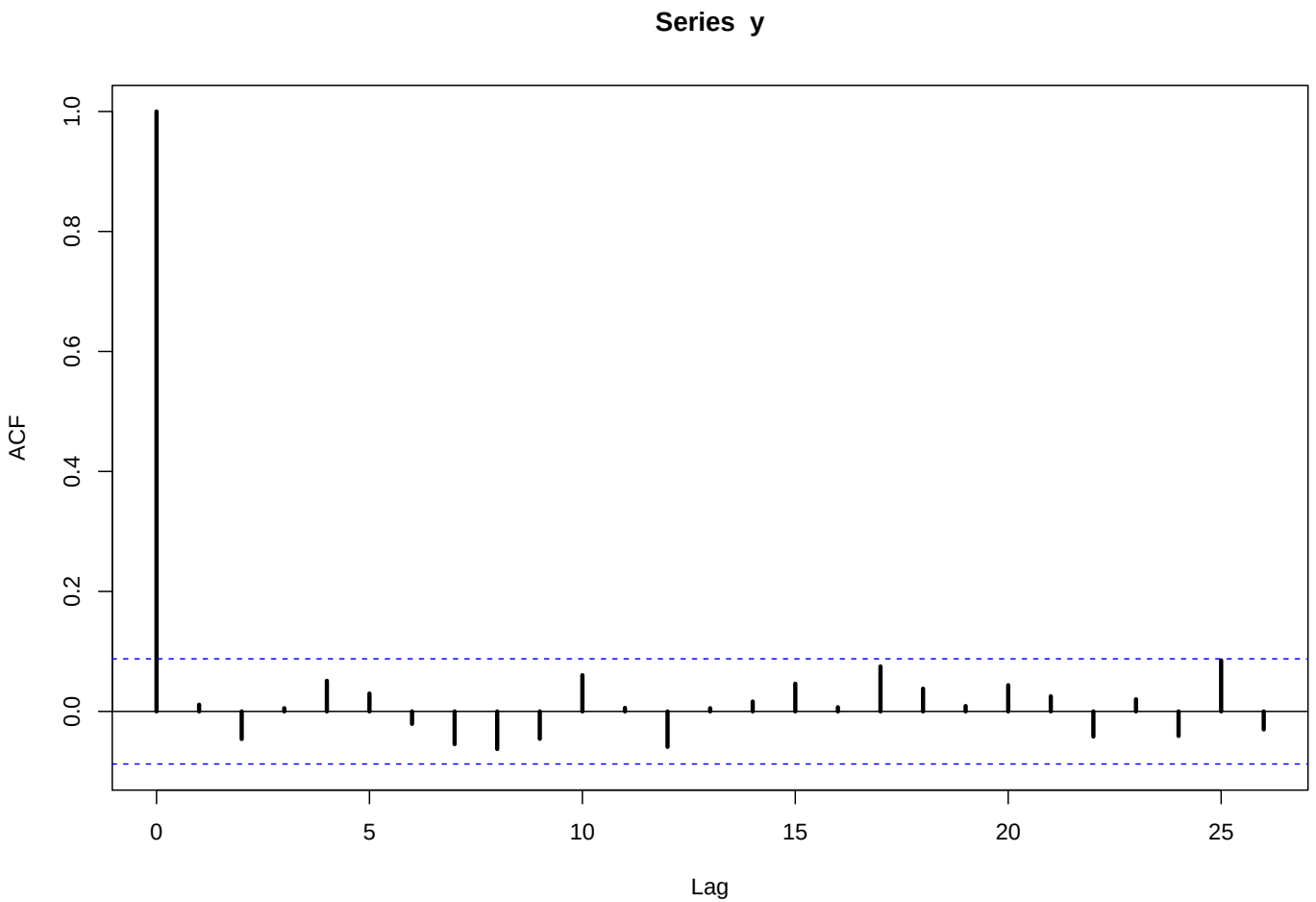
- *Zadanie:* Nech z_t je proces, ktorého hodnoty sú nezávislé náhodné premenné s rozdelením $N(0,1)$. Ukážte, že nasledujúci proces je stacionárny a vypočítajte jeho ACF:

$$y_t = \begin{cases} z_t & \text{pre } t \text{ nepárne} \\ \frac{1}{\sqrt{2}}(z_{t-1}^2 - 1) & \text{pre } t \text{ párne} \end{cases}$$

