Prognozowanie na podstawie szeregów czasowych.

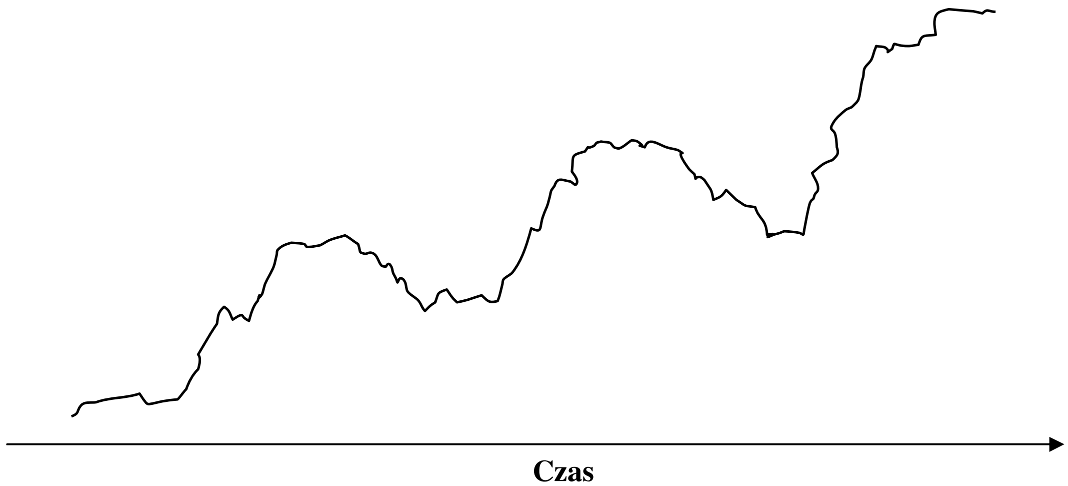
Składowe szeregów czasowych.

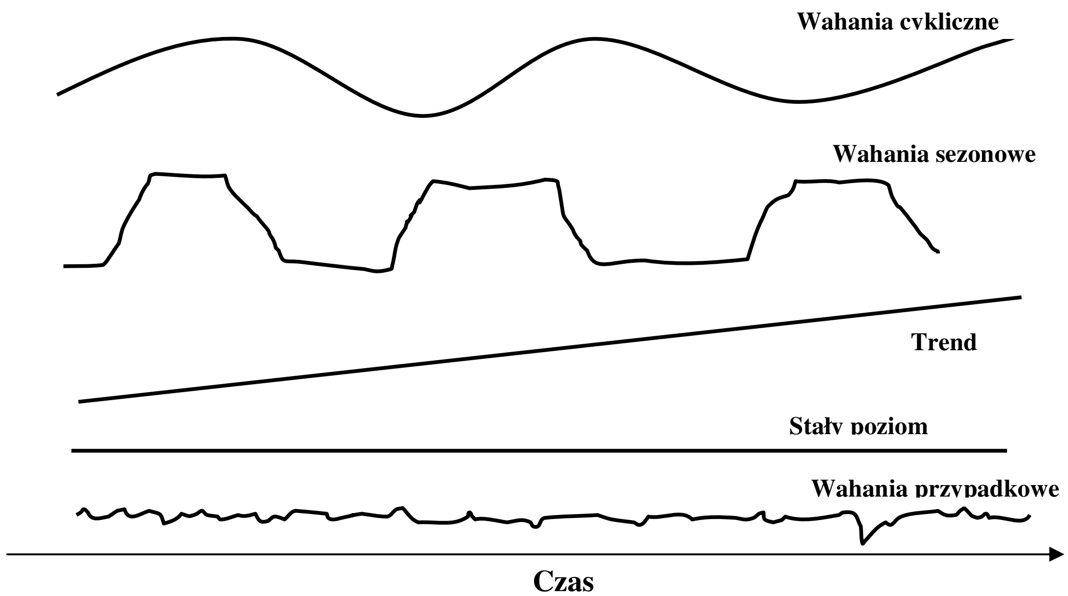
Trend (tendencja rozwojowa) - długookresowa skłonność do jednokierunkowych zmian (wzrostu lub spadku) wartości zmiennej badanej. Jest konsekwencją działania stałych czynników np. w przypadku sprzedaży - liczba potencjalnych klientów, ich dochody lub preferencje. Może być wyznaczony gdy mamy długi ciąg obserwacji.
Stały (przeciętny poziom) - występuje gdy w szeregu czasowym nie ma trendu, zaś wartości badanej zmiennej oscylują wokół pewnego stałego poziomu.

