# wprowadzenie do analizy szeregów czasowych

Patryk Filipiak

Instytut Informatyki, Uniwersytet Wrocławski

19 stycznia 2016

# Definicje

**Szeregiem czasowym** nazywamy ciąg $(x_1, x_2, \ldots, x_t, \ldots)$ następujących po sobie obserwacji pewnego zjawiska.

- charakter dyskretny – kolejne obserwacje w identycznych odstępach czasowych,
- formalnie jest realizacją pewnego **procesu stochastycznego** (tj. rodziny zmiennych losowych) postaci $X = (X_t)_{t \in T}$.

$$
\begin{array}{ccccccc}
X_1 & \longleftrightarrow & x_1, & x_1', & x_1'', & \ldots \\
X_2 & \longleftrightarrow & x_2, & x_2', & x_2'', & \ldots \\
\vdots & & \vdots & \vdots & \vdots & \\
X_t & \longleftrightarrow & x_t, & x_t', & x_t'', & \ldots \\
\vdots & & \vdots & \vdots & \vdots &
\end{array}
$$

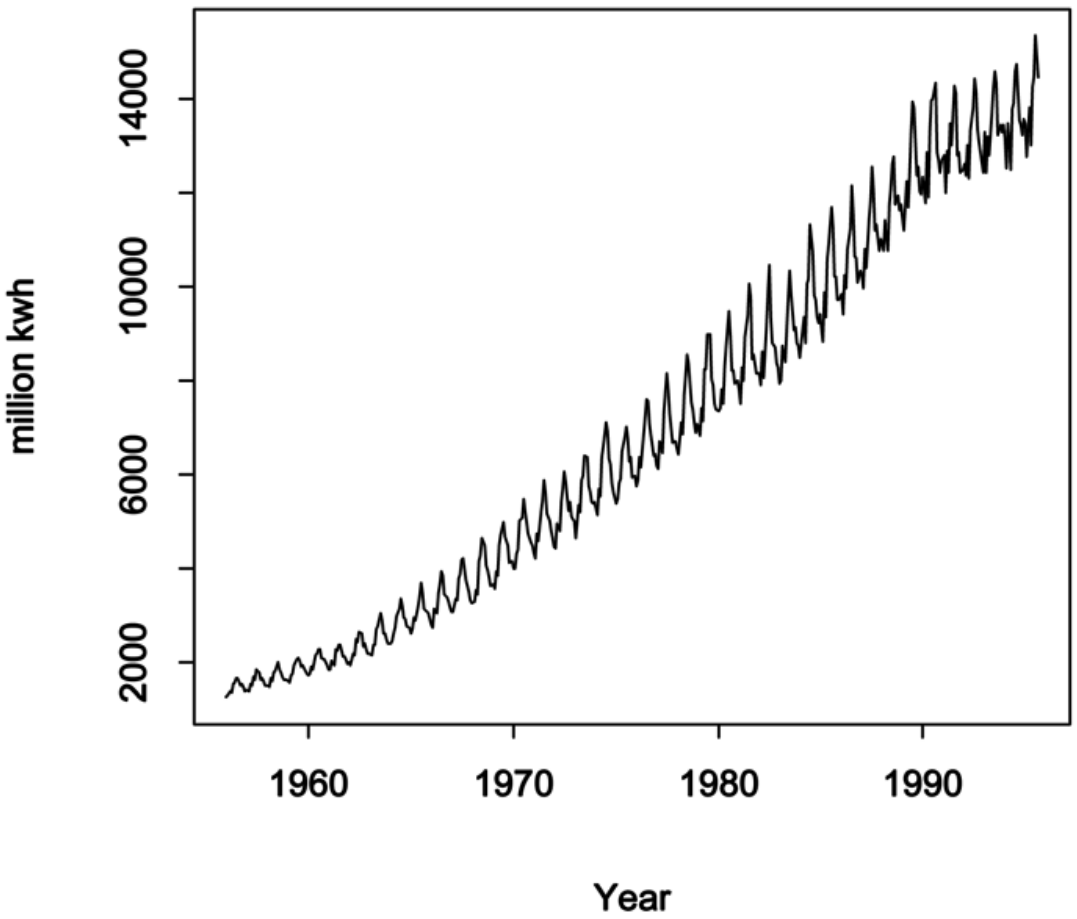
# Przykłady

## Wykres

miesięczne zestawienie ilości wytworzonej energii elektrycznej dla Australii

## Zjawiska

- trend,
- sezonowość,
- wysoka przewidywalność.



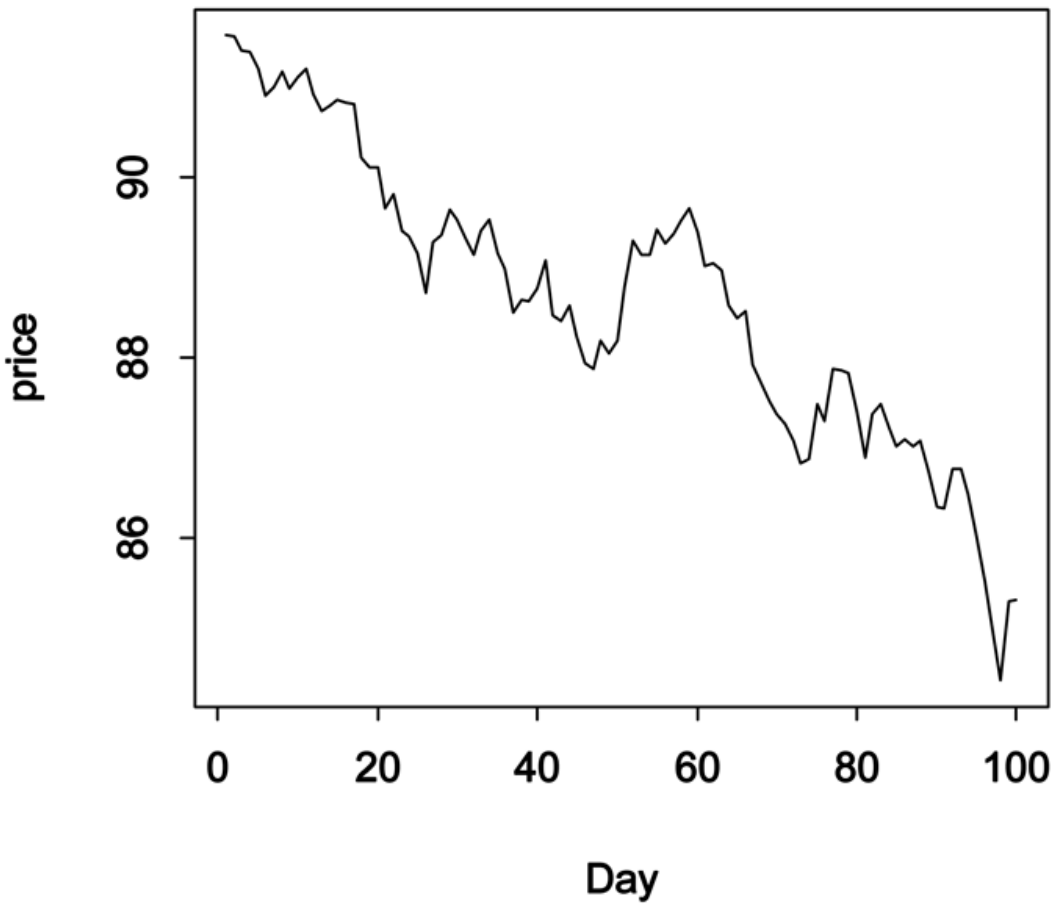
# Przykłady

## Wykres

notowania bonów skarbowych USA w ciągu 100 kolejnych dni roboczych 1981r.

## Zjawiska

- trend albo „chwilowy” spadek,
- ryzyko dużego błędu prognozy.



