

#Webinar

3 CZERWCA 2020 | GODZ. 18:00

**SZCZĘŚLIWI CZASU NIE
LICZĄ. MASZYNY LICZĄ,
CZYLI UCZENIE MASZYNOWE
DLA SZEREGÓW
CZASOWYCH**

facebook | twitter | linkedin:
@ sagespl

Kim jestem?

ROSTYSLAV APOSTOL

Data Scientist w NorthGravity

Certyfikowany specjalista ML AWS

Trener w Sages



sages

Agenda

1. Dlaczego szeregi czasowe są ważne?
2. Czym się różni uczenie maszynowe dla szeregów czasowych od regresji?
3. Czy jesteśmy pewni, że nasze dane to szeregi czasowe?
4. Co robić, jak nasze dane to nie szeregi czasowe, a muszą być?
5. Feature Engineering - jak wzbogacić nasze dane?
6. Naiwna prognoza - czy nasz model uczenia maszynowego w ogóle ma sens?
7. Dlaczego walidacja krzyżowa nie działa?
8. Czy tradycyjne metody statystyczne już do niczego?



Po co nam szeregi czasowe?

- Prognozy gospodarcze
- Prognozy sprzedażowe
- Analizy rynku akcji / surowców / nieruchomości
- Prognozy plonów
- Prognozowanie zużycia gazu
- Monitorowanie działanie sprzętu technologicznego

Podstawa kluczowych i strategicznych decyzji

PO CO NAM ML DLA SZEREGÓW CZASOWYCH?

ZALETY

- Wyższa dokładność prognoz
- Nieliniowe zależności
- Duża ilość danych,
- wiele wymiarów
- Wydajność obliczeń

WADY

- ML nie zawsze ma sens
- Problem z tłumaczeniem modeli (tzw. modele - czarne skrzynki)

SZEREGI CZASOWE VS. REGRESJA

NAPOLEON

**Czas jest
wszystkim.**

JOSEPH CONRAD

**Z naprawdę wielkich,
posiadamy
tylko jednego wroga -
czas.**

LEW TOŁSTOJ

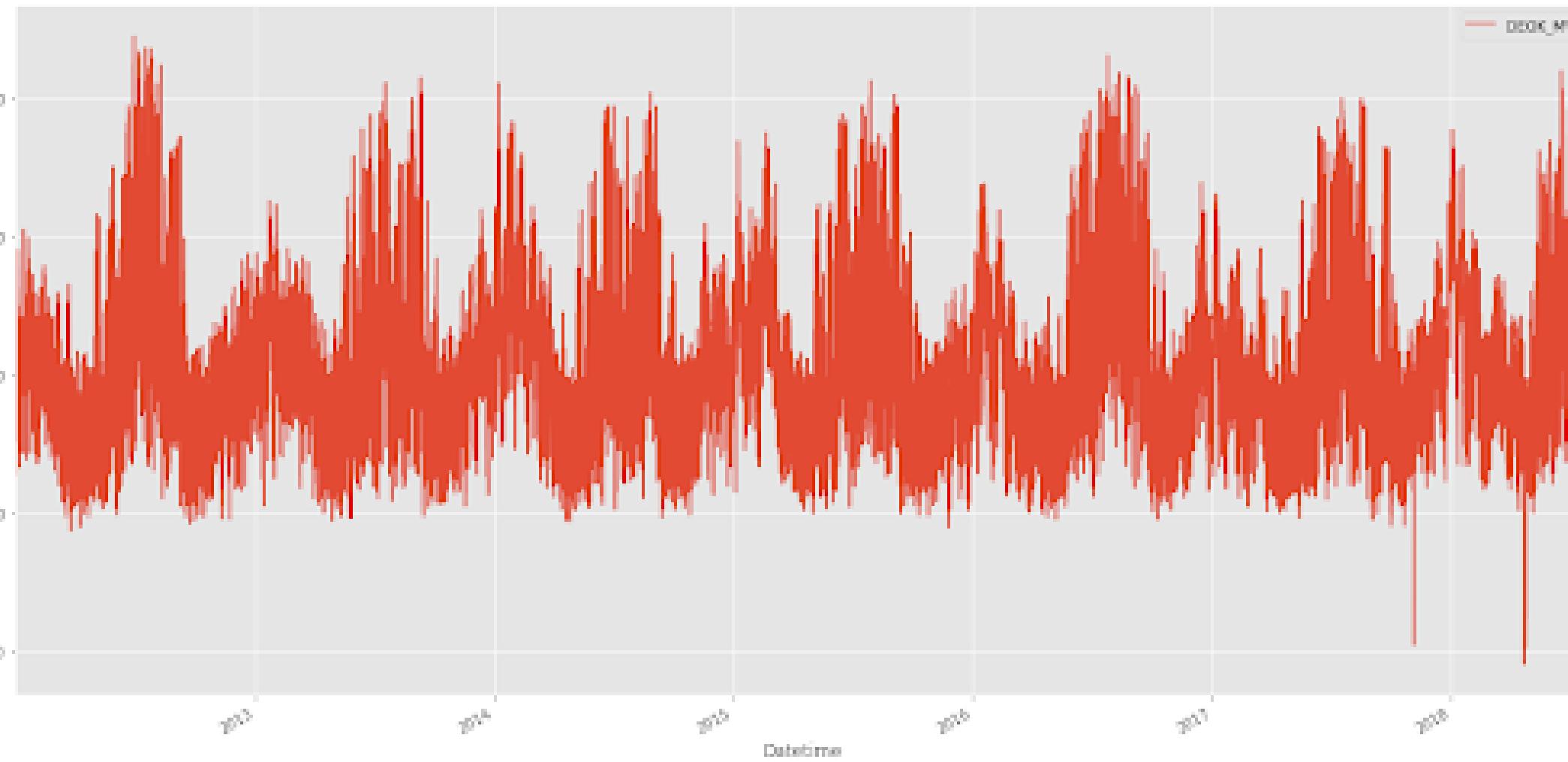
**Czas się nie spieszy
- to my nie nadążamy.**

