Analiza danych

prof. UAM dr hab. Tomasz Górecki

tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Wprowadzenie

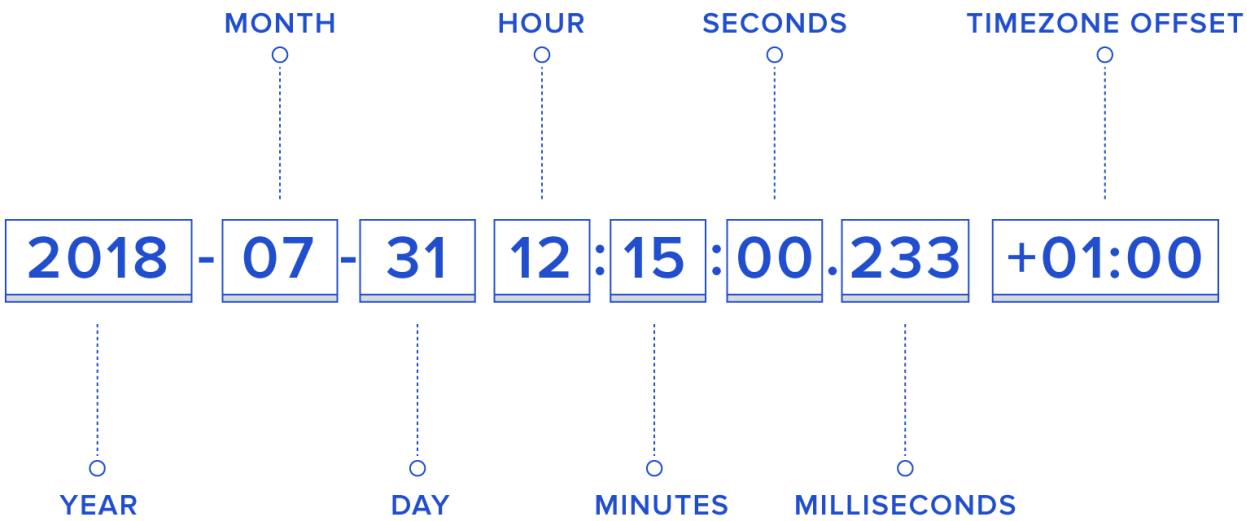
Analizę dynamiki zjawisk masowych przeprowadza się na podstawie szeregów czasowych (*ang. time series*). Są to ciągi (Y_t) wartości badanego zjawiska obserwowanego w kolejnych jednostkach czasu. Zmienną niezależną jest czas, a zmienną zależną – wartości liczbowe badanego zjawiska.

Szeregi czasowe w R

Do konstrukcji szeregu czasowego wykorzystywana jest funkcja ts. Jeśli mamy już szereg czasowy, to możemy uzyskać z niego wiele informacji. Wykorzystywane są do tego następujące funkcje: start (początkowy okres), end (końcowy okres), frequency (liczba podokresów), deltat (odstęp czasowy pomiędzy obserwacjami, np. dla miesięcy mamy 1/12), time (wektor czasów, w których mamy obserwacje z szeregu). Do wizualizacji danych zebranych w postaci szeregu czasowego służy funkcja ts.plot, której argumentem może być kilka szeregów czasowych (zostaną zwizualizowane na jednym wykresie).

Daty w R

Podstawowe funkcje to: Sys.time (data wraz z godziną), Sys.Date (data bez godziny). Do wprowadzania danych jako dat służy funkcja as.Date, której argumentem jest data. Domyślny format daty, to cztery cyfry na rok, dwie na miesiąc i dwie na dzień, oddzielone kreską lub ukośnikiem. Jeśli chcemy użyć niestandardowego formatu, należy go wyspecyfikować jako wartość parametru format według oznaczeń zawartych w poniższej tabeli.



Daty w R

Symbol	Meaning	Example
%a	Abbreviated weekday name	Tue
%A	Full weekday name	Tuesday
%b	Abbreviated month name	Apr
%B	Full month name	April
%C	Century: the integer part of the year divided by 100	20
%d	Day of the month	09
%H	Hours as decimal number (00–23)	13
%I	Hours as decimal number (01–12)	1
%m	Month as number (01–12)	04
%M	Minute as number (00–59)	12
%p	AM/PM indicator for 12-hour time (%I)	PM
%S	Second as integer (00–61)	12
%u	Weekday as a decimal number (1–7, Monday is 1)	2
%w	Weekday as decimal number (0–6, Sunday is 0)	2
%y	2-Digit Year (00–99)	19
%Y	4-Digit Year	2019

Daty w R

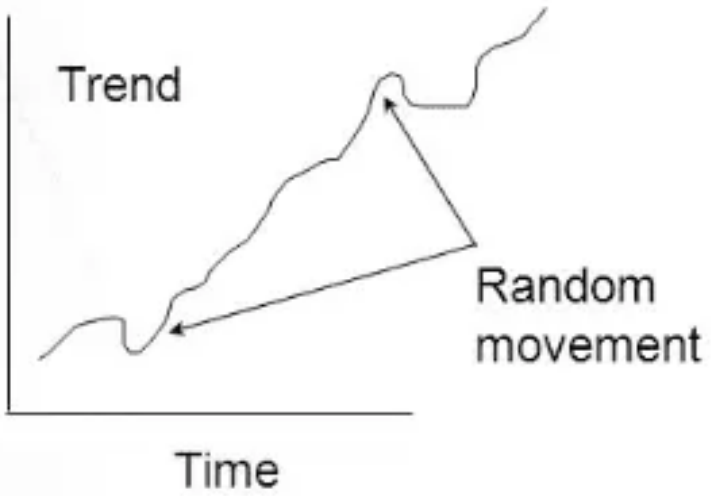
Data przechowywana jest jako liczba dni, jaka upłynęła od 1 stycznia 1970 roku (POSIXct). Można również podać datę jako liczbę dni, która upłynęła od pewnej daty początkowej. Jeśli chcemy się dowiedzieć, jakim dniem, miesiącem lub kwartałem jest dana data możemy użyć funkcji weekdays, months oraz quarters. Często możemy być zainteresowani jaka była różnica pomiędzy dwoma datami. W R różnicę tę możemy wyrazić w sekundach, minutach, godzinach, dniach i miesiącach używając funkcji difftime i określając parametr units na *secs*, *mins*, *hours*, *days*, *weeks* odpowiednio. Przy konstrukcji szeregów czasowych potrzebne nam są sekwencje dat. Można je z łatwością utworzyć korzystająca z poznanej wcześniej funkcji seq z wykorzystaniem jej parametru by, który może przyjmować wartości będące jednostkami czasowymi.