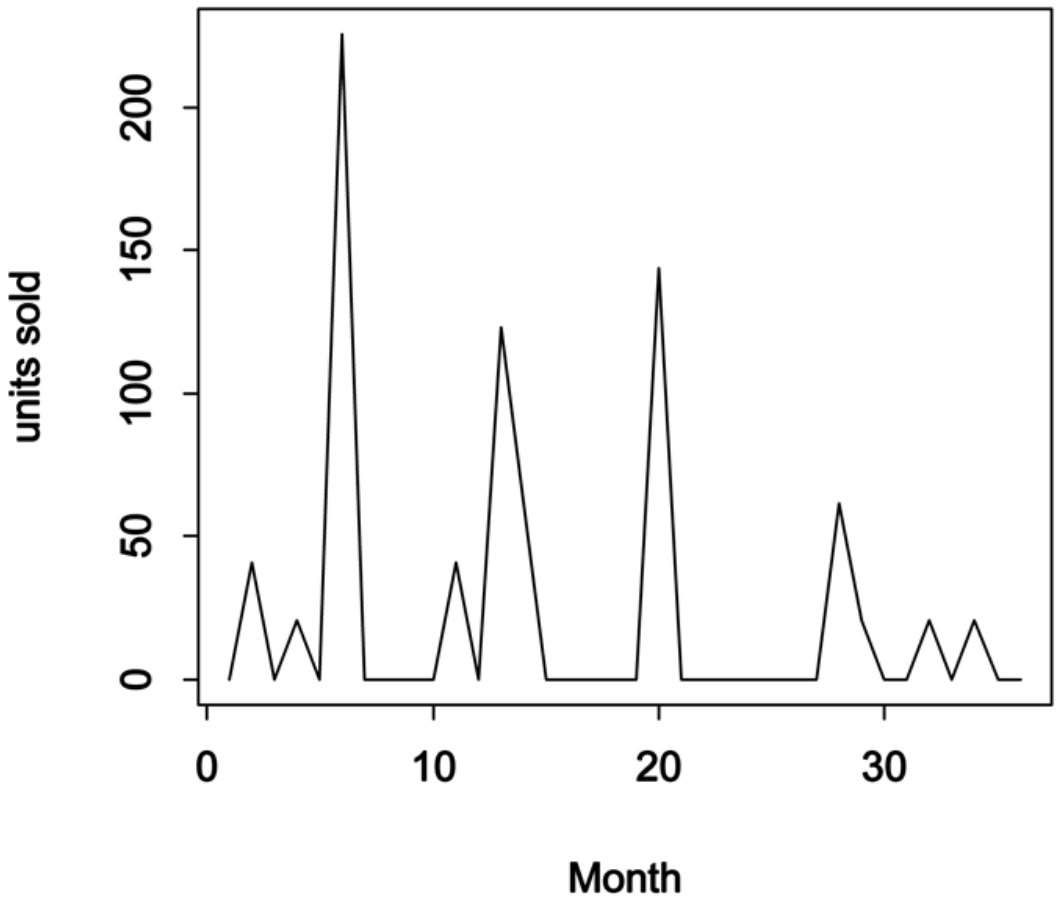
# Przykłady

## Wykres

sprzedaż produktu „C” notowana w dużej spółce naftowej

## Zjawiska

- *spikes*,
- predykcja wymaga analizy jakościowej.



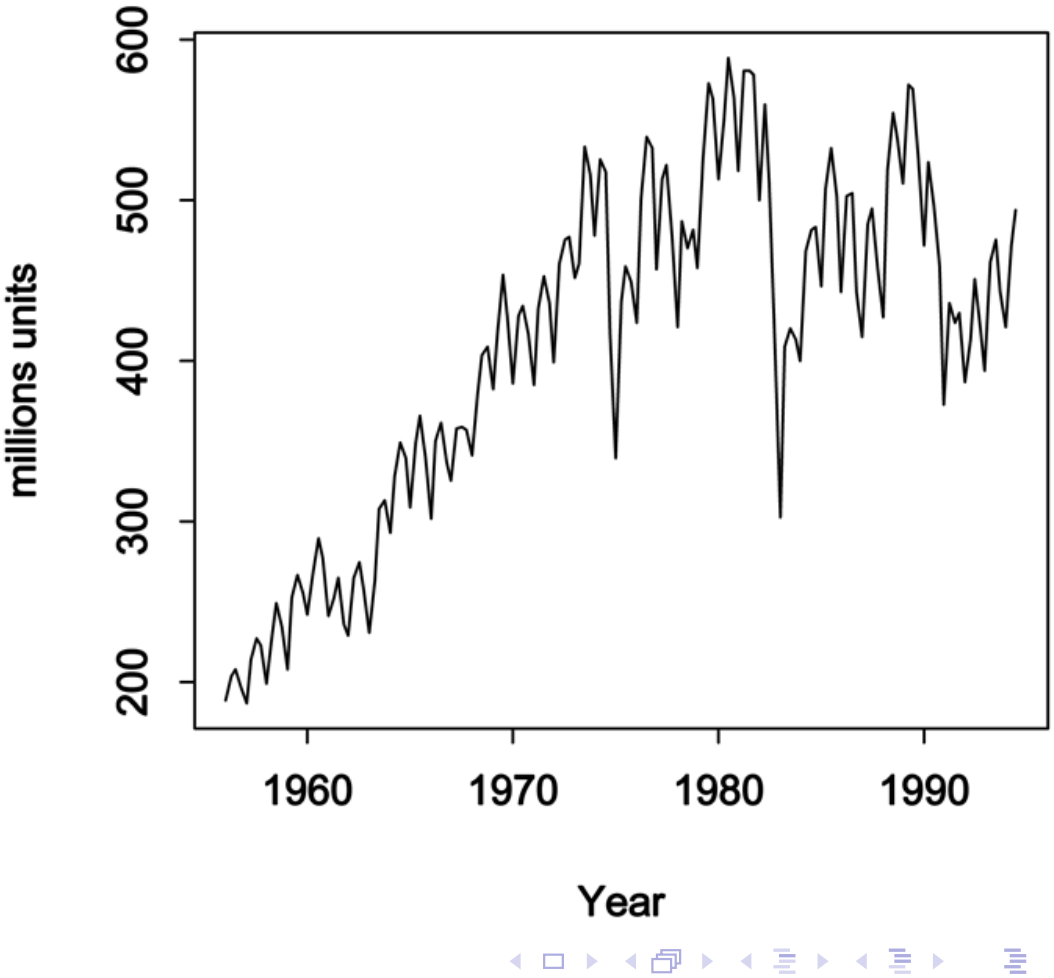
# Przykłady

## Wykres

miesięczne zestawienie produkcji gliny ceglarskiej dla Australii

## Zjawiska

- silne wahania,
- sezonowość o nieregularnej intensywności,
- predykcja jedynie po ustaleniu przyczyn fluktuacji.



# Zarys problematyki

## Cel

Dysponując historycznymi obserwacjami pewnego zjawiska $(x_1, x_2, \ldots, x_n)$, zrozumieć jego charakter i/lub przewidzieć przyszłe wartości, tj. $x_{n+1}, x_{n+2}, \ldots, x_{n+m}$ dla $m > 0$.

## Założenia

- Istnieje proces stochastyczny $(X_t)_{t \in T}$, którego instancją jest ciąg $(x_1, x_2, \ldots, x_n, \ldots)$.
- Na podstawie obserwacji $(x_1, x_2, \ldots, x_n)$ można estymować parametry rozkładów zmiennych losowych $X_1, X_2, \ldots, X_n$.
- Autokorelacja w procesie $(X_t)_{t \in T}$ pozwala na jakiekolwiek przewidywania na temat zmiennych $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$.

# Zarys problematyki

## Metody

- prezentacja danych – wykresy, statystyki,
- estymacja parametrów rozkładów $X_1, X_2, \ldots, X_n$ – wartości oczekiwanych, (auto)kowariancji, (auto)korelacji i innych,
- dobór i konstrukcja modelu $(X_t)_{t \in T}$,
- dekompozycja modelu,
- testowanie modelu – szacowanie błędów, dopasowania, etc.
- zastosowanie mechanizmu predykcji dla modelu – autoregresja, średnie kroczące i inne.

# Proste statystyki

Niech $(x_1, x_2, \ldots, x_n)$ będzie szeregiem czasowym.

- **średnia arytmetyczna**

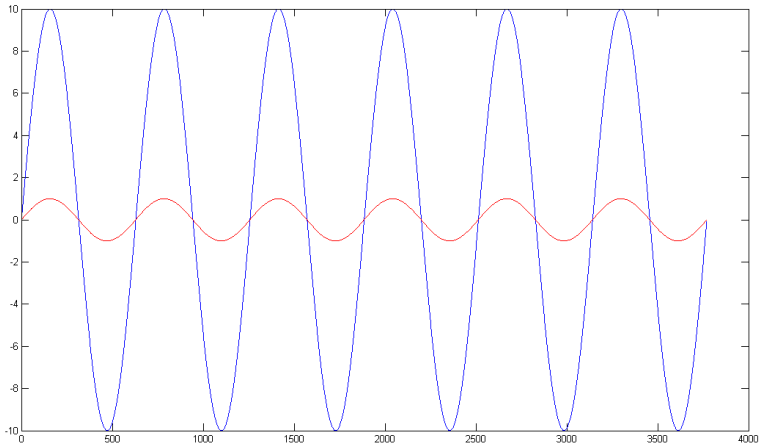
$$\overline{x} = \sum_{i=1}^{n} x_i$$

- **wariancja** (z próby)

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- **odchylenie standardowe**

$$s = \sqrt{s^2}$$




$\overline{x} = 0,\ s^2 \approx 50$

$\overline{x} = 0,\ s^2 \approx 0.5$

# Proste statystyki

Niech $1 \leqslant k < n$.

- **Autokowariancja** (z próby)

$$c_k = \frac{1}{n} \sum_{t=k+1}^{n} (x_t - \overline{x})(x_{t-k} - \overline{x})$$

- **Autokorelacja** (z próby)

$$r_k = \frac{c_k}{s^2} = \frac{\sum_{t=k+1}^{n} (x_t - \overline{x})(x_{t-k} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2},$$

czyli $-1 \leqslant r_k \leqslant 1$.

# Proste statystyki

Niech $k = 2m+1$.