- Vygenerujeme si daný proces a zobrazíme odhadnutú ACF (náš výpočet by mal dať podobný výsledok):

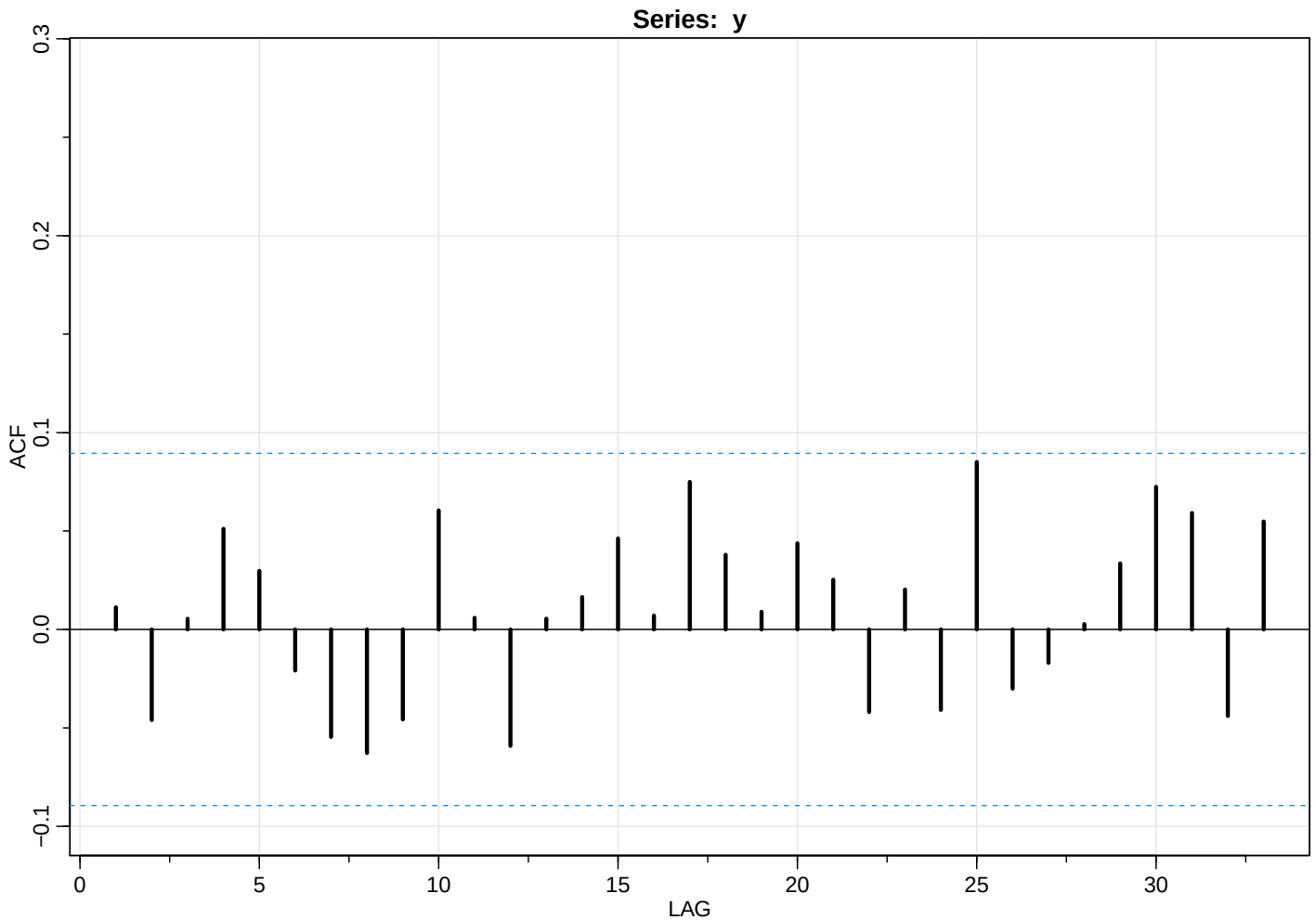
```
set.seed(1234)
N <- 500
z <- rnorm(N, mean = 0, sd = 1)
y <- z  # pre neparne indexy zostane, parne upravime
ind.parne <- seq(from = 2, to = N, by = 2)
y[ind.parne] <- (1/sqrt(2)) * (z[ind.parne - 1]^2 - 1)
```

```
acf(y)
```



