Szeregi czasowe

Martyna Śpiewak
Bootcamp Data Science

Szeregiem czasowym (ang. *time series*) nazywamy realizacje pewnej wielkości zarejestrowane w kolejnych odstępach czasu, np. w kolejnych dniach, miesiącach lub latach.

(Y_t)- ciąg zmiennych losowych indeksowany parametrem t (czas).

Charakterystyki szeregu

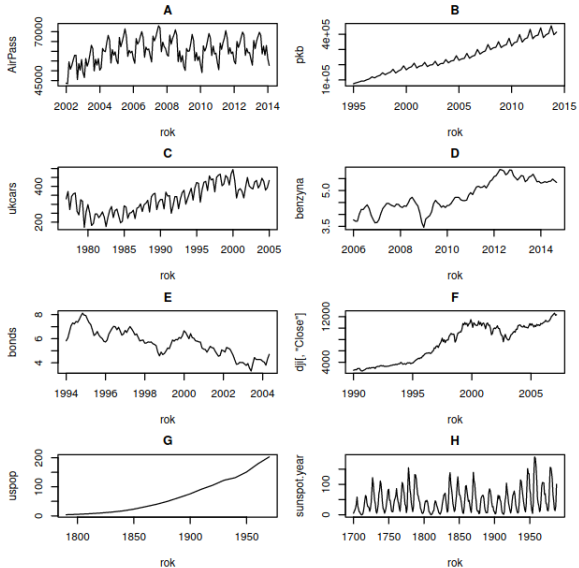
- Jak zachowuje się wartość średnia w funkcji czasu?
- Jak zależność między dwiema zmiennymi Y_s i Y_t zależy od odcinka czasu, który upłynął między tymi zdarzeniami?

Analiza szeregów czasowy ma na celu

- okrycie natury danego zjawiska losowego
- prognozowanie

Główne zadania analizy szeregów czasowych

- prognozowanie wielkości sprzedaży lub popytu na określony produkt/surowiec w kolejnych okresach,
- prognozy wartości wskaźników makroekonomicznych (np. inflacji lub PKB) w kolejnych kwartałach,
- analiza sytuacji na rynku pracy (w szczególności analiza tendencji dotyczących bezrobocia i zatrudnienia, w różnych grupach wiekowych),
- prognozowanie wartości akcji danej spółki, cen surowców, kursów walutowych itp. w kolejnych okresach,
- przewidywanie zmian cen danego produktu (np. paliw) w kolejnych miesiącach,
- analiza zmian demograficznych, socjologicznych, klimatycznych i ich wpływu na koniunkturę w określonej gałęzi przemysłu.



Źródło: Analiza i prognozowanie szeregów czasowych, Zagdański Adam, Suchwałko Artur, rok wydania 2015, wydawnictwo: Wydawnictwo Naukowe PWN.

Szereg A: miesięczna liczba pasażerów linii lotniczych (w tysiącach) w USA,w latach 2002–2014,

szereg B: kwartalne wartości produktu krajowego brutto (PKB) w Polsce,zarejestrowane w okresie 1995–2014,

szereg C: kwartalna wielkość produkcji samochodów osobowych w UK w okresie 1977:1–2005:1,

szereg D: miesięczne, średnie ceny 1 litra benzyny w Polsce w okresie 2006–2014,

szereg E: rentowność 10-letnich obligacji skarbowych USA, dane miesięczne w okresie styczeń 1994–maj 2004,

szereg F: miesięczne kursy zamknięcia indeksu Dow Jones w okresie styczeń1990–marzec 2007,

szereg G: populacja USA (w milionach) w okresie 1790–1970, dane 10-letnie,

szereg H: roczne liczby plam słonecznych w latach 1700–1988.