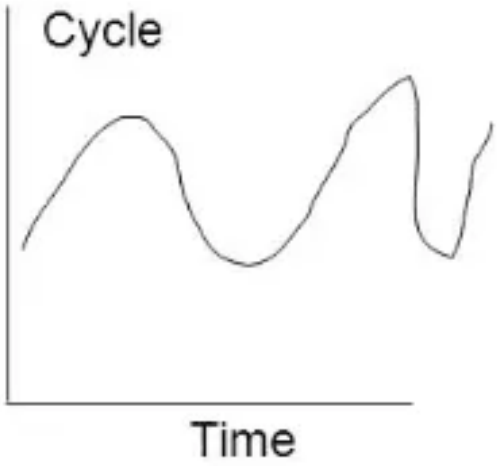
Model wahań w czasie

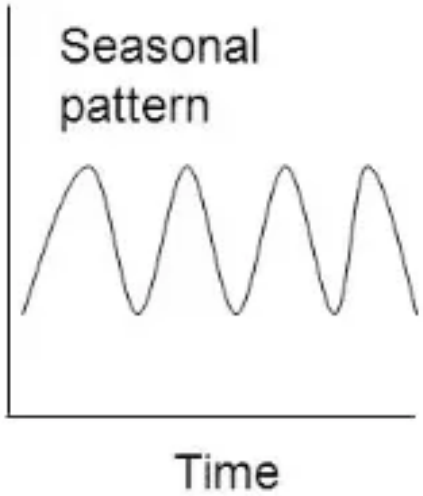
Modelem wahań w czasie nazywamy konstrukcję teoretyczną opisującą kształtowanie się określonego zjawiska jako funkcję czasu, wahań okresowych (periodycznych) i przypadkowych (nieregularnych). Tradycyjnie analizy prawidłowości w rozwoju zmiennej dokonuje się poprzez wyodrębnianie w szeregu czasowym jego elementów składowych, co nosi nazwę dekompozycji tego szeregu. W najogólniejszym przypadku zakłada się, że w szeregu czasowym mogą wystąpić cztery składniki:

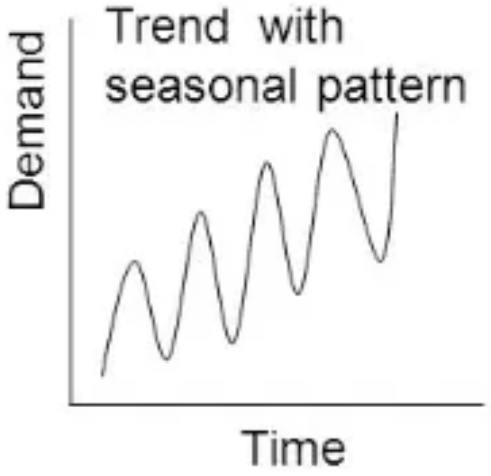
1. trend – T_t,
2. wahania cykliczne – C_t,
3. wahania sezonowe – S_t,
4. wahania nieregularne, przypadkowe – I_t.

Model wahań w czasie





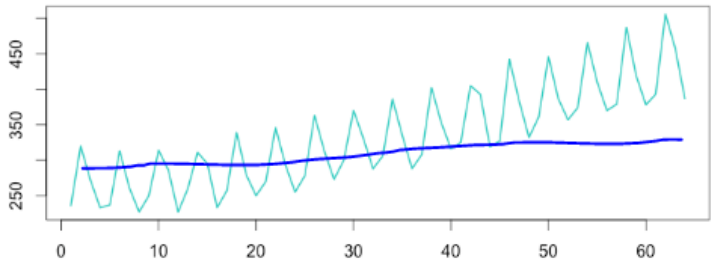




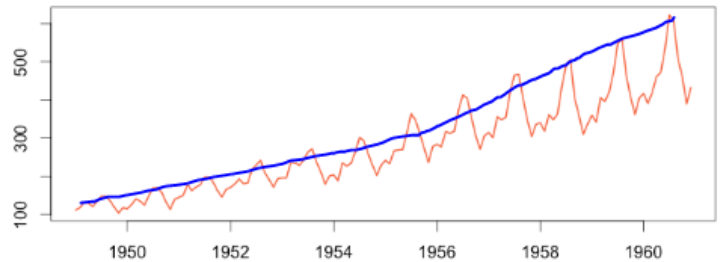
Model wahań w czasie

Wyróżniamy dwa podstawowe modele nakładania się na siebie poszczególnych elementów czyli powstawania szeregu czasowego:

- model addytywny: wahania (sezonowe i przypadkowe) są stałe i sumują się z trendem.
- model multyplikatywny: wahania są proporcjonalne do poziomu zjawiska co znaczy, że nakładają się na trend w sposób iloczynowy.



Australian beer production – The seasonal variation looks constant; it doesn't change when the time series value increases. We should use the **additive model.**



Airline Passenger Numbers – As the time series increases in magnitude, the seasonal variation increases as well. Here we should use the **multiplicative model.**

Additive:
Time series = Seasonal + Trend + Random

Multiplicative:
Time series = Trend * Seasonal *Random

Model wahań w czasie

Trend charakteryzuje długookresową tendencję zmian w szeregu czasowym. Może on oznaczać w miarę regularnie powtarzający się wzrost lub spadek wartości zmiennej Y lub też brak wyraźnej tendencji zmian. Pozostałe trzy składniki szeregu czasowego to różnego typu odchylenia od tendencji długookresowej. Wahania cykliczne oznaczają powtarzające się (niekoniecznie regularnie) wahania o czasie trwania dłuższym niż rok. Wahania sezonowe oznaczają takie odchylenia od trendu, które powtarzają się w czasie w sposób regularny i których pełen cykl zawiera się w ciągu jednego roku. Wahania sezonowe powtarzają się według pewnego „wzorca" każdego roku. Wahania sezonowe kształtowane są przez czynniki naturalne (pory roku, pogodę) oraz przez zwyczaje (np. różne święta). Wahania nieregularne (losowe) to te, które obejmują wszelkie odchylenia od trendu, będące efektem działania na badaną zmienną niepowtarzalnych, nie dających się przewidzieć ani prognozować zdarzeń.

Trend

Tendencją rozwojową (trendem) nazywamy powolne, regularne i systematyczne zmiany określonego zjawiska, obserwowane w dostatecznie długim przedziale czasu i będące wynikiem działania przyczyn głównych. Przyjmuje się, że aby wyodrębnić trend, niezbędne są co najmniej 10-letnie badania. Wyróżniamy dwie metody wyodrębniania tendencji rozwojowej szeregów czasowych:

- Metoda mechaniczna – opiera się na średnich ruchomych. Polega ona na zastąpieniu danych empirycznych średnimi poziomami z okresu badanego i kilku okresów sąsiednich. Średnie ruchome mogą być obliczane z parzystej bądź nieparzystej liczby wyrazów sąsiednich.
- Metoda analityczna – polega na dopasowaniu określonej funkcji matematycznej do całego szeregu czasowego za pomocą MNK. Istotny jest wybór klasy funkcji trendu oraz prawidłowe oszacowanie jej parametrów.