- **Prosta średnia krocząca**

$$SMA_k(t) = \frac{1}{k} \sum_{i=-m}^{m} x_{t+j},$$

gdzie $m < t \leqslant n - m$.

- Dla krańcowych wartości $t$ stosuje się zmodyfikowane warianty

$$\frac{1}{m+1} \sum_{i=0}^{m} x_{t+j} \quad \text{oraz} \quad \frac{1}{m+1} \sum_{i=-m}^{0} x_{t+j}.$$

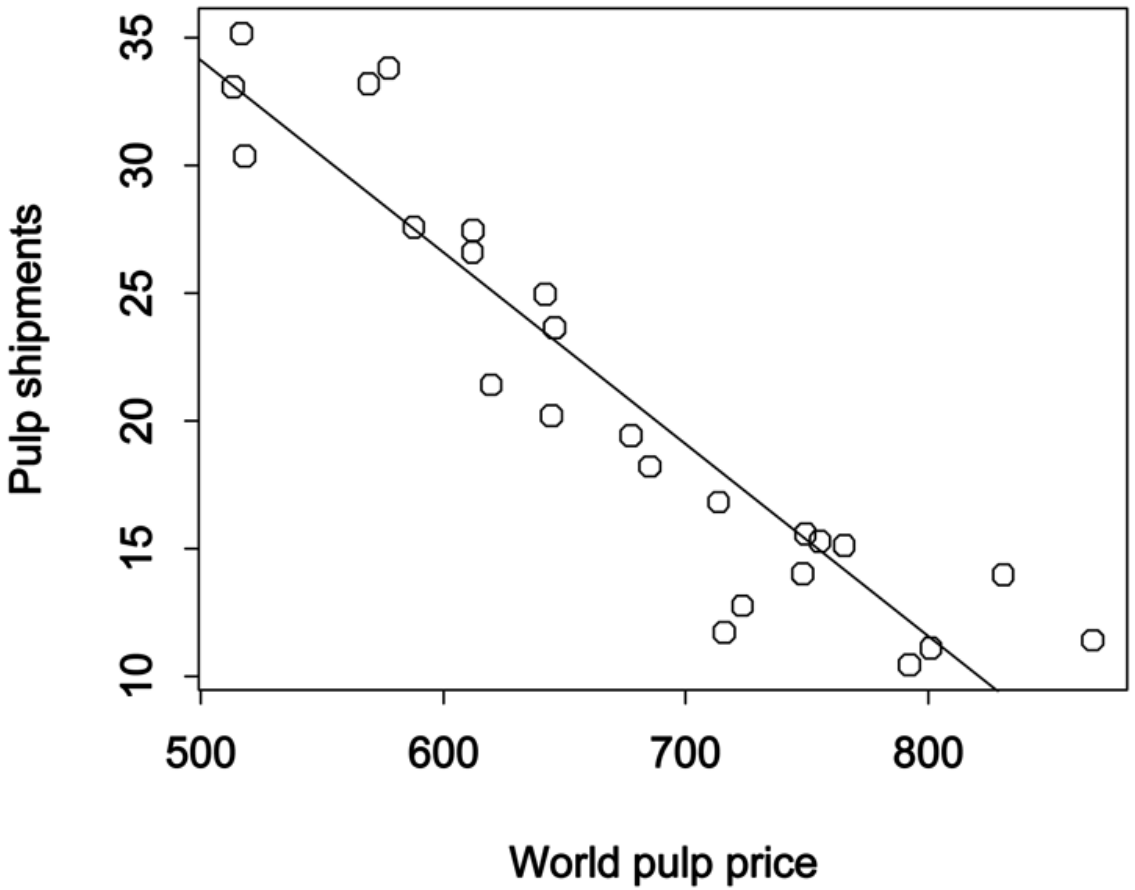
# Wykres punktowy (ang. *scatterplot*)

## Wykres

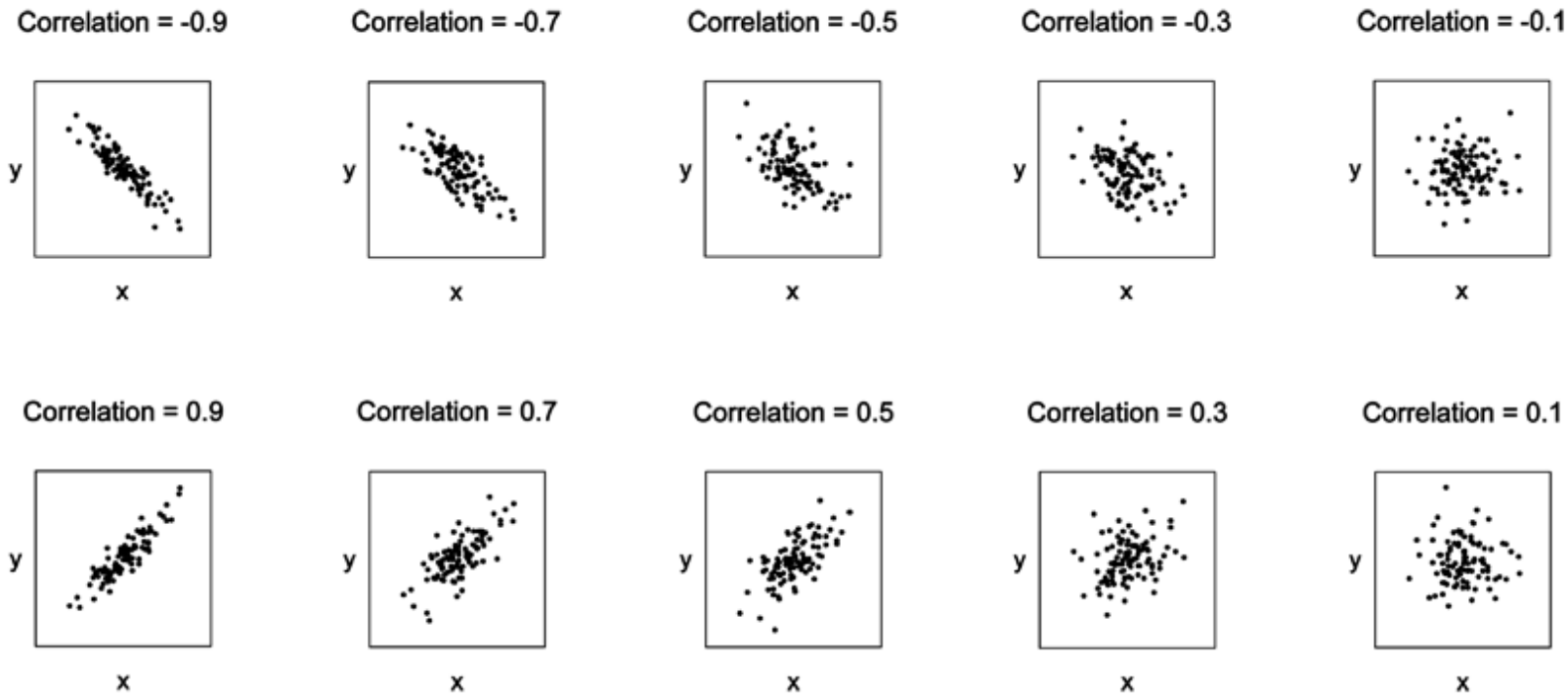
cena vs. ilość dostarczanej pulpy drzewnej

## Zjawiska

- zależność zbliżona do liniowej.

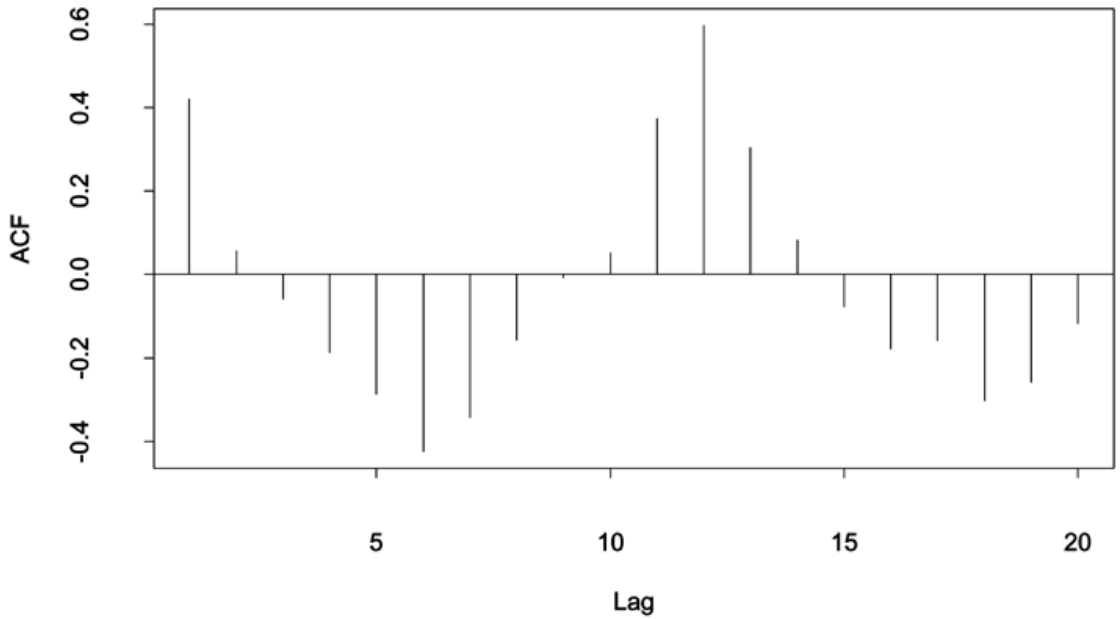


# Wykres punktowy – różne wartości (auto)korelacji



# Korelogram

miesięczne zestawienie sprzedaży piwa




- silna korelacja dodatnia dla $k = 12$,
- silna korelacja ujemna dla $k = 6$.