ALBERT EINSTEIN

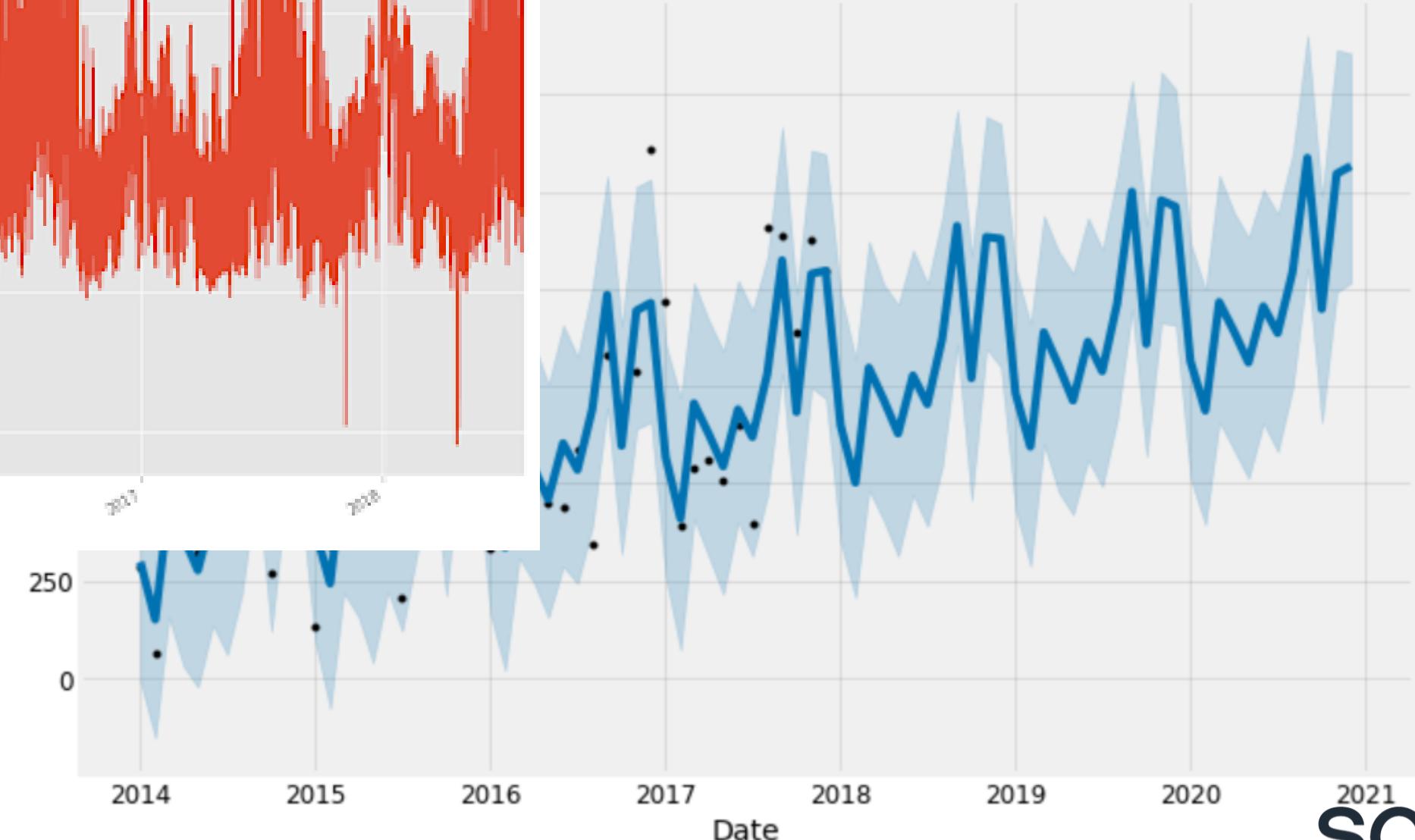
**Znane są tysiące
sposobów zabijania
czasu, ale nikt nie wie,
jak go wskrzesić.**