Wahania cykliczne - długookresowe wahania wokół trendu lub stałego poziomu. W ekonomii najczęściej związane z cyklem koniunkturalnym gospodarki.

Wahania sezonowe - wahania wokół trendu lub stałego poziomu. Wahania te mają skłonności do powtarzania się w określonym czasie nie przekraczającym jednego roku, odzwierciedlają wpływ pogody lub kalendarza na działalność gospodarczą.

Dekompozycja szeregu czasowego = wyodrębnienie poszczególnych składowych.





Modele szeregów czasowych.

f(t) - trend,

g(t) - wahania sezonowe,

h(t) - wahania cykliczne,

ξ_t - składnik losowy,

const - stały poziom.

Model **addytywny**:

$$y_t = f(t) + g(t) + h(t) + \xi_t \quad \text{lub} \quad y_t = const + g(t) + h(t) + \xi_t$$

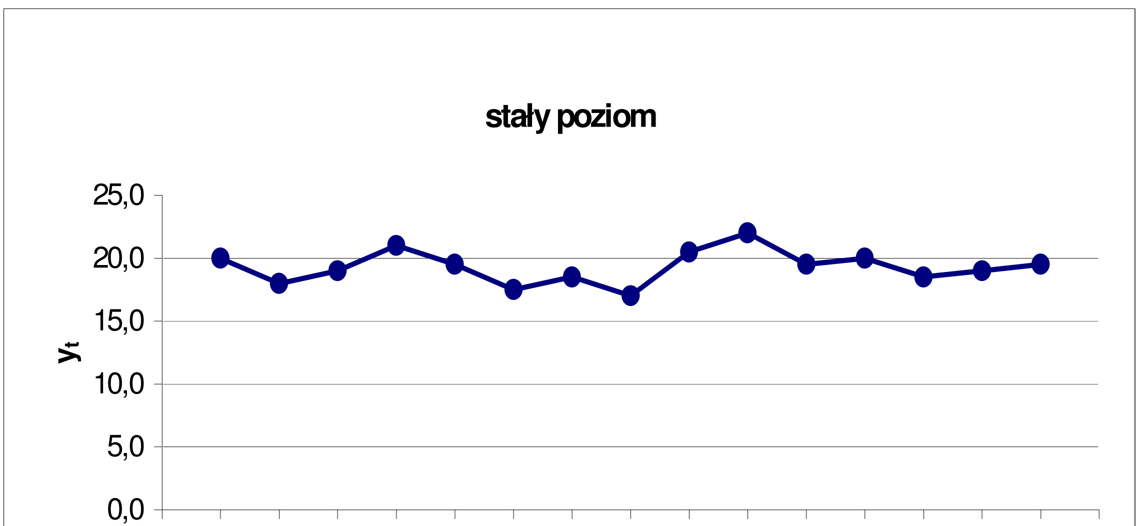
Model **multiplikatywny**:

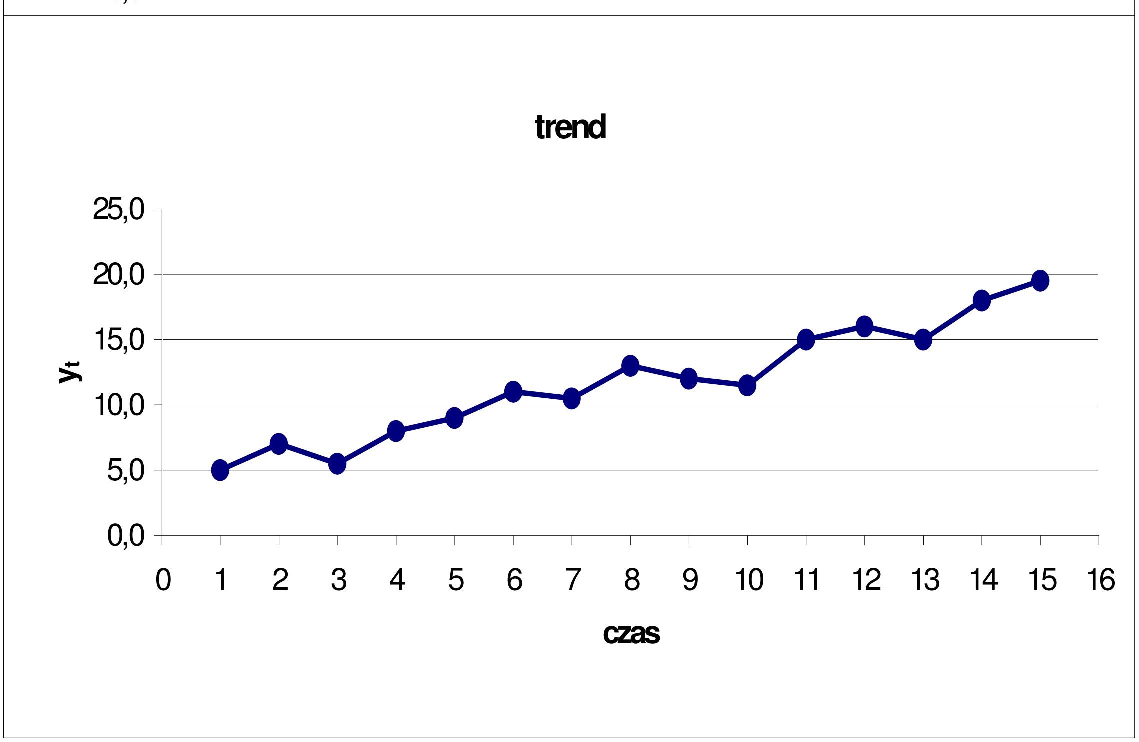
$$y_t = f(t) \cdot g(t) \cdot h(t) \cdot \xi_t \quad \text{lub} \quad y_t = const \cdot g(t) \cdot h(t) \cdot \xi_t$$

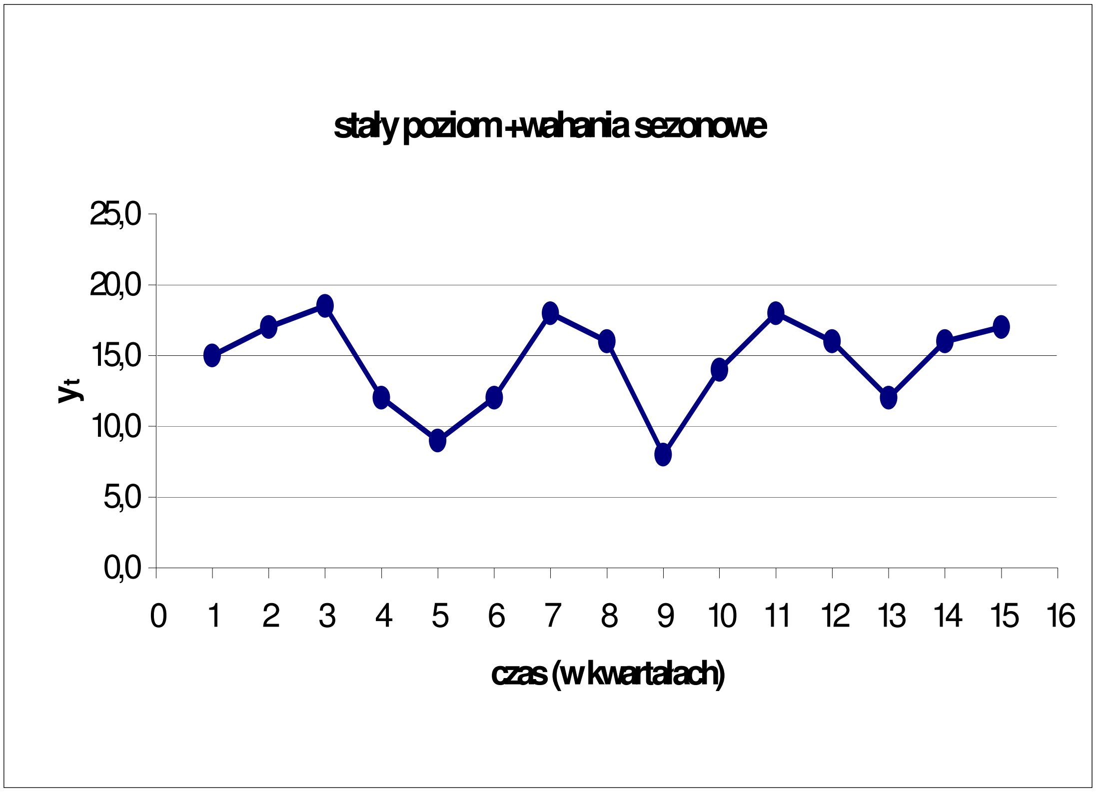
Modele **mieszane**:

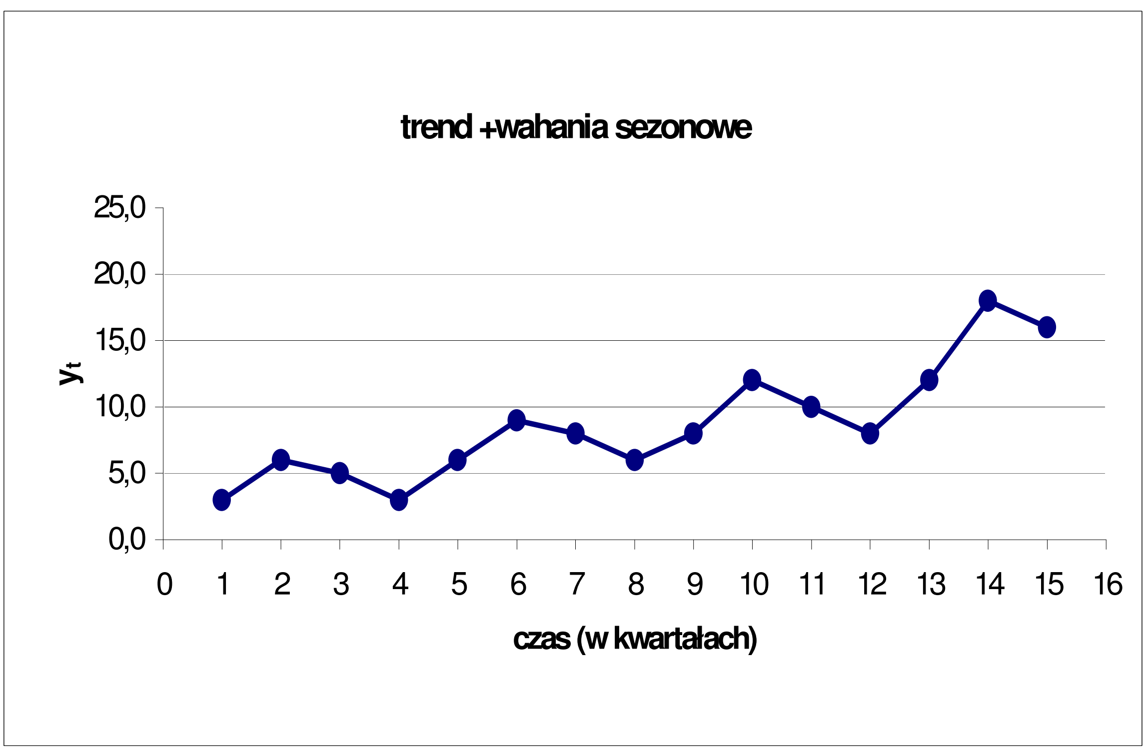
$$y_t = f(t) + g(t) + h(t) \cdot \xi_t \quad \text{lub} \quad y_t = f(t) \cdot h(t) + g(t) \cdot \xi_t$$

$$y_t = f(t) \cdot g(t) \cdot \xi_t + h(t) \quad \text{lub} \quad y_t = f(t) \cdot h(t) \cdot \xi_t + g(t)$$






stały poziom +wahania sezonowe
25,0
20,0
15,0
10,0
5,0
0,0
y_t
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
czas (w kwartałach)


trend +wahania sezonowe
25,0
20,0
15,0
10,0
5,0
0,0
y_t
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
czas (w kwartałach)