Średnia ruchoma

Dla przykładu średnie ruchome trzyokresowe ($k = 3$) obliczamy następująco:

$$\bar{Y}_2 = \frac{Y_1 + Y_2 + Y_3}{3},$$
$$\bar{Y}_3 = \frac{Y_2 + Y_3 + Y_4}{3},$$
$$\ldots,$$
$$\bar{Y}_{n-1} = \frac{Y_{n-2} + Y_{n-1} + Y_n}{3}.$$

Średnia ruchoma

Natomiast w przypadku średniej ruchomej dla parzystej liczby okresów ($k = 4$) obliczenia wykonujemy według wzorów:

$$\bar{Y}_3 = \frac{\frac{1}{2}Y_1 + Y_2 + Y_3 + Y_4 + \frac{1}{2}Y_5}{4},$$

$$\bar{Y}_4 = \frac{\frac{1}{2}Y_2 + Y_3 + Y_4 + Y_5 + \frac{1}{2}Y_6}{4},$$

$$\ldots,$$

$$\bar{Y}_{n-2} = \frac{\frac{1}{2}Y_{n-4} + Y_{n-3} + Y_{n-2} + Y_{n-1} + \frac{1}{2}Y_n}{4}.$$

Zaletą tej metody jest prostota obliczeń, wadą natomiast jest skracanie wyrównanego tą metodą szeregu czasowego. W naszym przypadku dla $k = 3$ tracimy element pierwszy i ostatni, a dla $k = 4$ tracimy dwa pierwsze i dwa ostatnie.

Filtry wygładzające

Oprócz najprostszej metody średniej ruchomej można zastosować, dużo bardziej wyrafinowane metody zwane filtrami. Do najpopularniejszych należą filtr liniowy oraz wykładniczy. Filtr liniowy ma postać:

$$\hat{Y}_t = \frac{1}{2a+1} \sum_{i=-a}^{a} Y_{t+i}.$$

Jest to w zasadzie nieco zmodyfikowana średnia ruchoma.

Filtry wygładzające

Filtr wykładniczy, który bywa również nazywany wygładzaniem wykładniczym Browna, opiera się na założeniu, że wartość szeregu czasowego powinna bardziej zależeć od obserwacji bliskich niż dalekich, co daje

$$\hat{Y}_{t+1} = \alpha Y_{t-1} + (1 - \alpha)\hat{Y}_t.$$

Istotny jak widać jest w tym przypadku wybór wartości startowej, najczęściej jest za nią przyjmowana wartość początkowa szeregu Y_1 lub jest to średnia z pierwszych czterech lub pięciu obserwacji początkowych. Takie proste wygładzanie wykładnicze używane jest w przypadku prognoz krótkoterminowych, gdy dane nie wykazują trendu ani sezonowości.

Filtry wygładzające

W przypadku wystąpienia trendu używa się podwójnego wygładzania wykładniczego Holta postaci:

$$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + b_{t-1}), \quad 0 < \alpha < 1$$
$$b_t = \beta(S_t - S_{t-1}) + (1-\beta)b_{t-1}, \quad 0 < \beta < 1$$

oraz

$$\hat{Y}_{t+1} = S_t + b_t$$

gdzie S_t jest wygładzoną wartością zmiennej prognozowanej w chwili t, a b_t wygładzoną wartością przyrostu trendu w okresie t. Za wartości startowe przyjmuje się $S_1 = Y_1$ oraz $b_1 = Y_2 - Y_1$ lub $b_1 = (Y_n - Y_1)/(n-1)$. Jeśli dodatkowo uwzględnimy sezonowość to dostaniemy potrójne wygładzanie wykładnicze, zwane również metodą Wintersa (pojawia się tam dodatkowy parametr γ). Ogólnie wygładzenie wykładnicze jest nazywane filtrem Holta-Wintersa.

Filtry wygładzające w R

Filtr liniowy realizuje funkcja filter, której pierwszym argumentem jest szereg czasowy, natomiast drugim argumentem jest wektor wag. Filtrowanie wykładnicze realizuje funkcja HoltWinters, której pierwszym argumentem jest szereg czasowy, następne trzy parametry alpha, beta i gamma określają wartości odpowiednich parametrów modelu. Jeśli nie zostaną podane (ustalenie na *NULL* wyklucza parametr z modelu), funkcja poszuka wartości minimalizujących błąd średniokwadratowy predykcji.

Metoda analityczna

Najczęściej stosowana jest funkcja liniowa postaci:

$$Y_t = \alpha_0 + \alpha_1 \cdot t + \varepsilon_t,$$

gdzie ε_t oznacza składnik losowy. Na podstawie danych z szeregu empirycznego wyznacza się oszacowanie tej funkcji:

$$\hat{Y}_t = a_0 + a_1 \cdot t,$$

gdzie estymatory parametrów wyznaczamy według wzorów:

$$a_1 = \frac{12 \sum_{t=1}^{n} Y_t \cdot t}{n^3 - n} - \frac{6 \sum_{t=1}^{n} Y_t}{n^2 - n},$$
$$a_0 = \bar{Y} - a_1 \cdot \bar{t}.$$

Autokorelacja

Aby powyższe wzory były poprawne, odchylenia resztowe muszą być losowe oraz nie może występować autokorelacja (ACF) (*ang. autocorrelation*) składnika losowego. 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. Autokorelacja rzędu k (popularnie zwana opóźnieniem) jest funkcją, która argumentowi naturalnemu k przypisuje wartość współczynnika korelacji Pearsona pomiędzy szeregiem czasowym, a tym samym szeregiem cofniętym o k jednostek czasu. Formalnie (dla procesów stacjonarnych):

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

gdzie

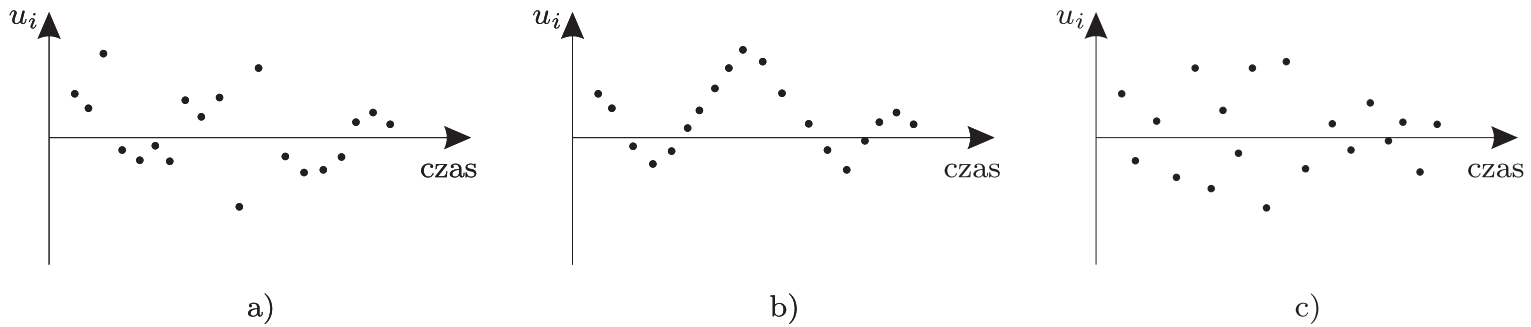
$$\gamma(k) = cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$$

jest autokowariancją rzędu k oraz $\mu = E(Y_t)$.