SZEREGI CZASOWE

KOLEJNOŚĆ JEST KLUCZOWA



Office Supplies Sales



Czy nasze dane to szeregi czasowe?



DANE ZAWIERAJĄ ZMIENNĄ
ODZWIERCIEDLAJĄCA CZAS LUB
KOLEJNOŚĆ - INDEKS CZASOWY.



DANE SĄ
POSORTOWANE.



INDEKS CZASOWY
MA STABILNĄ
CZĘSTOTLIWOŚĆ.

Tabela 1

Data	Wartość
2020-01-01	100
2020-01-02	200
2020-01-05	120
2020-01-09	170

Tabela 2

Data	Wartość
2020-01-01	120
2020-01-10	140
2020-01-05	160
2020-01-03	150

Tabela 3

Data	Wartość
2020-01-01	100
2020-01-02	130
2020-01-03	150
2020-01-04	140

Jak zmusić dane do bycia szeregiem czasowym?



Wprowadzić sztuczną kolumnę z kolejnością -
1,2,3,4,....



Posortować indeks
czasowy.



Ponowne próbkowanie - resampling z potrzebną częstotliwością.

Z pustego i Salomon nie naleje - wartości brakujące

Data	Wartość	Fill Forward	Backfilling	Interpolacja	Srednia
2020-01-01	100	100	100	100	100
2020-01-02	200	200	200	200	200
2020-01-03		200	120	160	148
2020-01-04	120	120	120	120	120
2020-01-05		120	170	137	148
2020-01-06		120	170	154	148
2020-01-07	170	170	170	170	170

FILL FORWARD

wartość brakująca taka jak poprzednia

BACKFILLING

wartość brakująca tak jak następująca

INTERPOLACJA

wartości pośrodku pomiędzy istniejącymi wartościami

WARTOŚĆ ŚREDNIA / MEDIANA

“

Feature Engineering

- STOSOWANE UCZENIA
MASZYNOWE TO W
ZASADZIE INŻYNIERIA
ZMIENNYCH

- ANDREW NG, PROFESOR

Wzbogacamy nasze dane

1. **Opóźnione wartości** - co było wczoraj, tydzień temu, miesiąc temu?