Do wyznaczania składowych szeregu czasowego oprócz wykresu przydatny jest ciąg współczynników autokorelacji rzędu k (k od 1 do około połowy liczby danych)

$$r_k = \frac{\sum_{t=1}^{n-k} (y_t - \overline{y}) \cdot (y_{t+k} - \overline{y})}{\sum_{t=1}^{n} (y_t - \overline{y})^2}$$

Jeśli r_k zbliżają się do zera i następnie oscylują wokół niego to szereg czasowy zawiera składową stałą.
Jeśli r_k maleją poprzez zero do wartości ujemnych to szereg ma trend.
Jeśli r_k oscylują wokół zera i co pewien okres są wyraźnie większe od zera to szereg czasowy zawiera wahania sezonowe.

Niekiedy przyjmuje się, że wahania przypadkowe są niewielkie, gdy ich współczynnik zmienności jest rzędu kilku, najwyżej kilkunastu procent.

Prognoza zmiennej Y jest wartością funkcji f zależnej od czasu, przeszłych wartości i prognoz tej zmiennej.

$$y_t^* = f(\underset{\text{czas}}{t}, \underbrace{y_{t-1}, \ldots, y_{t-p}}_{\text{wartości}}, \underbrace{y^*_{t-1}, \ldots, y^*_{t-p}}_{\text{prognozy}}, \underset{\text{składnik losowy}}{\xi_t})$$

p - wielkość opóźnienia.

Uwaga.

Jakość modelu oceniamy jak w ekonometrii.

Dopasowanie wielkości zjawiska wyznaczonych z modelu (prognozy wygasłe) do wielkości zaobserwowanych oceniamy na podstawie:

- **błędu średniokwadratowego** prognoz wygasłych

$$s^* = \sqrt{\frac{1}{k} \sum_{t=1}^{k} \left(y_t - y_t^*\right)^2}$$

k - liczba prognoz wygasłych

wielkość ta określa o ile średnio jednostek prognozy wygasłe odchylają się (plus-minus) od wartości zaobserwowanych.

- **względnego błędu średniokwadratowego** (procentowego błędu średniokwadratowego) prognoz wygasłych

$$s_w^* = \sqrt{\frac{1}{k}\sum_{t=1}^{k}\left(\frac{y_t - y_t^*}{y_t}\right)^2} \cdot 100\%$$

wielkość ta określa o ile średnio procent prognozy wygasłe odchylają się (plus-minus) od wartości zaobserwowanych.

- **średniego błędu względnego** (procentowego błędu względnego) prognoz wygasłych

$$\psi = \frac{1}{k}\sum_{t=1}^{k}\left|\frac{y_t - y_t^*}{y_t}\right| \cdot 100\%$$

(interpretacja jak wyżej)

Często ostatni z tych błędów służy do oceny jakości prognozy.

Naiwne i proste metody prognozowania.

(oparte na założeniu, że wahania przypadkowe są niewielkie i nie zmieni się dotychczasowy wpływ czynników kształtujących obserwowane zjawisko).

Zaletą metody naiwnej jest prostota, wadą brak oceny jakości prognozy na podstawie prognoz wygasłych.