Operator opóźnień (ang. *lag operator*)

$$\begin{aligned} \text{lag}_1(y_t) &= y_{t-1}, \\ \text{lag}_2(y_t) &= y_{t-2}, \\ &\vdots \\ \text{lag}_h(y_t) &= y_{t-h}. \end{aligned}$$

Do oceny zależności miedzy obserwacjami i ich opóźnieniami możemy wykorzystać wykres rozproszenia punktów

$$(y_t, \text{lag}_h(y_t))$$

.

Autokorelacja

Funkcja autokorelacji (**ACF**) mierzy zależności elementu szeregu z jej opóźnieniem h-tego rzędu, tzn. autokorelacja występuje wtedy, gdy skutki działania zmienności losowej nie wygasają w danym okresie t, lecz są przenoszone na okresy przyszłe $t+1$ (autokorelacja rzędu pierwszego), $t+2$ (autokorelacja rzędu drugiego) itd.

Autokowariancja

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(y_i - \overline{y})(y_{i+h} - \overline{y})$$

Autokorelacja

$$\text{ACF}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Własności funkcji autokorelacji — ACF

- dodatnie, powoli zanikające wartości funkcji ACF sugerują, że dane zawierają deterministyczną składową trendu,
- zanikająca bardzo powoli i cykliczne wskazuje na obecność trendu sezonowego,

Funkcja autokorelacji cząstkowej

Funkcja cząstkowej autokorelacji (ang. *Partial Autocorrelation Function*) rzędu h mierzy bezpośrednią zależność pomiędzy obserwacjami odległymi o h jednostek czasowych (to znaczy Y_t i Y_{t+h}) eliminując korelację pomiędzy Y_t i Y_{t+h}, która pochodzi z obserwacji pośrednich: $Y_{t+1}, \ldots, Y_{t+h-1}$.

Liczbowo jest równa oszacowaniu współczynnika ρ_h (PACF(h) $\approx \rho_h$) w modelu:

$$y_{t+h} = \mu + \rho_1 y_{t+h-1} + \ldots + \rho_h y_t + \varepsilon_{t+h}$$

Stacjonarność

Stacjonarność – własność procesu stochastycznego (Y_t), polegająca na tym, że rozkład danego procesu stochastycznego jest stały w czasie.

Stacjonarność w szerszym sensie

- stała w czasie wartość oczekiwana

$$\mathbb{E}(y_t) = \mu,$$

- stała w czasie wariancja

$$\text{Var}(y_t) = \sigma^2 < \infty,$$

- kowariancja zależna od przesunięcia h (nie od czasu)

$$\text{Cov}(y_t, y_{t+h}) = \mathbb{E}(y_t - \mu)(y_{t+h} - \mu) = \lambda_h.$$

Biały szum

Biały szum (ang. *white noise*) jest przykładem procesu **stacjonarnego**:

$$y_t = \varepsilon_t,$$

gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma)$.

Ponieważ:

- $\mathbb{E}(y_t) = \mathbb{E}(\varepsilon_t) = 0$
- $\mathrm{Var}(y_t) = \mathrm{Var}(\varepsilon_t) = \sigma^2$
- $\mathrm{Cov}(y_t, y_{t+h}) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = \mathbb{E}(\varepsilon_t \varepsilon_{t+h}) - \mathbb{E}(\varepsilon_t)\mathbb{E}(\varepsilon_{t+h}) = 0$

Błądzenie losowe

Błądzenie losowe (ang. *random walk*) jest przykładem procesu **niestacjonarnego**:

$$y_0 = a = \text{const}$$

$$y_t = y_{t-1} + \varepsilon_t = y_0 + \sum_{i=1}^{t} \varepsilon_i,$$

gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma)$.

Ponieważ:

- $\mathbb{E}(y_t) = y_0 + \sum_{i=1}^{t} \mathbb{E}(\varepsilon_i) = y_0 = \text{const}$,
- $\text{Var}(y_t) = \text{Var}(\sum_{i=1}^{t}(\varepsilon_i) = \sum_{i=1}^{n} \text{Var}(\varepsilon_i) = t\sigma^2$.