# Model addytywny i multiplikatywny

Rozważamy zwykle dwa zasadnicze modele:

- **addytywny**

$$X_t = T_t + S_t + E_t,$$

- **multiplikatywny**

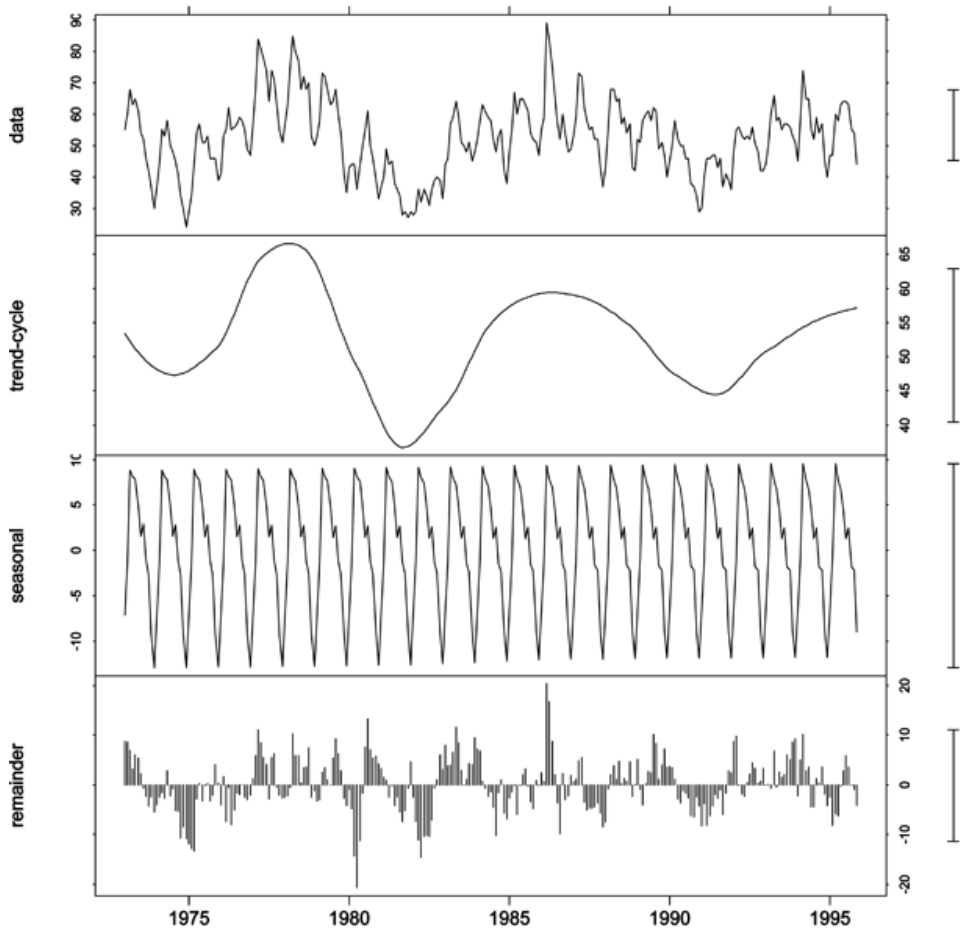
$$X_t = T_t \times S_t \times E_t,$$

gdzie

- $T_t$ – składowa *trend-cykl*,
- $S_t$ – składowa *sezonowa*,
- $E_t$ – reszta.

# Przykład – model addytywny

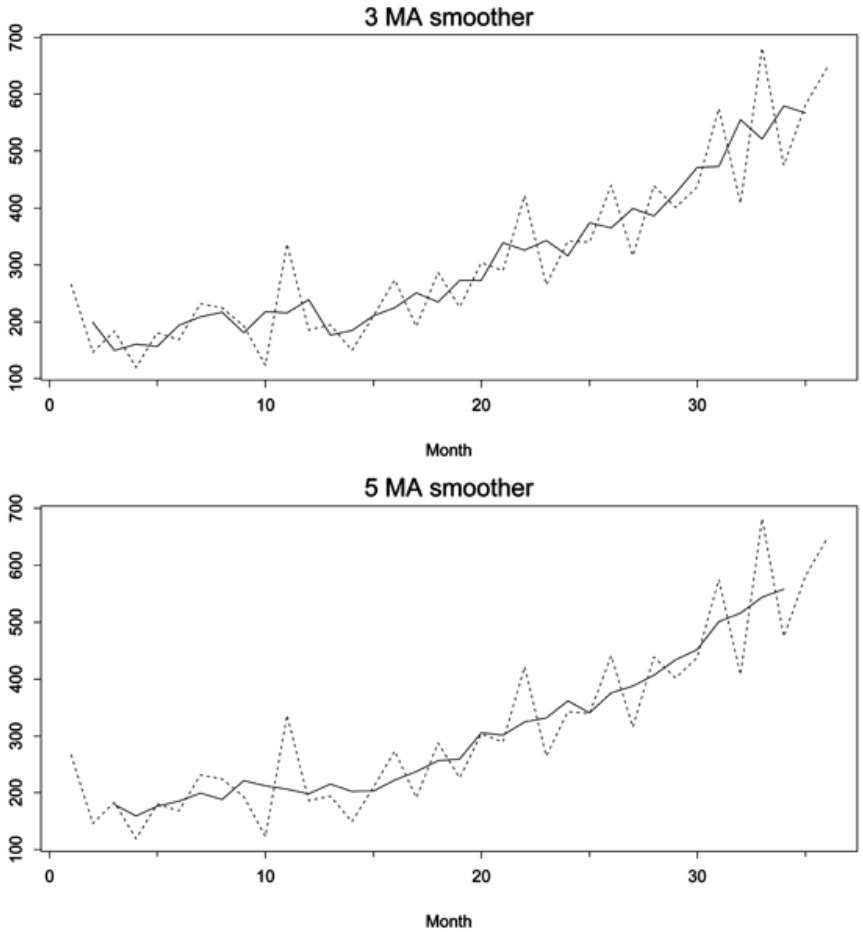
- obserwacje
- składowa *trend-cykl*
- składowa *sezonowa*
- reszta



# Wykrywanie trendu

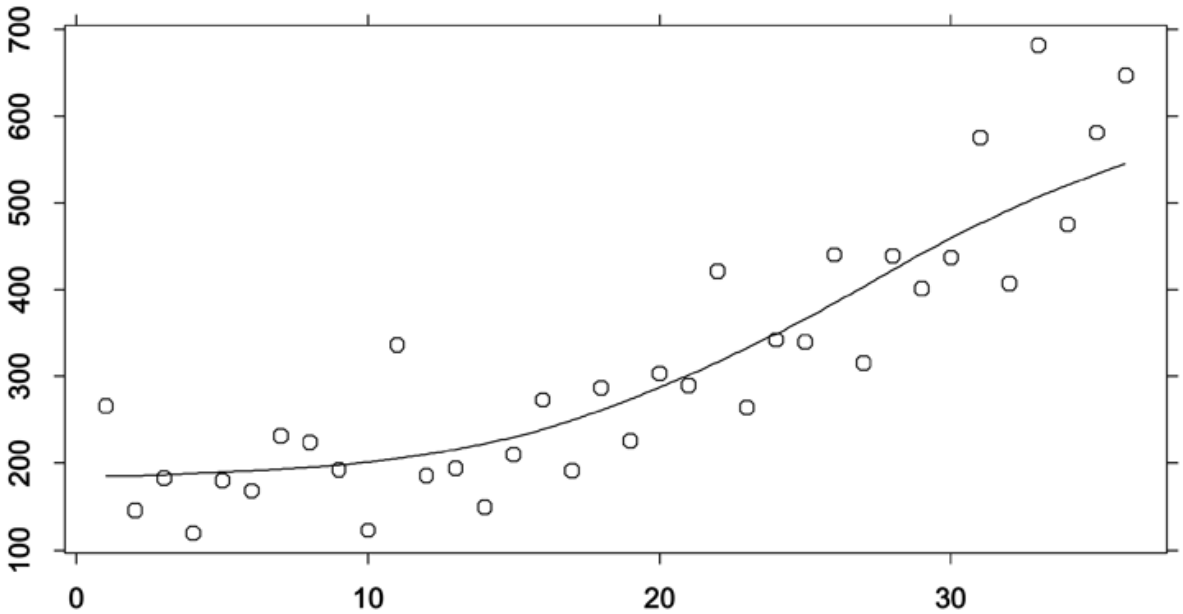
Najprostszym mechanizmem wykrywania trendu jest wygładzanie za pomocą średniej kroczącej.