Alternatíva: funkcia acf1 z balíka astsa

```
library(astsa); acf1(y)
```



Testovanie nulovosti autokorelácií

Testovanie nulovosti autokorelácií - každej samostatne

- Biely šum má nulovú ACF $\Rightarrow$ pri testovaní, či sú dáta bielym šumom, budeme testovať, či majú nulové autokorelácie
- Odhad ACF v prípade bieleho šumu
  - asymptoticky nevychýlený
  - disperzia $\approx 1/T$
  - $\Rightarrow$ približný 95 % interval spoľahlivosti: $\pm 1.96/\sqrt{T}$, resp. $\pm 2/\sqrt{T}$ - často sa zobrazuje spolu s odhadnutými autokoreláciami (aj v prípade funkcií `acf` a `acf1`)
- Pre každú autokoreláciu samostatne:
  - Testujeme, či sa rovná nule.
  - Nulovú hypotézu zamietame, ak je jej odhad mimo intervalu spoľahlivosti

Ak testujeme nulovosť autokorelácie, nie pre biely šum:

- V prípade procesu, pre ktorý platí $\rho(\tau) = 0$ pre $\tau > k$, pre tieto τ platí

$$\mathbb{D}(\hat{\rho}(\tau)) \approx \frac{1}{T}\left(1 + 2\sum_{j=1}^{k}\hat{\rho}(j)^2\right)$$

Testovanie nulovosti autokorelácií - Ljung-Boxov test

- Netestujeme nulovosť každej autokorelácie samostatne, ale testujeme hypotézu $\rho(1) = \rho(2) = \cdots = \rho(m) = 0$
- Box & Pierce, 1970: ak platí H_0, asymptoticky

$$Q = \sum_{j=1}^{m} \left(\frac{\hat{\rho}(j)}{1/\sqrt{T}} \right)^2 = T \sum_{j=1}^{m} \hat{\rho}(j)^2 \sim \chi^2_m$$

- Ljung & Box, 1978: modifikácia s lepšími vlastnosťami pri menšom počte dát

$$Q = T \sum_{j=1}^{m} \frac{T+2}{T-j} \hat{\rho}(j)^2 \sim \chi^2_m$$

- *Poznámka, ktorú využijeme neskôr: Počet stupňov voľnosti sa zmení, ak ide o rezíduá z modelu*