Testowanie stacjonarności – test Dickey-Fullera

Idea **testu Dickey-Fullera** opiera się na modelu autokorelacji pierwszego rzędu następującej postaci:

$$y_t = \varphi y_{t-1} + \varepsilon_t,$$

gdzie φ - parametr modelu autoregresji i ε_t - składnik losowy o własnościach białego szumu.

Hipoteza

$$H_0 : \text{szereg jest niestacjonarny} (\varphi = 1)$$

Operator różnicowania (ang. *difference operator*)

$$\Delta(y_t) = y_t - y_{t-1},$$
$$\Delta^2(y_t) = \Delta(\Delta(y_t)) = \Delta(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$
$$\Delta^k(y_t) = \underbrace{\Delta\Delta\ldots\Delta}_{k}(y_t).$$

Operator różnicowania z opóźnieniem sezonowym

$$\Delta(y_t)_s = y_t - y_{t-s}$$

Własności różnicowania

- różnicowanie z opóźnieniem 1 usuwa z szeregu trend liniowy
- aby usunąć trend wielomiany stopnia k, należy k-krotnie zróżnicować szereg z opóźnieniem 1,
- różnicowanie z opóźnieniem sezonowym s usuwa trend sezonowy o okresie s oraz jednocześnie usuwa trend liniowy,

Sprowadzanie do stacjonarności - różnicowanie

Rozważmy szereg postaci:

$$y_t = at + \varepsilon_t,$$

gdzie $\varepsilon_t \sim \mathcal{N}(0, \sigma)$.

Szereg jest **niestacjonarny**, ponieważ wartość oczekiwana zależy od czasu:

$$\mathbb{E}(y_t) = at \neq \text{const.}$$

Różnicowanie

$$\Delta(y_t) = a + \varepsilon_t - \varepsilon_{t-1},$$

wtedy szereg spełnia:

- $\mathbb{E}(\Delta(y_t)) = a$,
- $\text{Var}(\Delta(y_t)) = \text{Var}(\varepsilon_t - \varepsilon_{t-1}) = \text{Var}(\varepsilon_t) + \text{Var}(\varepsilon_{t-1}) - 2 \cdot \text{Cov}(\varepsilon_t, \varepsilon_{t-1}) = 2\sigma^2$,
- $\text{Cov}(\Delta(y_t), \Delta(y_{t+h})) = \text{Cov}(a + \varepsilon_t - \varepsilon_{t-1}, a + \varepsilon_{t+h} - \varepsilon_{t+h-1}) = 0$.

Cel analizy szeregów czasowych

Zbudowanie modelu pewnego zjawiska w oparciu o obserwowane zmiany w czasie pewnych mierzalnych wielkości.

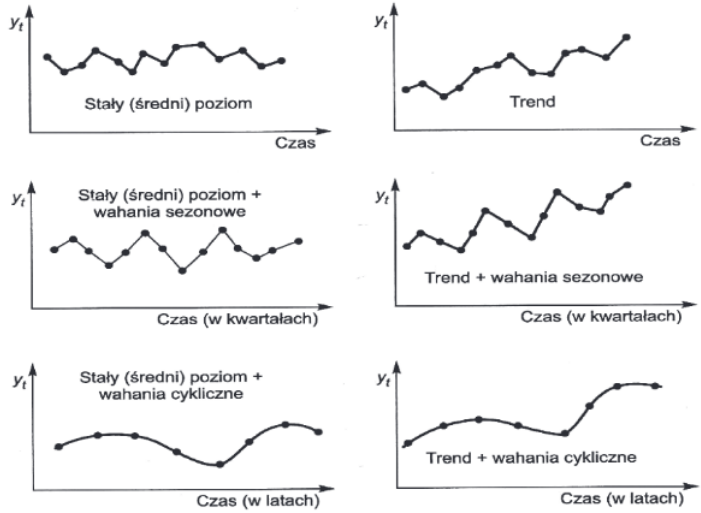
Założenie: Szereg składa się z:

- **części systematycznej** - składowa stała, trend, wahania sezonowe i cykliczne (w oparciu, o które buduje się model),
- **części przypadkowej** - szumu, wahań przypadkowych.