- im większa wartość $k$, tym gładszy wykres,
- im mniejsza wartość $k$, tym mniej danych potrzebnych do wyliczenia średniej kroczącej.



# Wykrywanie trendu

Optymalnie wyznaczony trend dla $k = 19$ ze zmodyfikowaną średnią kroczącą dla krańcowych wartości $t$.



# Usuwanie sezonowości

Po usunięciu składowej *trend-cykl*, otrzymujemy

$$X_t - T_t = S_t + E_t, \qquad X_t / T_t = S_t \times E_t.$$

- Załóżmy, że badany szereg czasowy jest ciągiem comiesięcznych obserwacji danego zjawiska.
- Średnie arytmetyczne obserwacji (po usunięciu trendu) z poszczególnych miesięcy wyznaczają przybliżenie składowej *sezonowej*.

# Usuwanie trendu i sezonowości – przykład

Niech $n = 5 \times 12$ mies. $= 60$

- Dane są obserwacje $(x_1, x_2, \ldots, x_{60})$.
- Przyjmujemy model addytywny.

## Algorytm

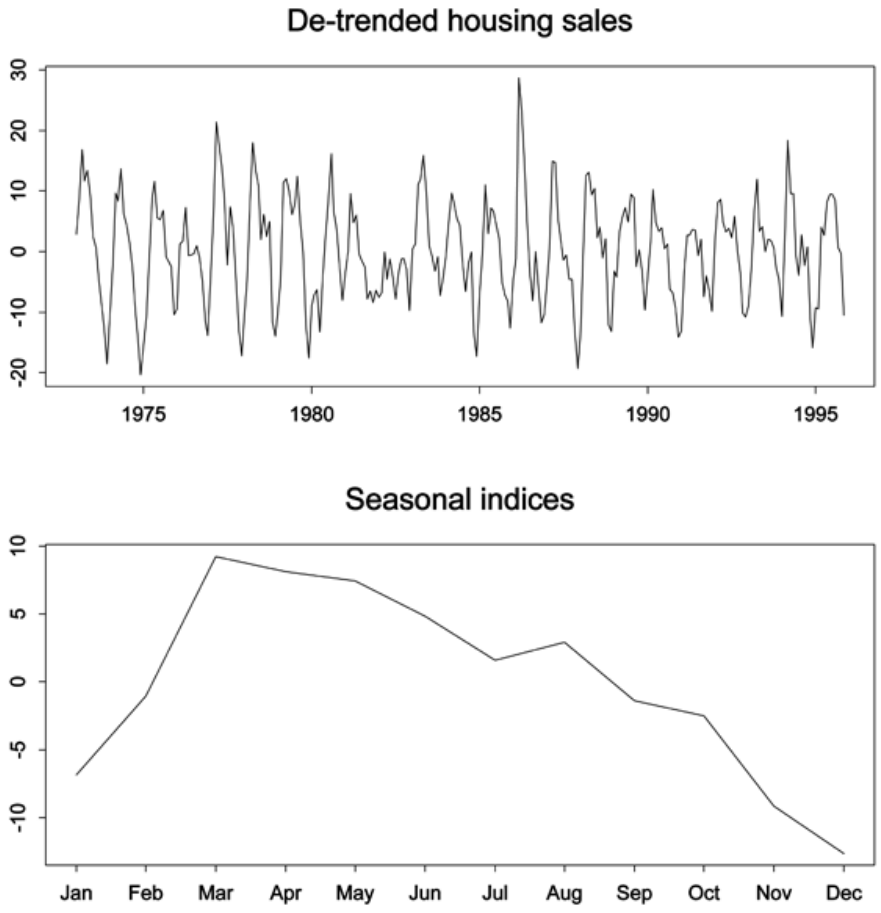
(1) Obliczamy $\tau_t = \mathrm{SMA}_k(t)$, np. dla $k = 15$.

(2) Wyznaczamy średnie dla miesięcy

$$\sigma_{Jan} = \frac{\sum_{i=0}^{4}(x_{1+12\cdot i} - \tau_{1+12\cdot i})}{5}, \ldots, \sigma_{Dec} = \frac{\sum_{i=0}^{4}(x_{12+12\cdot i} - \tau_{12+12\cdot i})}{5}$$

(3) Obliczamy resztę (błąd) $e_t = x_t - \tau_t - \sigma_{mies.}$

# Usuwanie trendu i sezonowości – przykład



# Dekompozycja – uwagi

- Pomyślnie przeprowadzona dekompozycja dostarcza istotnych informacji o postaci szeregu czasowego.
- Wiele metod predykcji wymaga usunięcia trendu i sezonowości.

## Problemy

- Czy składowe *trend-cykl* i *sezonową* również należy przewidywać czy przyjąć, że nie ulegną zmianie w przyszłości?
- Jak dużo danych (np. $\tau_t, \sigma_t$) należy dodatkowo gromadzić dla skutecznej predykcji?

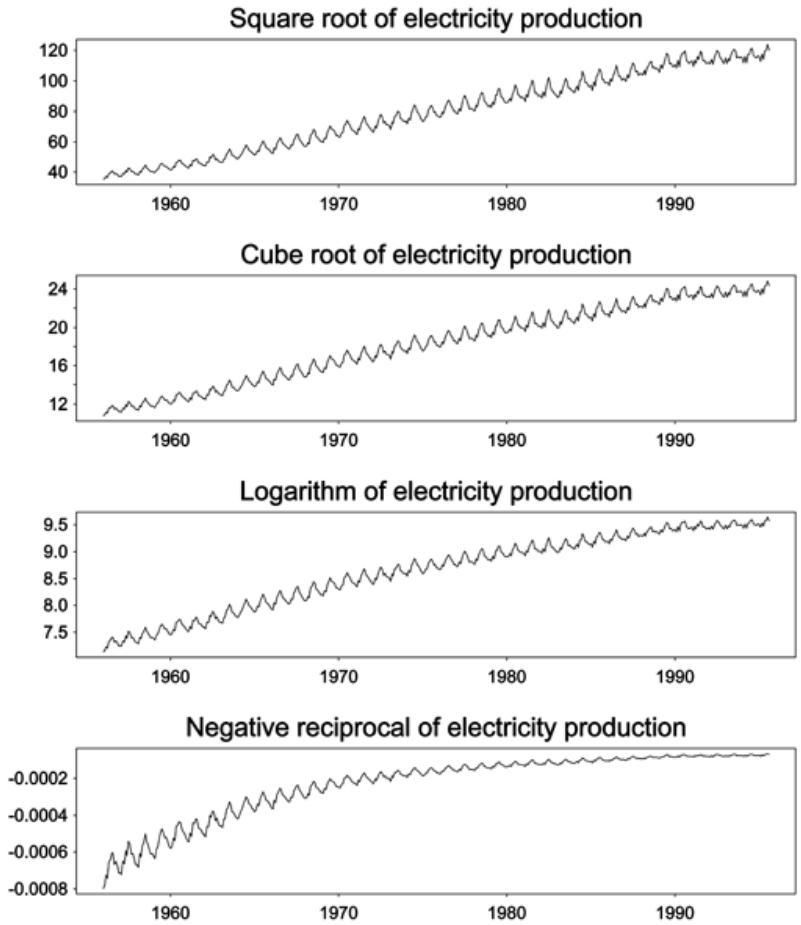
# Zaszumienie danych

Wszelkie obserwacje narażone są na liczne zaszumienia, np.:

- przerwy w działaniu urządzenia pomiarowego (braki danych),
- znaczne błędy pomiaru (nadspodziewanie duże lub małe wartości, ang. *outliers*),
- wpływ kalendarza, w szczególności:
  - liczba dni w miesiącu,
  - liczba weekendów,
  - święta (w tym święta „ruchome”).