Rodzaje prognoz naiwnych:

- wg stałego poziomu

$$y_{t+1}^* = y_t \qquad \text{lub} \qquad y_{t+1}^* = \frac{y_t + y_{t-1} + y_{t-2}}{3}$$

- wg stałych przyrostów bezwzględnych (np. trend zbliżony do liniowego)

$$y_{t+1}^* = y_t + \left(y_t - y_{t-1}\right)$$

- wg stałych przyrostów względnych (niektóre trendy nieliniowe)

$$y_{t+1}^* = y_t\left(1 + \frac{y_t - y_{t-1}}{y_{t-1}}\right) = \frac{y_t^{\ 2}}{y_{t-1}}$$

- wg wahań sezonowych i stałego poziomu

$y_{t+1}^* = y_{t-r+1}$, gdzie r długość cyklu sezonowego (liczba faz cyklu),

- wg wahań sezonowych i trendu

  $y_{t+1}^{*} = y_{t-r+1} + r\Delta\overline{y}_r$, gdzie r długość cyklu sezonowego (liczba faz cyklu),

  $\Delta\overline{y}_r$ - przyrost średnich w dwóch ostatnich cyklach.

Przykład.

Dla poszczególnych serii danych miesięcznych wyznacz prognozę naiwną na kolejny miesiąc.

a) 115, 119, 126, 131, 136,

b) 1, 4, 12, 24, 72,

c) 125, 120, 110, 115, 120.

Przykład.

Wiedząc, że zjawisko ma charakter sezonowy (r = 4), dla poszczególnych serii danych kwartalnych wyznacz prognozę naiwną na dwa kolejne kwartały.

a) 150, 200, 160, 120, 170,

b) 150, 220, 190, 170, 200, 270, 240, 220, 270

Metoda średniej globalnej.

$$y_{n+1}^{*} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Metoda średniej ruchomej.

Metodę tą wykorzystujemy zarówno do wygładzania szeregu czasowego jak i do prognozowania.

Prognoza jest średnią arytmetyczną z k ostatnich obserwacji (k - stała wygładzania).

$$y_t^{*} = \frac{1}{k}\sum_{i=t-k}^{t-1} y_i$$

k wyznaczamy tak aby średni kwadratowy błąd ex post $s^{*^2} = \frac{1}{n-k}\sum_{t=k+1}^{n}\left(y_t - y_t^{*}\right)^2$ był minimalny.

Prognozę oceniamy za pomocą średniego błędu względnego prognoz przeszłych

$$\Psi_k = \frac{1}{n-k}\sum_{t=k+1}^{n}\frac{\left|y_t - y_t^{*}\right|}{y_t}100\%$$

Uwaga.

Gdy k = 1 to metoda naiwna.

Gdy k = n to średnia globalna.

Gdy k duże to średnia ruchoma silniej wygładza szereg czasowy lecz jednocześnie wolniej reaguje na zmiany poziomu badanego zjawiska.

Gdy k małe to średnia ruchoma szybciej odzwierciedla zmiany zjawiska lecz większy wpływ wywierają na nią wahania przypadkowe.

Aby stosować średnią ruchomą powinniśmy zwykle dysponować co najmniej kilkunastoma danymi.

Średnia ważona.

Ustalamy wagi $0 < w_1 \leq w_2 \leq .... \leq w_k < 1$ takie, że $\sum_{i=1}^{k} w_i = 1$ (oznacza to, że do wcześniejszych informacji przywiązujemy mniejszą wagę).

Prognozę wyznaczamy na podstawie wzoru:

$$y_{t+1}^{*} = \sum_{i=t-k+1}^{t} w_{i-t+k}\, y_i$$

Model Browna (prosty model wygładzania wykładniczego).

Zwykle stosujemy ten model dla szeregów czasowych o stałym poziomie lub bardzo słabym trendzie i umiarkowanych wahaniach przypadkowych.

Model pozwala wyznaczyć prognozę wg wzoru:

$$y_t^{*} = \alpha y_{t-1} + (1-\alpha) y_{t-1}^{*}, \qquad t = 2, 3, ..., n + 1$$

prognoza jest kombinacją wypukłą (średnią ważoną) przeszłej wartości zjawiska i przeszłej prognozy. $\alpha \in \langle 0,1 \rangle$ – parametr wygładzania.