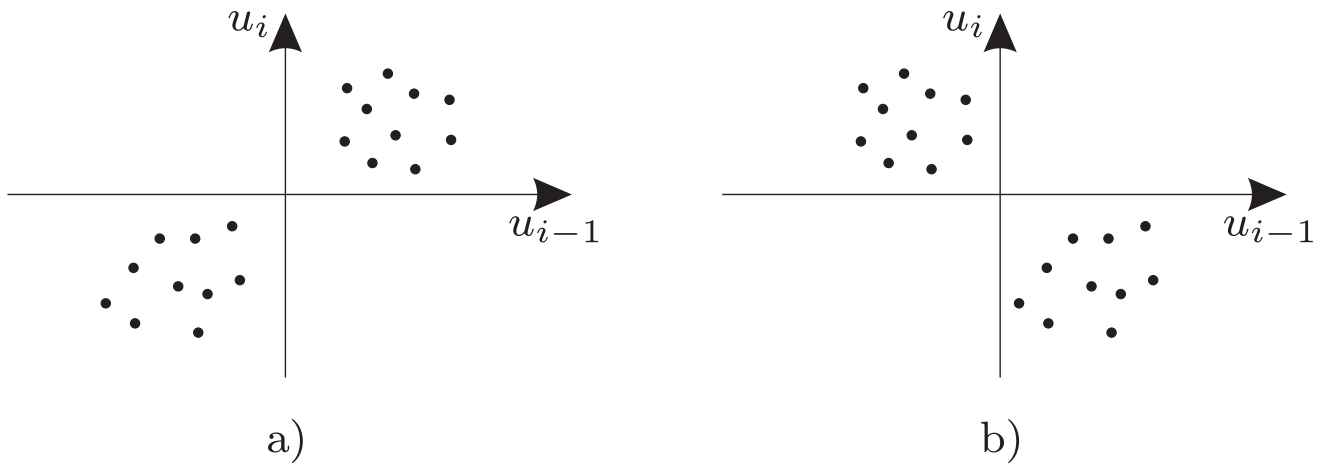
Autokorelacja

Najczęściej spotykaną formą autokorelacji jest autokorelacja dodatnia. Dodatnio skorelowane zaburzenia losowe nie zachowują się całkowicie chaotycznie. Jeśli w okresie t błąd losowy był dodatni, to prawdopodobieństwo, że w okresie $t+1$ będzie on także dodatni, jest wyższe niż prawdopodobieństwo, że w okresie tym będzie on ujemny. Spowodowana jest ona zwykle rozciągnięciem na dłużej niż jeden okres skutków zdarzeń losowych wpływających na poziom zmiennej objaśnianej. Rzadziej spotykaną formą autokorelacji jest autokorelacja ujemna. W takim przypadku prawdopodobieństwo wystąpienia po dodatnim błędzie losowym ujemnego błędu jest wyższe niż prawdopodobieństwo wystąpienia dodatniego błędu. Autokorelacja może być także spowodowana przyjęciem błędnej postaci funkcyjnej dla estymowanego modelu. Sprawdzenie istotności autokorelacji składnika losowego następuje najczęściej za pomocą testu Durbina-Watsona, w którym hipoteza zerowa zakłada brak autokorelacji.

Autokorelacja – wykresy



a) Brak autokorelacji b) autokorelacja dodatnia c) autokorelacja ujemna



a) Autokorelacja dodatnia b) autokorelacja ujemna

Modelowanie szeregów czasowych z autokorelacją

Jeśli autokorelacja występuje szereg czasowy modeluje się poprzez:

- Proces średniej ruchomej (MA) rzędu q postaci:

$$Y_t = c + \sum_{j=0}^{q} \beta_j \varepsilon_{t-j},$$

gdzie ε_t jest czynnikiem losowym (o wartości oczekiwanej 0 oraz wariancji σ^2), przy czym ε_i oraz ε_{i+1} są niezależne dla każdej wartości i.

Modelowanie szeregów czasowych z autokorelacją

- Proces autoregresji (AR) rzędu p postaci:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t.$$

W procesie AR(p) uwzględniamy wpływ p poprzednich wartości szeregu na jego wielkość w momencie t.

Modelowanie szeregów czasowych z autokorelacją

- Proces autoregresji i średniej ruchomej (ARMA) rzędu (p, q) postaci:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=0}^{q} \beta_j \varepsilon_{t-j},$$

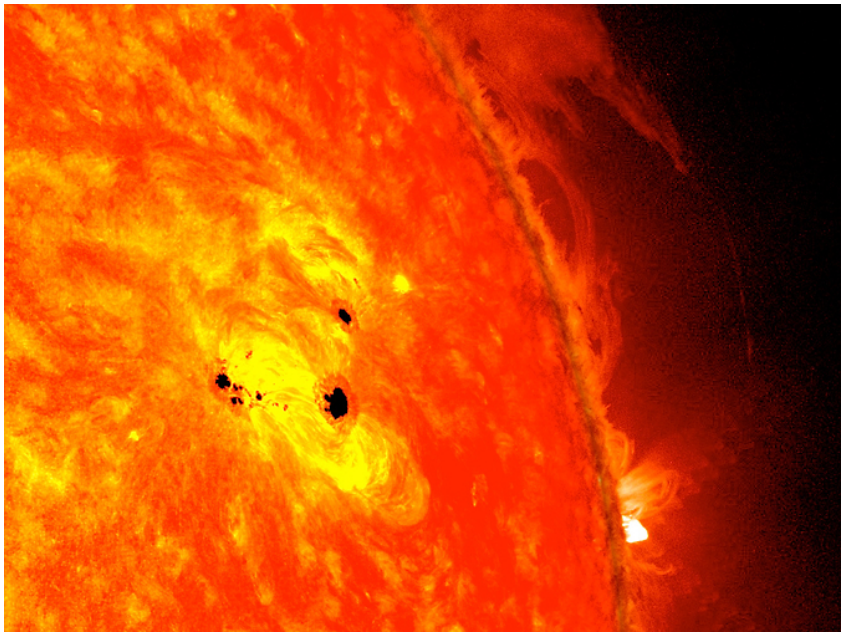
w którym dodajemy dodatkowo efekt wpływu czynnika losowego z poprzednich momentów czasowych na wartość szeregu w momencie t.