Ljung-Boxov test v R-ku: funkcia Box.test

- Testujme pre dáta `x`, že prvé štyri autokorelácie sú nulové

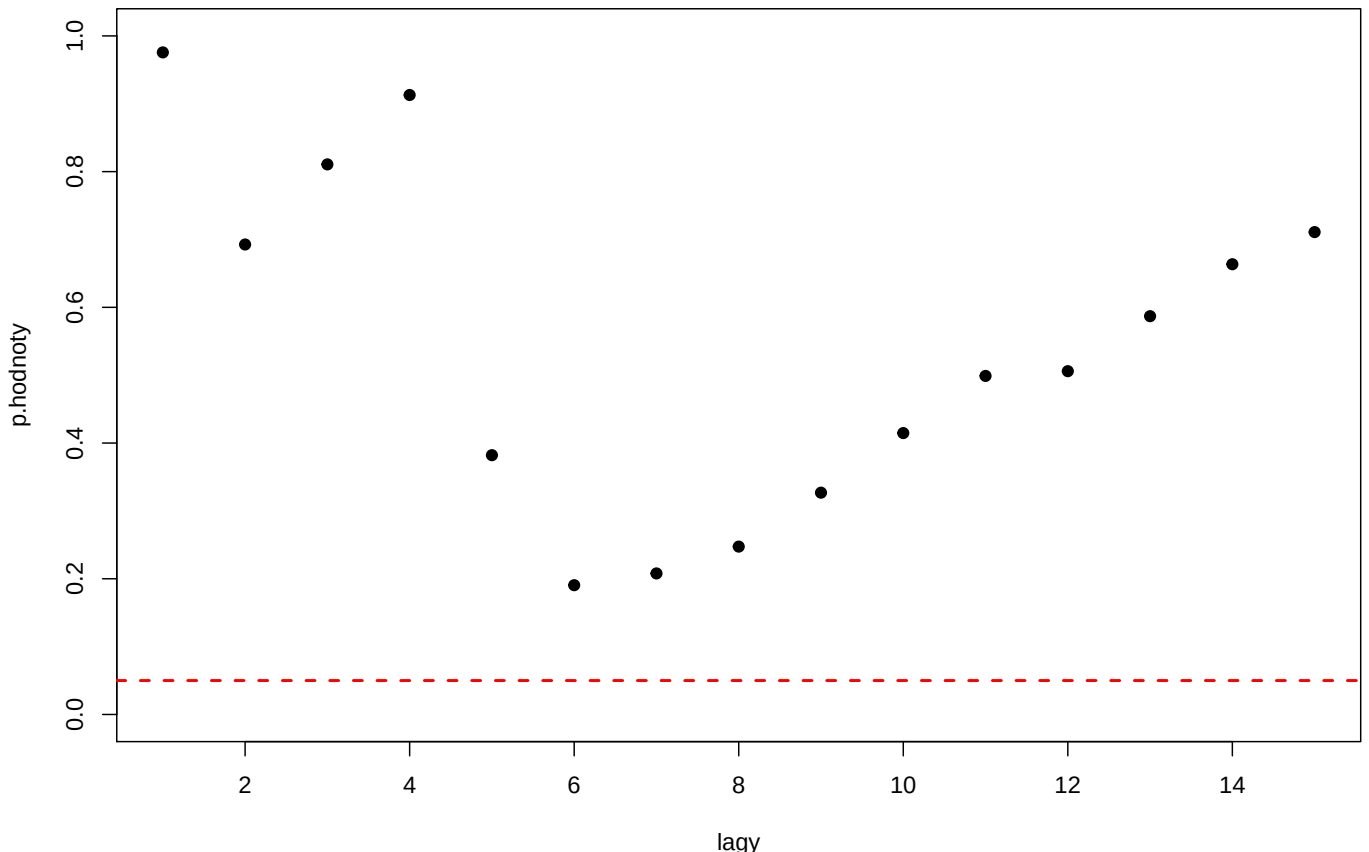
```
set.seed(12345)
N <- 500
x <- rnorm(N)

Box.test(x, lag = 4, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  x
## X-squared = 0.97917, df = 4, p-value = 0.9129
```