W analizie szeregów czasowym dążymy do **dekompozycji szeregu czasowego** – wyodrębnienia wymienionych składowych szeregu.

Składowe szeregu czasowego

- **stała** – przeciętny poziom zmiennej,
- **trend** – charakteryzuje długookresową tendencję zmian w szeregu czasowym, może on oznaczać w miarę regularnie powtarzający się wzrost lub spadek wartości lub też brak wyraźnej tendencji zmian,
- **wahania okresowe** – wahania okresowe, regularne odchylenia od tendencji rozwojowej, składnik powtarzający się cyklicznie,
  - **cykliczne** – długookresowe, rytmiczne wahania (cykl koniunkturalny gospodarki),
  - **sezonowe** – krótkookresowe do 1 roku,
- **szum** – zakłócenia, wahania przypadkowe.



Źródło: M. Cieślak, Prognozowanie gospodarcze. Metody i zastosowania, PWN, Warszawa 2005.

Dekompozycja szeregu czasowego

W trakcie budowy modelu przeprowadza się dekompozycje szeregu czasowego w zależności od przyjętych założeń.

Założenia: m_t - trend, s_t - sezonowość, ε_t - szum.

Model addytywny

$$y_t = m_t + s_t + \varepsilon_t$$

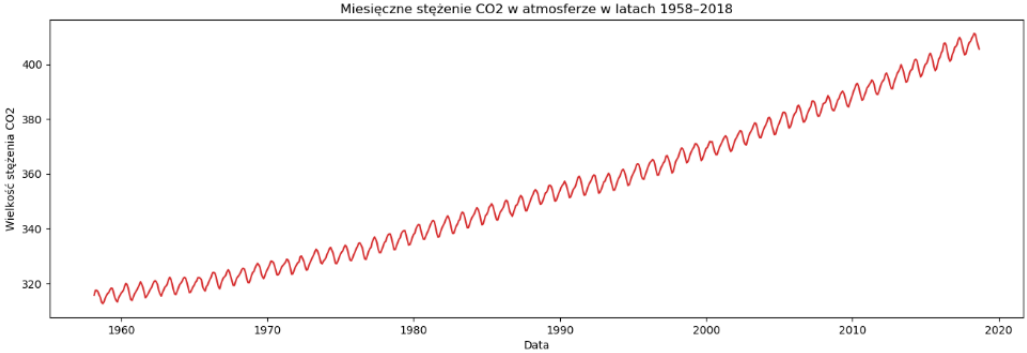
Model multiplikatywny

$$y_t = m_t \cdot s_t \cdot \varepsilon_t$$

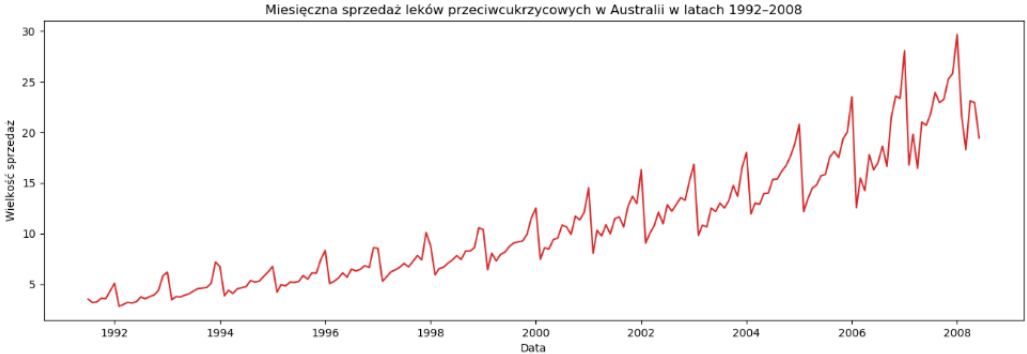
Dekompozycja szeregu czasowego

- **dekompozycja addytywna** – wielkość wahań sezonowych lub wariancja danych wokół tendencji długoterminowej (trendu) nie zmienia się wraz z poziomem szeregu,
- **dekompozycja multiplikatywna** – amplituda wahań sezonowych lub wariancja danych proporcjonalna do poziomu szeregu.

Model addytywny szeregu czasowego z trendem



Model multiplikatywny szeregu czasowego z trendem



Algorytm dekompozycji szeregu czasowego

Krok 1 Wyznaczamy oszacowany składnik trendu $\hat{m}_t$, wykorzystując np. metodę średniej ruchomej z jednakowymi wagami.

Krok 2 Oszacowany trend jest eliminowany z danych. Wyznaczamy szereg:

- dekompozycja addytywna: $y_t - \hat{m}_t$,
- dekompozycja multiplikatywna: $y_t / \hat{m}_t$.