Modelowanie szeregów czasowych z autokorelacją

- Zintegrowany proces autoregresji i średniej ruchomej (ARIMA) rzędu (p, d, q). Jeśli w danych występuje wyraźny trend (proces jest niestacjonarny), należy taki trend usunąć przed dalszą analizą. Trend usuwany jest poprzez różnicowanie d razy. Stopień różnicowania określony jest przez stopień wielomianu opisującego trend (pojedyncze różnicowanie usuwa trend liniowy, podwójne kwadratowy itd.). Operacja różnicowania polega na d krotnym zastępowaniu szeregu szeregiem różnic wyrazów sąsiednich. Przy każdej takiej operacji długość szeregu zmniejsza się o jeden. Gdy metodą różnicowania dojdziemy do szeregu stacjonarnego obliczając różnice rzędu d, taki szereg nazywamy szeregiem zintegrowanym stopnia d.

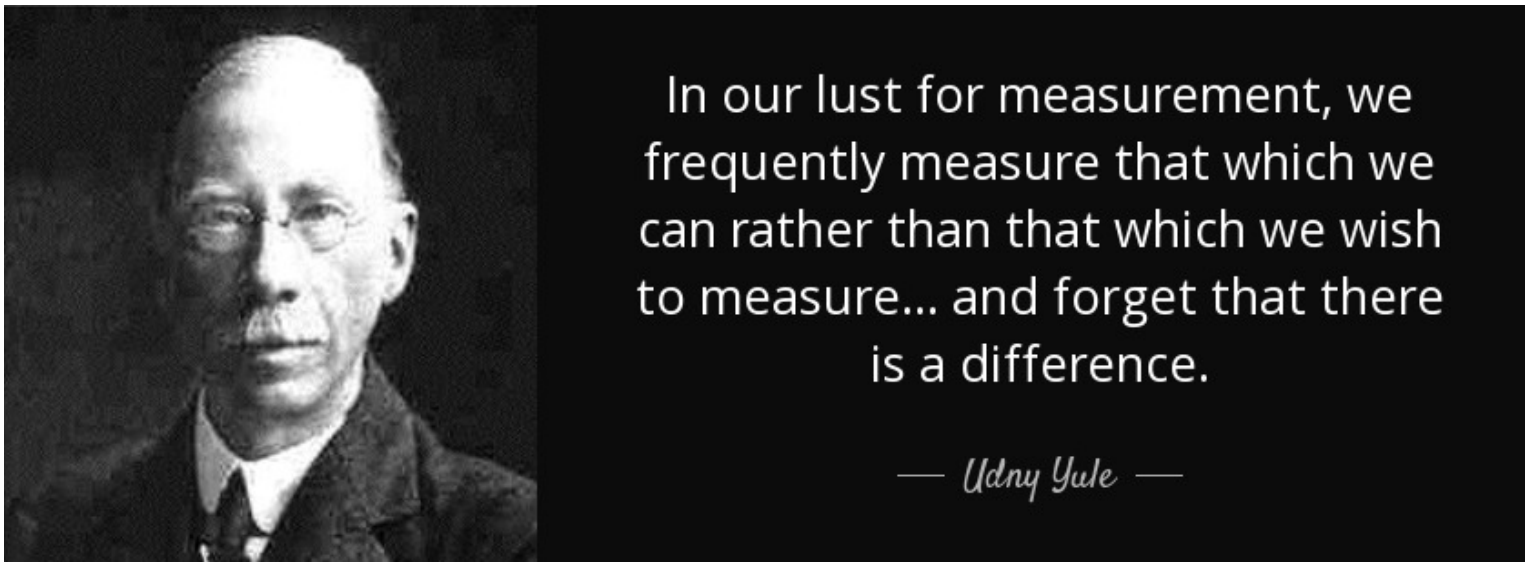
Modelowanie szeregów czasowych z autokorelacją

Pierwsze użycie modelu autoregresyjnego zostało podane przez YULE'a w 1927 roku. Zastosował on ten model do modelowania szeregów czasowych liczby plam na Słońcu.



George Udny Yule
(1871-1951)

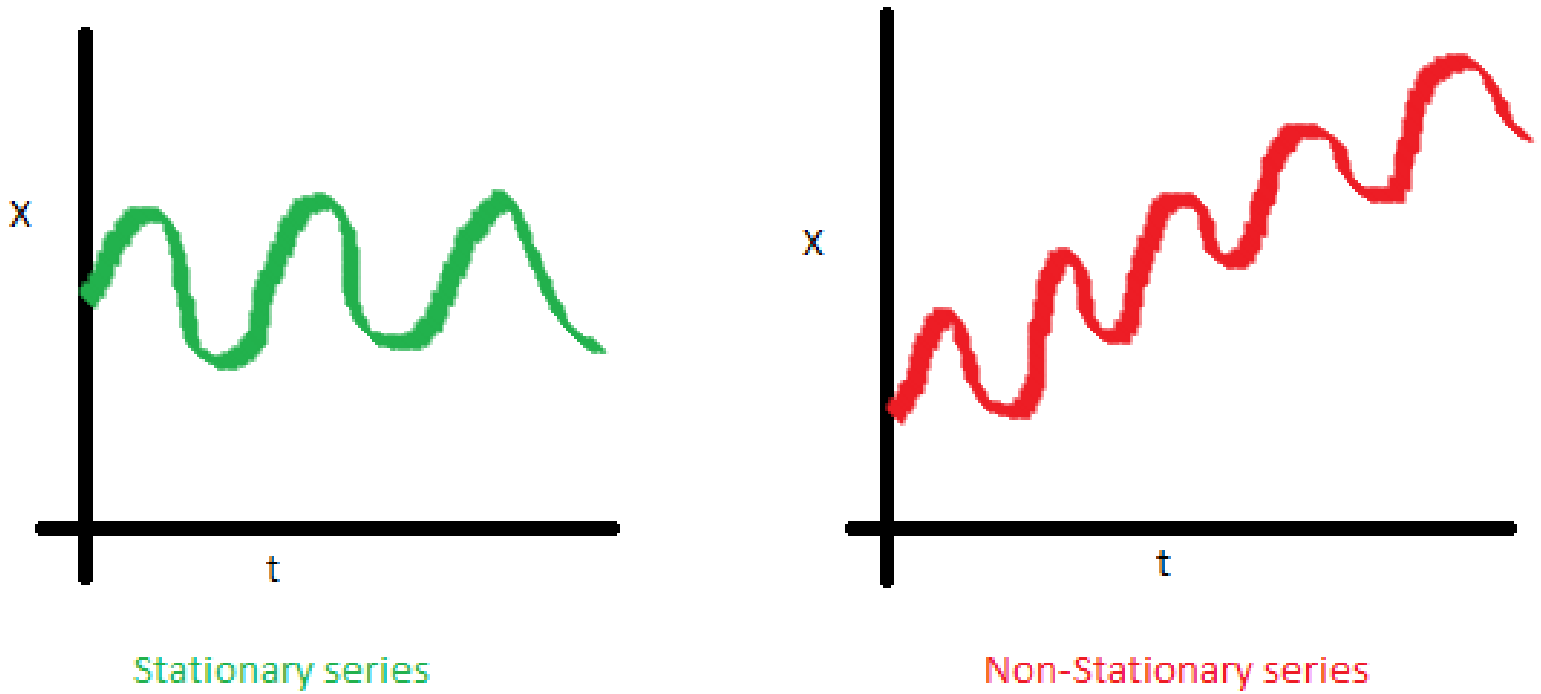
Modelowanie szeregów czasowych z autokorelacją