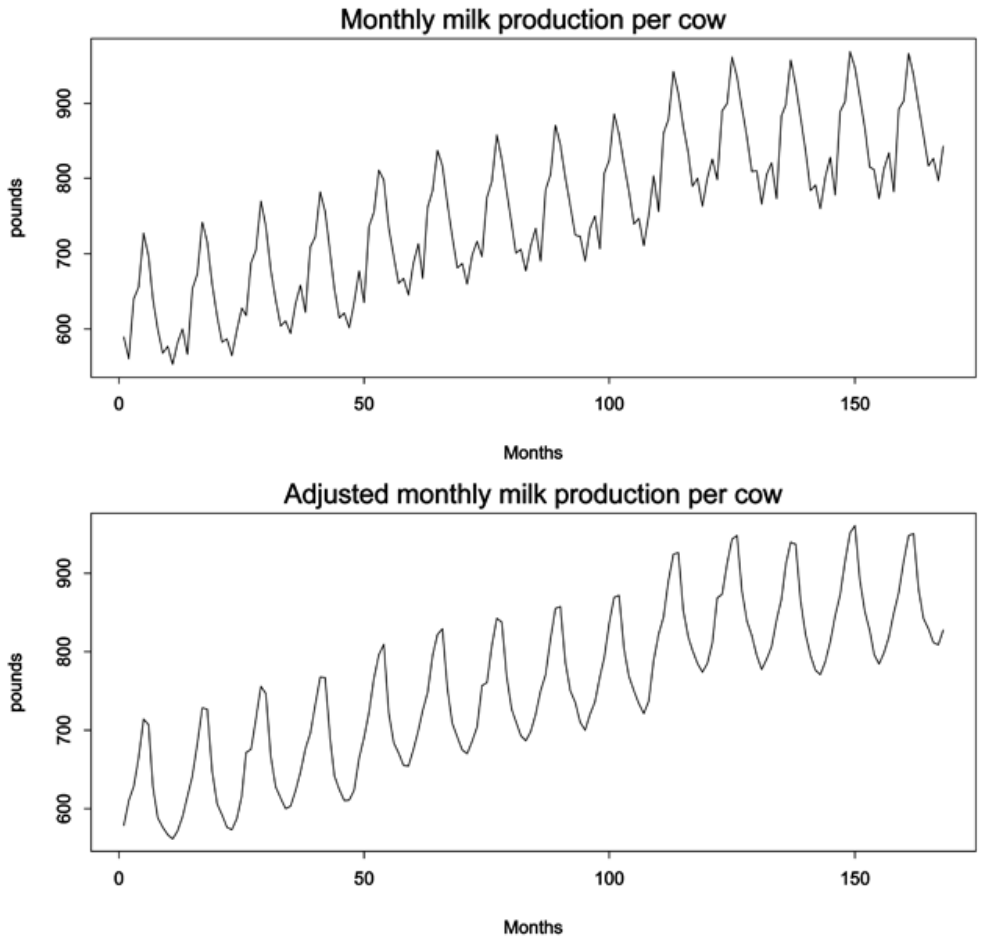
# Proste przekształcenia matematyczne

Szczególną rolę odgrywa logarytmowanie, które pozwala przejść od modelu multiplikatywnego do addytywnego.



# Proste przekształcenia kalendarzowe

W wykresie na dole uwzględniono liczbę dni w miesiącu.



# Predykcja

Istnieje wiele metod prognozowania szeregów czasowych, m.in.:

- analiza jakościowa (np. kondycja firm, stan gospodarki),
- wiedza i doświadczenie (np. odwołanie się do podobnych zjawisk opisanych w literaturze),
- intuicja (np. w kontekście postępowania drugiej osoby),
- użycie jako prognozy ostatniej odnotowanej wartości,
- **analiza ilościowa** (metody matematyczne).

# Analiza ilościowa

Wykorzystanie metod matematycznych wymaga spełnienia pewnych warunków.

- Dane obciążone trendem i/lub sezonowością utrudniają predykcję.
- W modelu $X_t = T_t + S_t + E_t$ lub $X_t = T_t \times S_t \times E_t$ racjonalna jest analiza ilościowa wyłącznie dla $E_t$.
- Podstawowym wymogiem nakładanym na prognozowany szereg czasowy jest jego **stacjonarność**, a więc stacjonarność procesu, którego jest on instancją.

# Proces stacjonarny

Proces nazywamy **stacjonarnym**, jeżeli dla dowolnych $t, k \in T$:

- wartość oczekiwana jest stała

$$\mathbb{E}(X_t) = \mathbb{E}(X_{t+k}) = \mu$$

- autokowariancja zależy jedynie od $k$

$$\mathrm{Cov}(X_t, X_{t+k}) = \mathrm{Cov}(X_1, X_{k+1}) = \gamma(k)$$

$$X_1, \ldots, X_{t-1}, \underbrace{X_t, X_{t+1}, \ldots, X_{t+k-1}, X_{t+k}}_{\mu, \gamma(k)}, X_{t+k+1}, X_{t+k+2}, \ldots$$

# Biały szum

Proces stacjonarny $(X_t)_{t\in T}$ złożony wyłącznie ze zmiennych losowych parami nieskorelowanych nazywamy **białym szumem**.

Wówczas dla dowolnego $t \in T$

- $\mathbb{E}(X_t) = \mu$,
- $\mathrm{Cov}(X_t, X_{t+k}) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$.

Biały szum jest więc *„nieprzewidywalny”*.

# Model autoregresywny AR($p$)

Model procesu, w którym każda zmienna losowa jest w liniowej zależności z $p > 0$ poprzedzającymi ją zmiennymi, nazywamy **modelem autoregresywnym** rzędu $p$, AR($p$):

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t,$$

gdzie $c$ = const, $-1 \leqslant \phi_1, \phi_2, \ldots, \phi_p \leqslant 1$, zaś $e_t$ – biały szum.

## Random walk *(„spacer losowy")*

Szczególny przypadek modelu AR(1) dla $c = 0, \phi_1 = 1$

$$X_t = X_{t-1} + e_t.$$

# Częściowa autokorelacja

Rozważmy model AR(1), czyli $X_t = c + \phi_1 X_{t-1} + e_t$.

- $X_t$ jest związane z $X_{t-1}$, więc występuje silna autokorelacja $r_1$.
- $X_{t-1}$ jest z kolei związane z $X_{t-2}$, więc pośrednio $X_t$ jest związane z $X_{t-2}$, czyli występuje istotna autokorelacja $r_2$.
- ...

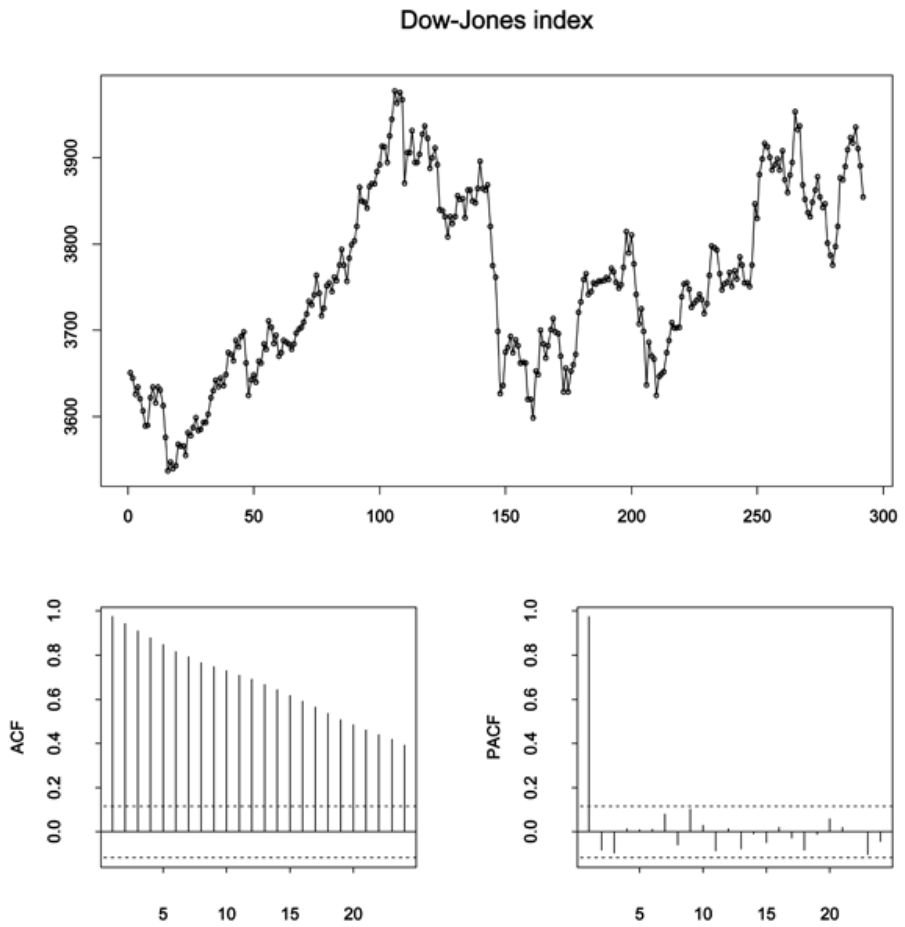
Problem:
Jaka jest bezpośrednia zależność pomiędzy $X_t$ a $X_{t-k}$ dla $k > 1$?

Odpowiedzią jest **częściowa autokorelacja**, czyli współczynnik $\phi_k$ w modelu AR($k$)

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_k X_{t-k} + e_t,$$

# Częściowa autokorelacja – przykład

- Notowania indeksu Dow-Jones łudząco przypominają *random walk*.
- Model ten jest często najbardziej adekwatny w zagadnieniach finansowych.



# Model średniej kroczącej MA($q$)

W praktyce, składowa $e_t$ w stanowi **residuum**, tzn. różnicę pomiędzy teoretycznym modelem a odnotowaną obserwacją (błąd).

**Model średniej kroczącej** MA($q$) jest kombinacją liniową residuów

$$X_t = c + e_t - \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q},$$

gdzie $c$ = const, $-1 \leqslant \theta_1, \theta_2, \ldots, \theta_q \leqslant 1$, zaś $e_t$ – biały szum.