Krok 3 Wyznaczamy indeksy sezonowe dla poszczególnych miesięcy, kwartałów, itd. Indeksy sezonowe są wyznaczane poprzez uśrednienie wartości szeregu z kroku 2 dla każdej jednostki czasu (np. miesiąc, kwartał) i wszystkich okresów (lat).

Krok 4 Standaryzujemy wskaźniki sezonowe, tak aby nie miały one wpływu na tendencję długoterminową.

Krok 5 Wyznaczamy reszty, usuwając trend i sezonowość

- dekompozycja addytywna: $\varepsilon_t = y_t - \hat{m}_t - \hat{s}_t$,
- dekompozycja multiplikatywna: $\varepsilon_t = y_t / \hat{m}_t / \hat{s}_t$.

Modelowanie szeregów czasowych z autokorelacją

Jeśli występuje autokorelacja szereg czasowy modeluje się przy użyciu:

- modelu autoregresji rzędu p (**AR**).
- modelu średniej ruchomej rzędu q (**MA**).

Model autoregresji rzędu p – AR(p)

W statystyce model autoregresyjny jest typem modelu opisującego procesy losowe, który jest często używany do modelowania i przewidywania zjawisk o charakterze **naturalnym** i **społecznym** .

Procesy autoregresyjne typu AR(p) mogą być analizowane różnymi metodami, w tym standardową liniową metodą MNK, oraz posiadają dość prostą interpretację . Powyższe właściwości wynikają z faktu, że modele AR(p) są po prostu **regresją liniową wartości bieżącej szeregu czasowego od poprzedzających ją wartości tego szeregu**.

Model autoregresji rzędu p – AR(p)

Modelem autoregresji rzędu p nazywamy *stacjonarny* szereg czasowy Y_t spełniający równanie:

$$Y_t = \sum_{i=1}^{p} \phi_i \cdot Y_{t-i} + \varepsilon_t,$$

gdzie $\phi_1, \phi_2, \ldots, \phi_p$ są współczynnikami modelu oraz ε_t to biały szum.

Wartość szeregu w chwili T jest przedstawiona jako liniowa kombinacja p wcześniejszych wartości $Y_{t-1}, \ldots, Y_{t-p}$, do której dodajemy zakłócenie w postaci białego szumu.

Model średniej ruchomej — MA(q)

Założeniem dla modelu średniej ruchomej jest uznanie, że wartość bieżąca danego szeregu zależy od wartości białego szumu lub nieprzewidywalnych szoków, co przedstawić można w postaci liniowej regresji. Dodatkowo uznaje się, że te losowe szoki pochodzą z tego samego rozkładu prawdopodobieństwa, z reguły rozkładu normalnego.