Bibliografia

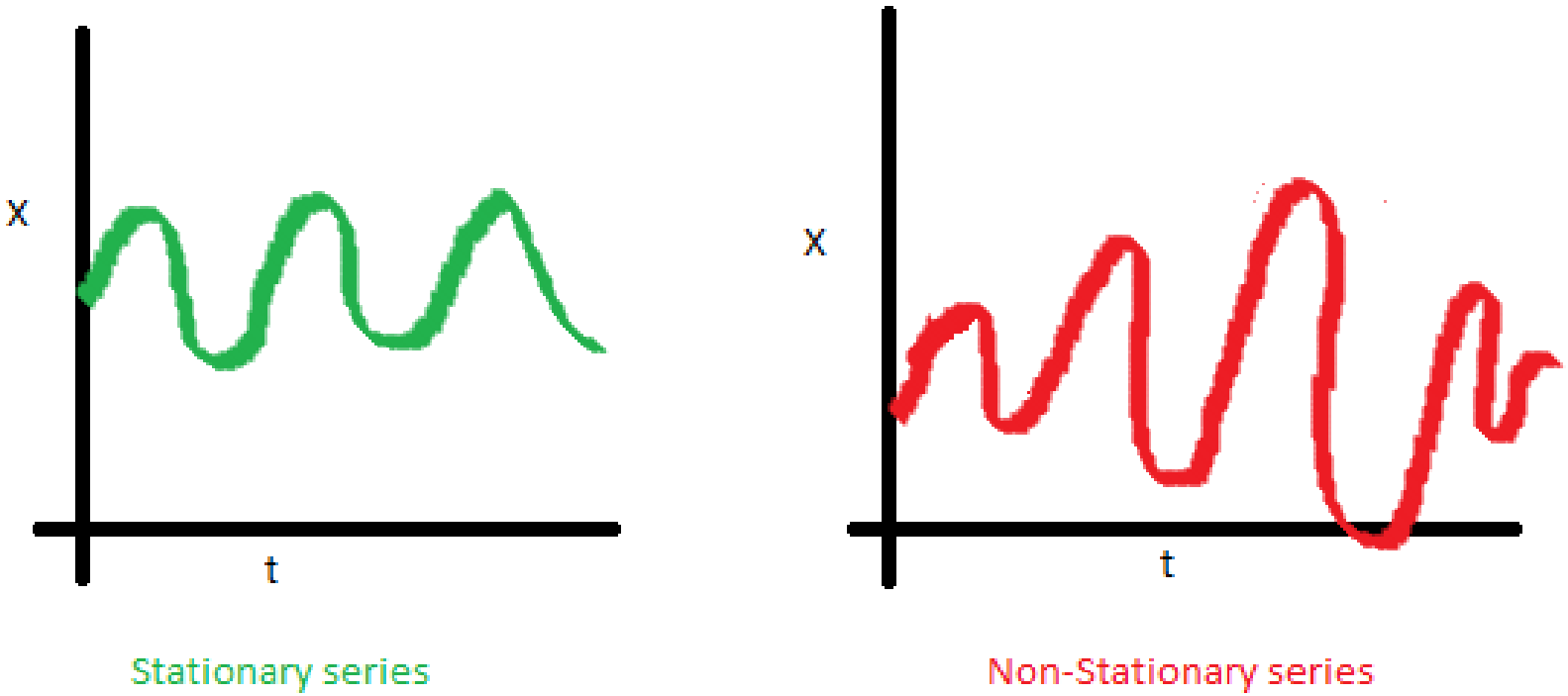
Yule, G.U. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society A* 226:267–298.

Stacjonarność



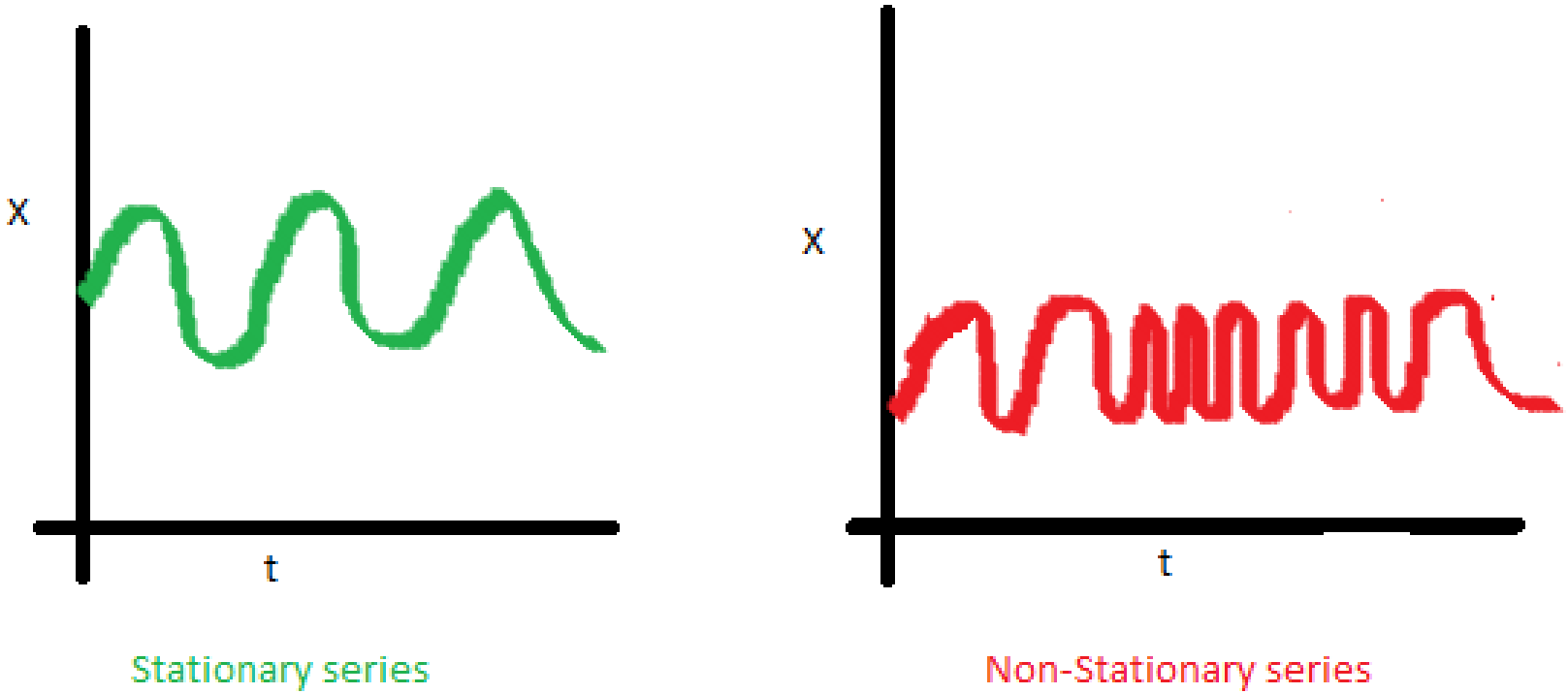
Średnia szeregu czasowego (trend) nie powinna być funkcją czasu, raczej powinna być stała.