## Uwaga

Modelu MA($q$) nie należy mylić ze średnią kroczącą obserwacji (np. $\mathrm{SMA}_k$) wykorzystywaną do wykrywania trendu. W modelu MA($q$) mamy do czynienia ze średnią kroczącą ciągu residuów.

# Model ARIMA$(p, d, q)$

AR$(p)$ i MA$(q)$ wzajemnie się dopełniają, tworząc model ARMA$(p, q)$

$$X_t = c + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

- W bardzo wielu praktycznych problemach, aby usunąć niestacjonarność z szeregu czasowego, wystarczy rozważyć szereg różnic $X_t - X_{t-1}$.
- Model ARIMA$(p, d, q)$ to model ARMA$(p, q)$ dla szeregu czasowego poddanego $d$-krotnemu różnicowaniu (w celu zagwarantowania stacjonarności).

# Kryterium Akaike

$$AIC = 2k - 2\sum_{i=0}^{N} ln(p_i),$$

gdzie $k$ to liczba parametrów modelu, $N$ to liczba rozpatrywanych obserwacji, zaś $p_i$ to estymowane prawdopodobieństwo, że przy założeniach rozpatrywanego modelu $i$-ta obserwacja będzie miała taką wartość jaka została zarejestrowana.

# Błędy dopasowania modelu

- mean absolute error:

$$MAE = mean(|e_i|)$$

- root mean squared error:

$$RMSE = \sqrt{mean(e_i^2)}$$

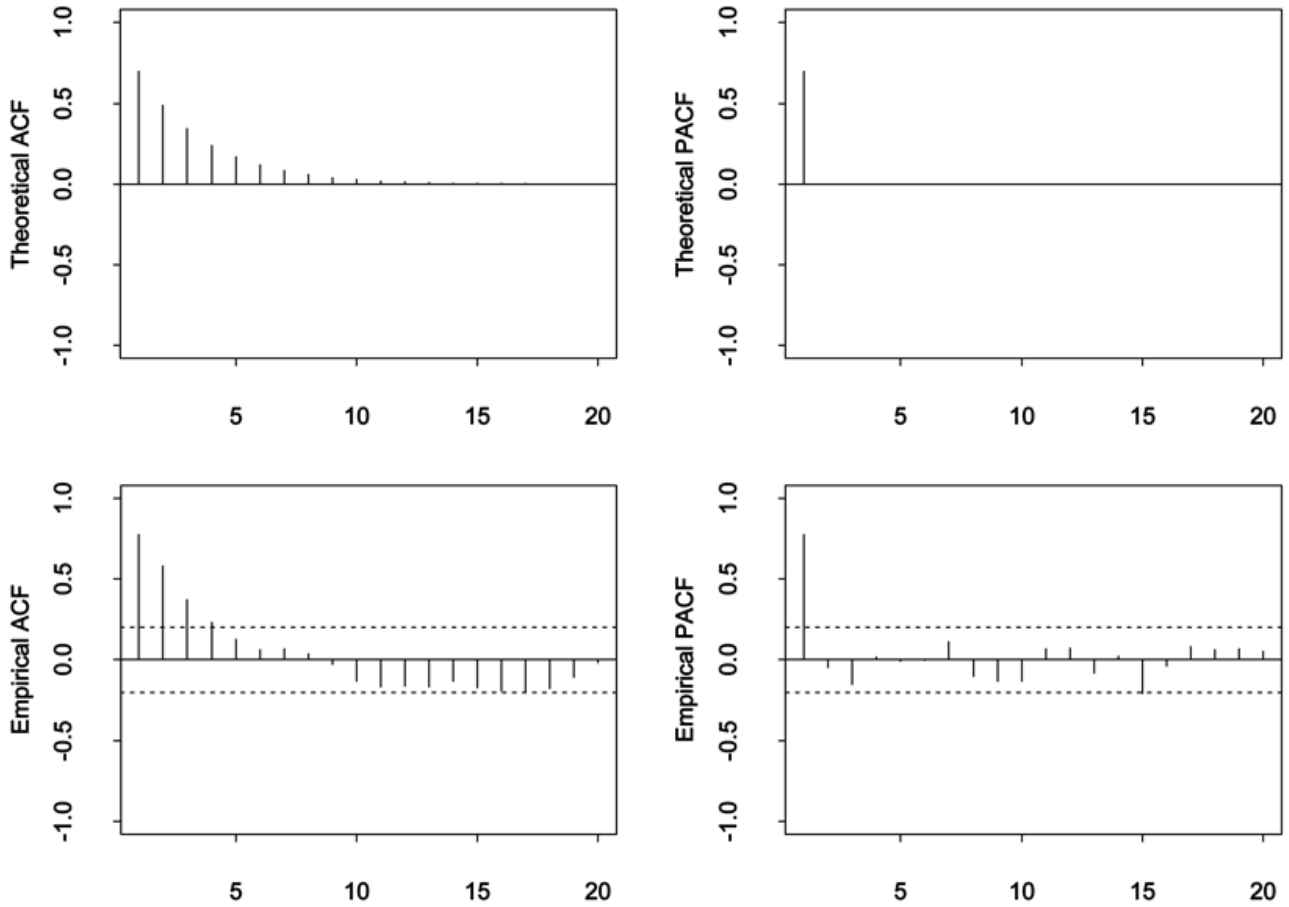
- mean absolute percentage error:

$$MAPE = mean(|p_i|), \quad \text{where } p_i = \frac{100 e_i}{y_i}$$

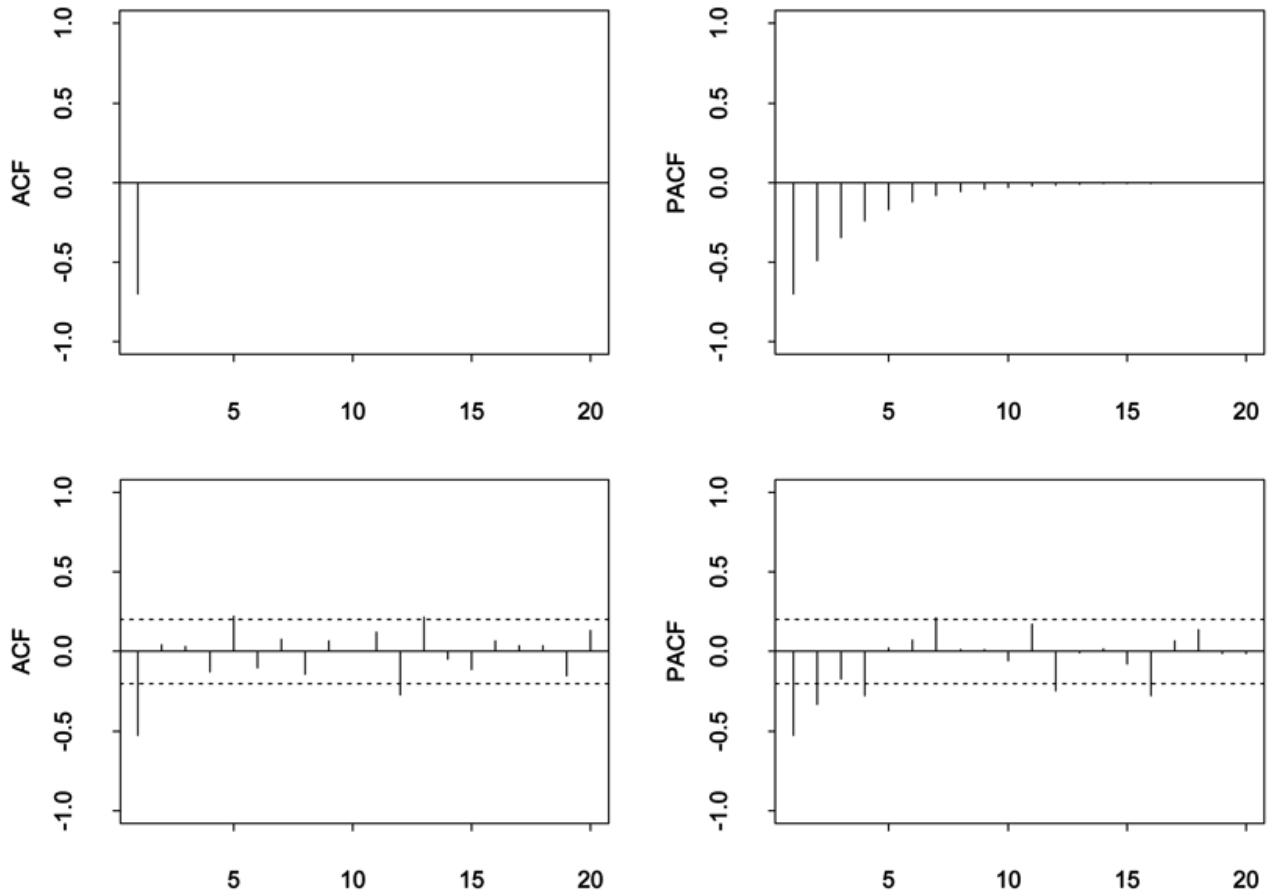
- mean absolute scaled error:

$$MASE = mean(|q_j|), \quad \text{where } q_j = \frac{e_j}{(1/(T-1)\sum_{t=2}^{T}|y_t - y_{t-1}|)}$$