Ljung-Boxov test v R-ku: p-hodnoty

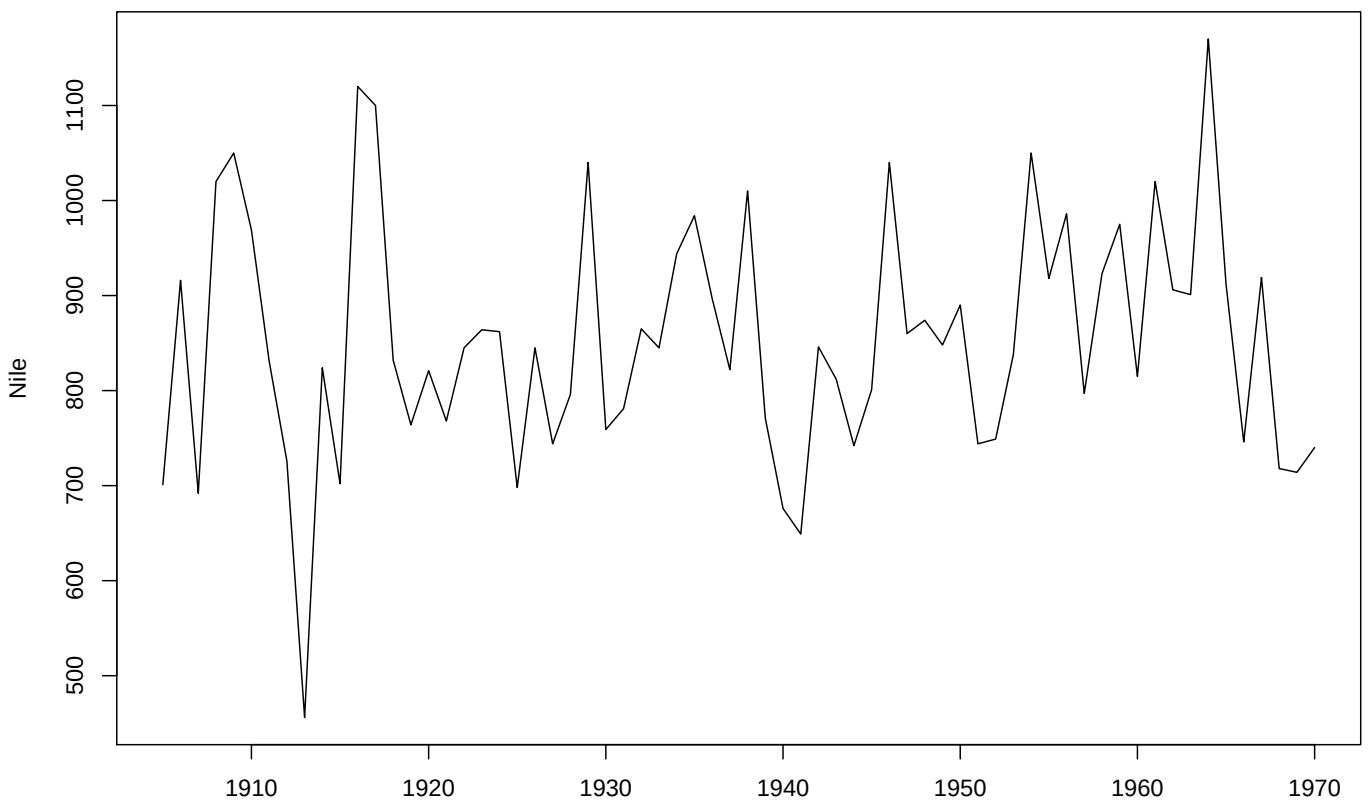
- `Box.test(x, lag = 3, type = "Ljung")$p.value`
- Grafický výstup s p-hodnotami, parameter `lag` od 1 do 15:



Použitie ADF testu

- Testujeme, či je zadaný časový rad biely šum, napríklad:

```
x <- window(Nile, start = 1905)
plot(x, ylab = "Nile")
```



- Rezíduá modelu majú byť bielym šumom → LB test rezíduí testuje, či je model vyhovujúci
- *D. Rawat et al. (2022). Modeling of rainfall time series using NAR and ARIMA model over western Himalaya, India. Arabian Journal of Geosciences 15(23), 1696.*

The best ARIMA model for Jammu and Kashmir data series is depicted in Table 6. For the state of Jammu and Kashmir, the ARIMA (1, 1, 1) model performed best for most of the series, except for June, August, September, December, and MAM data series. The best model for the series June, August, and MAM is ARIMA (2, 1, 1) and for September and December is ARIMA (3, 1, 1). For checking the autocorrelation of residuals obtained from the best ARIMA model, the Ljung–Box Q test was applied, which clearly pointed toward the June, August, and December series having non-zero autocorrelation as the p value is less than 0.05. From the Jarque–Bera test, most of the series

Month/season/year	Model	Box–Ljung test (p-value)
Jan	ARIMA (1, 1, 1)	0.82
Feb	ARIMA (1, 1, 1)	0.44
Mar	ARIMA (1, 1, 1)	0.81
April	ARIMA (1, 1, 1)	0.98
May	ARIMA (1, 1, 1)	0.97
June	ARIMA (2, 1, 1)	0.02
July	ARIMA (1, 1, 1)	0.94
Aug	ARIMA (2, 1, 1)	0.04
Sep	ARIMA (3, 1, 1)	0.99
Oct	ARIMA (1, 1, 1)	1.00
Nov	ARIMA (1, 1, 1)	1.00
Dec	ARIMA (3, 1, 1)	0.02
JF	ARIMA (1, 1, 1)	0.40
MAM	ARIMA (2, 1, 1)	0.69
JJAS	ARIMA (1, 1, 1)	0.60
OND	ARIMA (1, 1, 1)	0.85
Annual	ARIMA (1, 1, 1)	0.70