Stacjonarność



Wariancja szeregu czasowego nie powinna być funkcją czasu.

Stacjonarność



Kowariancja i-tego i $(i+m)$-ego składnika nie powinna być funkcją czasu.

Stacjonarność

Jeśli nie mamy pewności co do stacjonarności szeregu, możemy spróbować zbadać to jednym z dostępnych testów stacjonarności. Do najpopularniejszych należy test Dickeya-Fullera (DF) i jego rozszerzona wersja (ADF). Hipoteza zerowa stanowi, że szereg jest niestacjonarny. Również popularny jest test Kwiatkowskiego-Phillipsa-Schmidta-Shina (KPSS). W nim hipoteza zerowa stanowi, że szereg czasowy jest stacjonarny. Proces AR oraz ARMA są stacjonarne jeżeli wszystkie pierwiastki równania charakterystycznego są większe co do wartości bezwzględnej od 1. Jeżeli w modelu zawarto funkcję zależną od czasu t, to proces jest niestacjonarny. Proces MA jest zawsze stacjonarny.

Stacjonarność

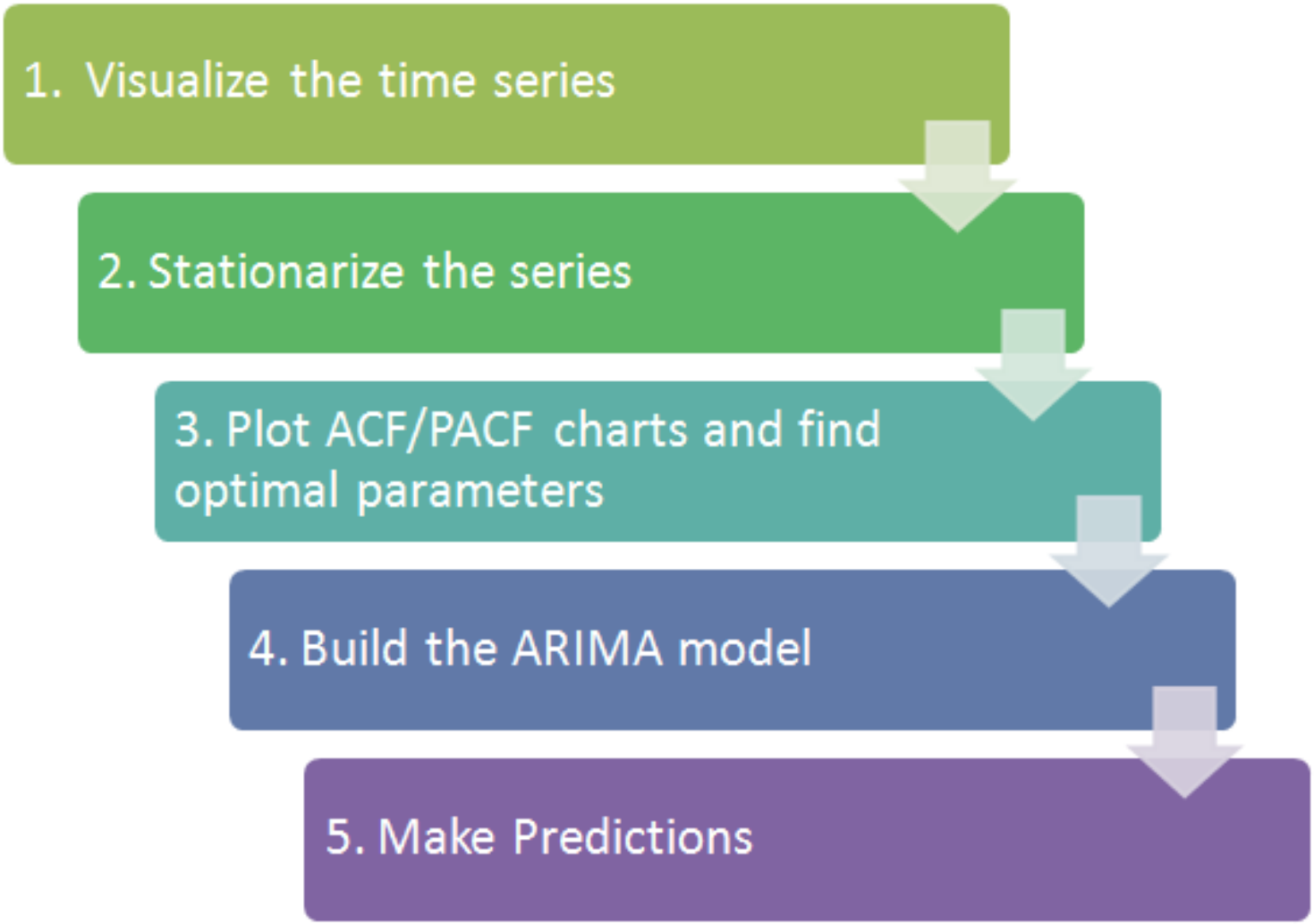
- Proces $Y_t = \frac{1}{2} Y_{t-1} + \varepsilon_t$ ma równanie charakterystyczne postaci $x - 2 = 0$, które ma pierwiastek równy 2. Jest to zatem proces stacjonarny.
- Proces $Y_t = Y_{t-1} - \frac{1}{4} Y_{t-2} + \varepsilon_t$ ma równanie charakterystyczne postaci $x^2 - 4x + 4 = 0$, które ma pierwiastek podwójny równy 2. Zatem jest to również proces stacjonarny.
- Proces $Y_t = \frac{1}{2} Y_{t-1} + \frac{1}{2} Y_{t-2} + \varepsilon_t$ ma równanie charakterystyczne postaci $x^2 + x - 2 = 0$, które ma pierwiastki równe -2 i 1. Ponieważ nie są oba większe co do wartości bezwzględnej od 1, zatem proces jest niestacjonarny.
- Proces $Y_t = -\frac{1}{4} Y_{t-2} + \varepsilon_t$ ma równanie charakterystyczne postaci $x^2 + 4 = 0$, które ma dwa pierwiastki zespolone postaci $\pm 2i$, dla których $|2i| = \sqrt{2^2 + 0^2} = 2$. Czyli proces jest stacjonarny.