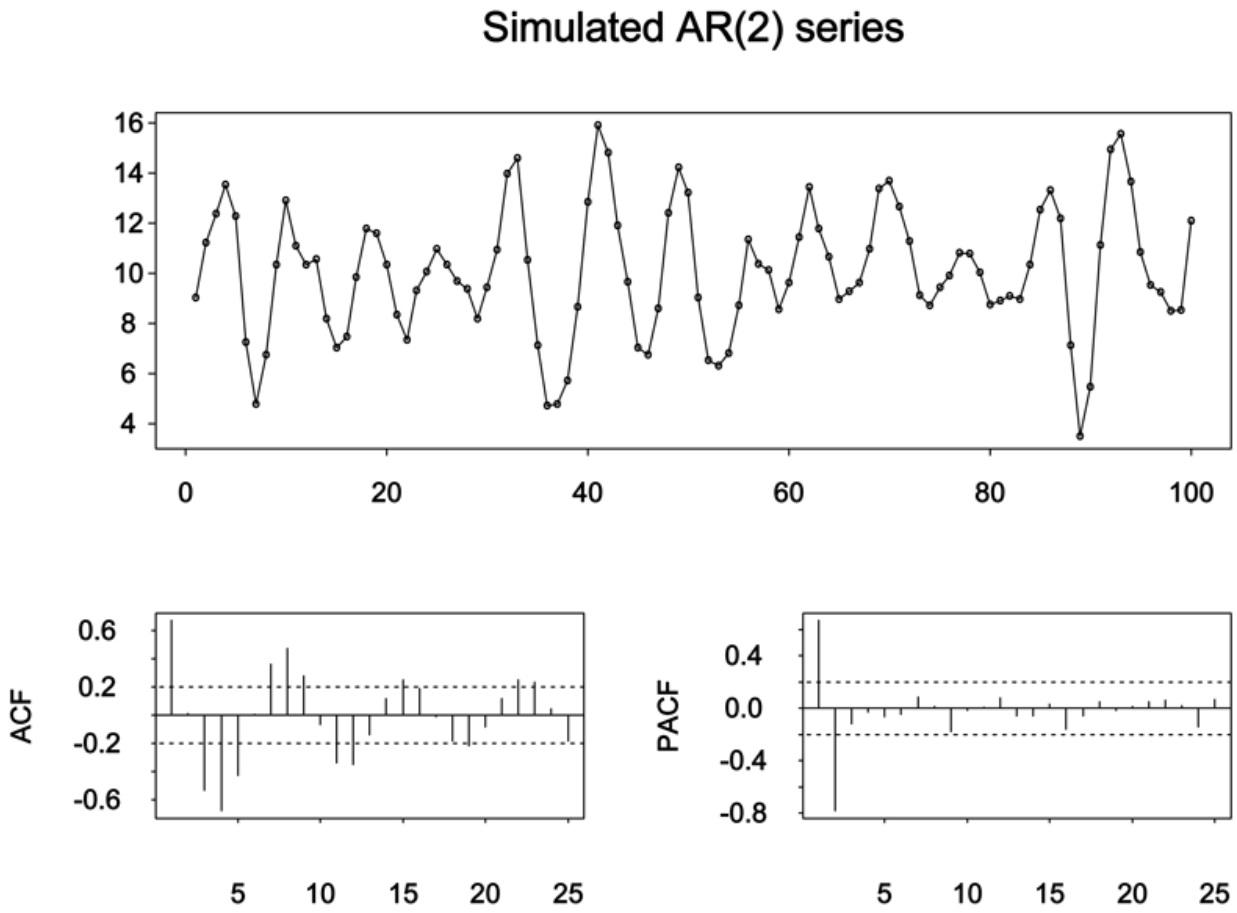
# Przykład – AR(1)



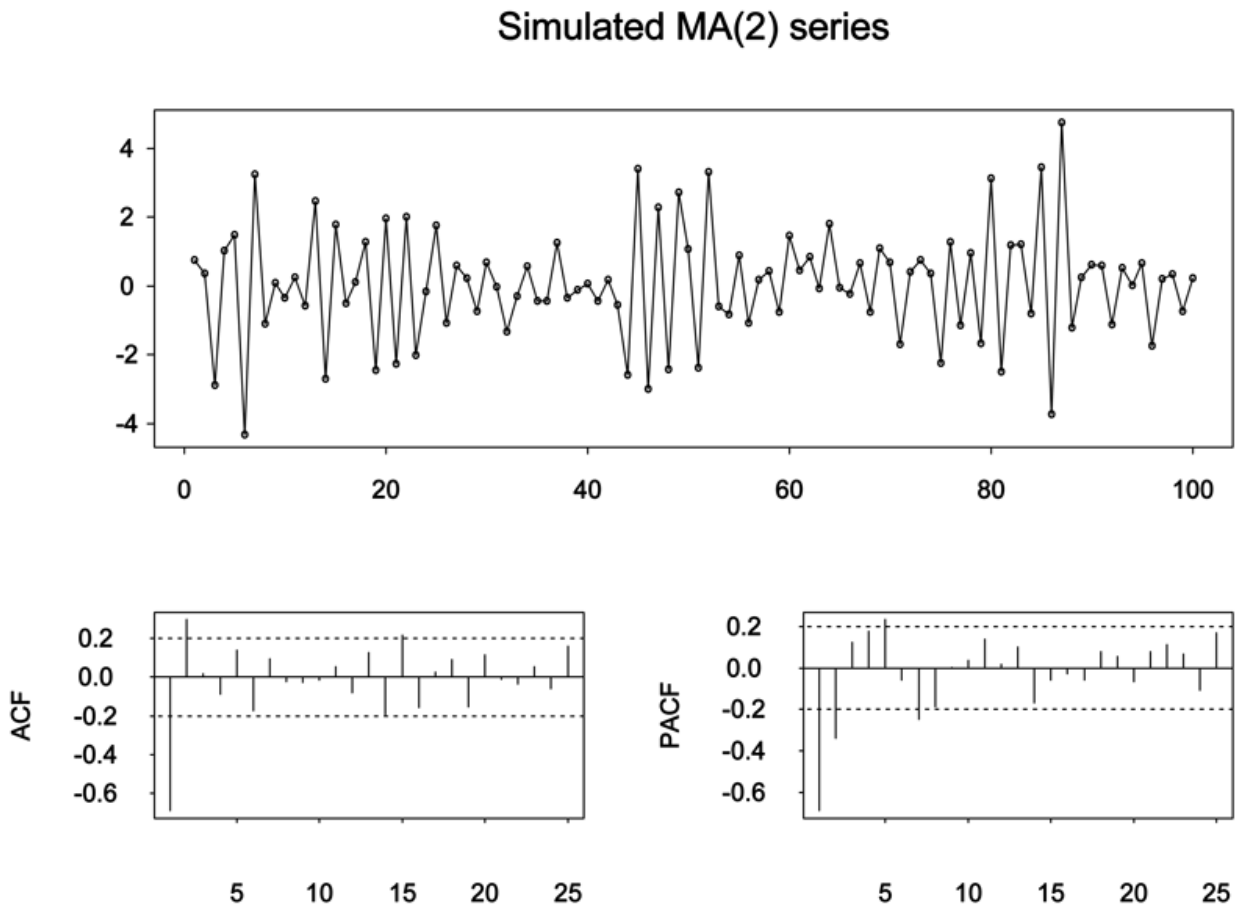
# Przykład – MA(1)



# Przykład – AR(2)



# Przykład – MA(2)



# Model mieszany

W przypadku mieszanym dobór parametrów ARIMA$(p, d, q)$ nie jest natychmiastowy.

- zwykle dobór metodą empiryczną,
- podobne rezulataty mogą być osiągane dla różnych parametrów,
- na ogół wartości $p, d, q$ ograniczają się do zbioru $\{0, 1, 2\}$,

Powszechną praktyką jest podział danych historycznych na dane treningowe (*„uczące”*) i testowe.

- dobór parametrów na danych treningowych,
- weryfikacja skuteczności predykcji na danych testowych,
- powszechnie stosowany błąd średniokwadratowy jako miara dopasowania.

# Podsumowanie

- Szereg czasowy jest realizacją pewnego procesu stochastycznego.
- Metody matematyczne dostarczają wiele modeli pozwalających opisywać szeregi czasowe występujące w praktyce.
- Stacjonarność szeregu czasowego jest podstawowym wymogiem do zbudowania skutecznego modelu predykcji.
- ARIMA$(p, d, q)$ jest uniwersalnym modelem autoregresywnym i średniej kroczącej z opcją różnicowania (w celu usuwania niestacjonarności).

# Literatura


- Box, G., E., P., Jenkins, G., M., Reinsel, G., C., *Time series analysis: Forecasting and control*, 1994.
- Makridakis, S., G., Wheelwright, S., C., Hyndman, R., F., *Forecasting: Methods and applications*, 1997.
- Falk, M. et al., *A first course on time series analysis – examples with SAS*, 2011.
- Taylor, S., J., *Modelling financial time series*, 2008.