Charakterystyczną cechą tego modelu jest również fakt, że występujące szoki mogą wpływać na przyszłe wartości szeregu czasowego. Niestety, ponieważ te składniki losowe nie są obserwowalne, nie można stwierdzić, kiedy się zaczęły lub kiedy się skończą, co znacznie utrudnia dopasowanie modelu MA(q) w porównaniu do wcześniej opisanego modelu AR(p).

Model średniej ruchomej — MA(q)

Modelem średniej ruchomej rzędu q nazywamy *stacjonarny* szereg Y_t spełniający równanie:

$$Y_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \cdot \varepsilon_{t-i},$$

gdzie $\theta_1, \theta_2, \ldots, \theta_q$ są współczynnikami modelu oraz ε_t to biały szum.

MA(1): $Y_t = \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1}$.

Ważne: MA(q) jest przykładem modelu, dla którego **korelacja czasowa** dla opóźnień większych niż rząd q jest równa zeru.

Model autoregresji ruchomej średniej – ARMA(p, q)

Modelem autoregresji ruchomej średniej nazywamy *stacjonarny* szereg czasowy Y_t spełniający równanie:

$$Y_t = \sum_{i=1}^{p} \phi_i \cdot Y_{t-i} + \sum_{j=1}^{q} \theta_j \cdot \varepsilon_{t-j} + \varepsilon_t,$$

gdzie $\phi_1, \phi_2, \ldots, \phi_p$ oraz $\theta_1, \theta_2, \ldots, \theta_q$ są współczynnikami modelu oraz ε_t to biały szum.

Identyfikacja modelu

1. przekształcenie szeregu do postaci stacjonarnej:
   - transformacje potęgowe,
   - eliminacja trendu i sezonowości na podstawie dekompozycji,
   - eliminację trendu i sezonowości przy użyciu różnicowania;
2. identyfikacja białego szumu;
3. identyfikacja MA(q) na podstawie funkcji ACF — jeśli autokorelacja próbkowa ACF(h) znajduje się pomiędzy przedziałami ufności $\pm 1,96\sqrt{n}$ dla $h > q$, to spodziewamy się, że dane są realizacją procesu MA(q);
4. identyfikacja AR(p) na podstawie funkcji PACF — jeśli cząstkowa autokorelacja próbkowa PACF(h) znajduje się pomiędzy przedziałami ufności $\pm 1,96\sqrt{n}$ dla $h > p$, to spodziewamy się, że dane są realizacją procesu AR(p);

Podsumowanie

Modele z klasy ARMA mogą być zastosowane w przypadku, gdy analizowane dane można uważać za realizację szeregu stacjonarnego. Oznacza to:

- brak trendów (długoterminowych i sezonowych),
- jednorodną wariancję,
- odpowiednio szybko zanikającą funkcję autokorelacji ACF.

Jeśli analizowany szereg wykazuje odstępstwa od stacjonarności, możemy zastosować odpowiednie przekształcenie danych, które pozwolą przekształcić szereg do postaci stacjonarnej:

- metody dekompozycji
- różnicowanie

Innym podejściem jest uwzględnienie składowych regularnych bezpośrednio w modelu, w ten sposób możemy otrzymać modele dla szeregów **niestacjonarnych**.

Model ARIMA(p, d, q)

ARIMA – Auto-Regressive Integrated Moving Averages

Szereg Y_t nazywamy procesem ARIMA(p, d, q), jeśli po d-krotnym $(d \geq 0)$ zróżnicowaniu z opóźnieniem 1 jest już procesem ARMA(p, q).

Materiały na podstawie

- Analiza i prognozowanie szeregów czasowych, Zagdański Adam, Suchwałko Artur, rok wydania 2015, wydawnictwo: Wydawnictwo Naukowe PWN.
- Peter J Brockwell and Richard A Davis. 1986. Time Series: Theory and Methods. Springer-Verlag, Berlin, Heidelberg.