Stacjonarność

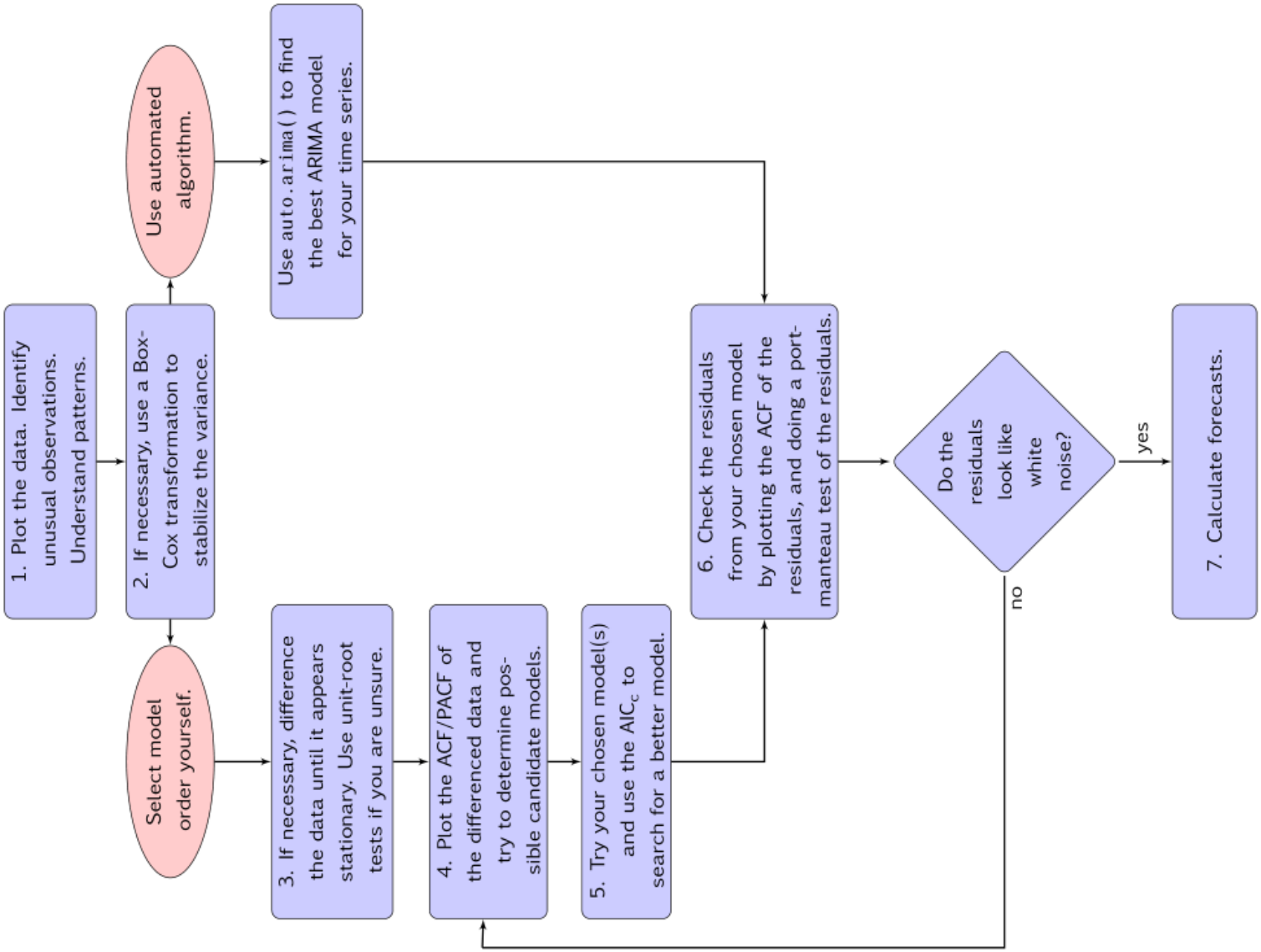
Bibliografia

- Dickey, D.A., Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association* 74(366):427–431.
- Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics* 54(1-3):159–178.

Flowchart


1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Flowchart


1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Use automated algorithm.
Use auto.arima() to find the best ARIMA model for your time series.
Select model order yourself.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AIC$_c$ to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
yes
no
7. Calculate forecasts.

Inne modele

Istnieje oczywiście znacznie więcej metod modelowania szeregów czasowych (oczywiście można spróbować też każdej metody regresyjnej), np.

- ETS – bardziej wyrafinowane wygładzanie wykładnicze.
- BATS – modele o złożonej sezonowości.
- THETA – wygładzanie wykładnicze z dryftem.
- Prophet – metoda zaproponowana przez naukowców z Facebooka, mająca dawać dobre prognozy bez „ręcznego" dopasowywania parametrów. Uwzględnia trend, sezonowość, święta oraz punkty zmian.

Inne modele

Bibliografia

Hyndman, R.J., Athanasopoulos, G. (2021). Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. `OTexts.com/fpp3.Accessed03.01.2024`

Taylor, S.J., Letham, B. (2017). *Forecasting at scale*. PeerJ Preprints, Tech. Rep. e3190v2

Korelacja krzyżowa

Korelacja krzyżowe

Korelacja krzyżowa lub wzajemna (*ang. cross correlation*) – funkcja wartości współczynnika korelacji Pearsona dwóch szeregów czasowych. Oczywiście możemy również przesuwać szeregi względem siebie.

Przyczynowość w sensie Grangera

Przyczynowość w sensie GRANGERA (*ang. Granger causality*) – X powoduje Y wtedy i tylko wtedy, gdy włączenie do modelu przewidującego zmienną objaśnianą Y wartości zmiennej objaśniającej X zwiększa trafność predykcji.

Test Grangera

Test Grangera

H_0 : Szereg czasowy X_t nie wpływa na prognozy szeregu czasowego Y_t.
H_1 : Szereg czasowy X_t wpływa na prognozy szeregu czasowego Y_t.

Bibliografia

Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 37(3):424–438.

sir Clive William John Granger (1934-2009)


I'VE FINALLY FOUND IT... AFTER 15 YEARS
THE SCROLL OF TRUTH!
Robotatertotcomics
It's impossible to predict stock prices
NYEHHH
imgflip.com