Wartość α dobieramy np. na podstawie kryterium najmniejszego błędu średniokwadratowego prognoz wygasłych s^* tzn. $\min\limits_{\alpha} s^*(\alpha)$ gdzie

$$s^* = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - y_t^{*}(\alpha)\right)^2}$$

Jeśli nie mamy możliwości wyznaczenia optymalnej wartości parametru wygładzania zwykle zaleca się stosowania wartości 0,1 – 0,3.

Uwaga

Równoważny wzór na prognozę w tym modelu ma postać:

$$y_t^{*} = y_{t-1}^{*} + \alpha(y_{t-1} - y_{t-1}^{*})$$

zatem dla małych α prognoza w małym stopniu uwzględnia błąd ex post prognoz przeszłych.

Uwaga
Jako wartość y_1^* przyjmujemy jedną z wartości:

a) pierwszą wartość szeregu czasowego, $y_1^* = y_1$,

b) średnią z trzech początkowych wartości szeregu czasowego, $y_1^* = \dfrac{y_1 + y_2 + y_3}{3}$ **,**

c) średnią z pięciu początkowych wartości szeregu czasowego, $y_1^* = \dfrac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$.

Model Browna jest rozwinięciem metody średnich ważonych.
Wagi maleją wykładniczo przy coraz starszych danych.
Widać to gdy przekształcimy wzór na prognozę w tym modelu:

$$y_n^* = \alpha y_{n-1} + (1-\alpha)\, y_{n-1}^*$$

podstawiając $y_{n-1}^* = \alpha y_{n-2} + (1-\alpha)\, y_{n-2}^*$ otrzymamy

$$y_n^* = \alpha y_{n-1} + (1-\alpha)(\alpha y_{n-2} + (1-\alpha)\, y_{n-2}^*) = \alpha y_{n-1} + \alpha(1-\alpha)\, y_{n-2} + (1-\alpha)^2\, y_{n-2}^*$$

następnie podstawiając $y_{n-2}^* = \alpha y_{n-3} + (1-\alpha)\, y_{n-3}^*$ otrzymamy

$$y_n^* = \alpha y_{n-1} + \alpha(1-\alpha)\, y_{n-2} + (1-\alpha)^2 (\alpha y_{n-3} + (1-\alpha)\, y_{n-3}^*) =$$
$$= \alpha y_{n-1} + \alpha(1-\alpha)\, y_{n-2} + \alpha(1-\alpha)^2\, y_{n-3} + (1-\alpha)^3\, y_{n-3}^*$$

ostatecznie

$$y_n^* = \alpha y_{n-1} + \alpha(1-\alpha)\, y_{n-2} + ... + \alpha(1-\alpha)^k\, y_{n-k+1} + ...$$

Wagi przy poszczególnych elementach szeregu czasowego

$$\alpha > \alpha(1-\alpha) > .... > \alpha(1-\alpha)^k > ...$$

stanowią kolejne wyrazy ciągu geometrycznego o ilorazie $0 < 1-\alpha < 1$. Dla dużych n ich suma jest prawie równa 1 bowiem

$$\alpha + \alpha(1-\alpha) + ... + \alpha(1-\alpha)^k + ... = \frac{\alpha}{1-(1-\alpha)} = 1$$

Uwaga.

a) jeśli wygładzenie szeregu czasowego (zwłaszcza dla dużych α) nie jest zadawalające to możemy powyższe wygładzanie powtórzyć,

b) chociaż dla małych α wygładzenie jest lepsze, to nie zawsze wtedy jest najmniejszy błąd średniokwadratowy dla prognoz przeszłych, widać to w następującym przykładzie

Przykład (*L. Kowalski*, „Statystyka”, 2003, 2005), .
Liczba sprzedanych żarówek (tys. szt.) w hurtowni „LUMEN” w kolejnych kwartałach lat 1998-2000:

37, 36, 34, 33, 34, 33, 35, 34, 35, 33, 34, 36
Badając wielkość błędu średniokwadratowego dla różnych wartości α otrzymamy:

α	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
błąd	1,47	1,39	1,36	1,35	1,35	1,36	1,37	1,38	1,41

Jak widać najlepsze (z tego punktu widzenia) wartości α są w przedziale 0,4 ÷ 0,5.