2. **Kroczące statystyki** - średnia, mediana, odchylenie standardowe, suma itd. o różnych wartościach

• “rolling window”

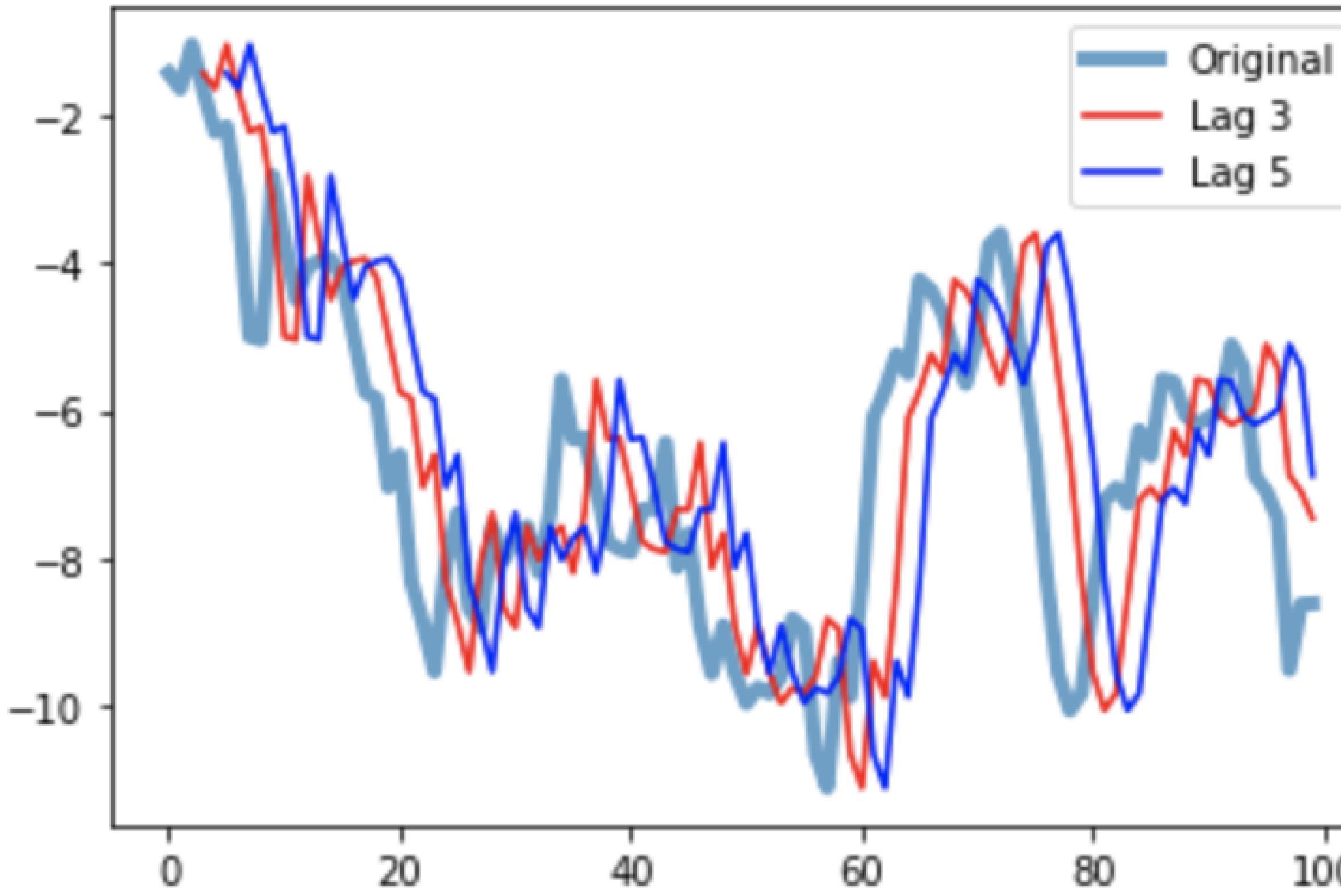
3. **Trend i sezonowość**

4. **Zmienne związane z datą:**

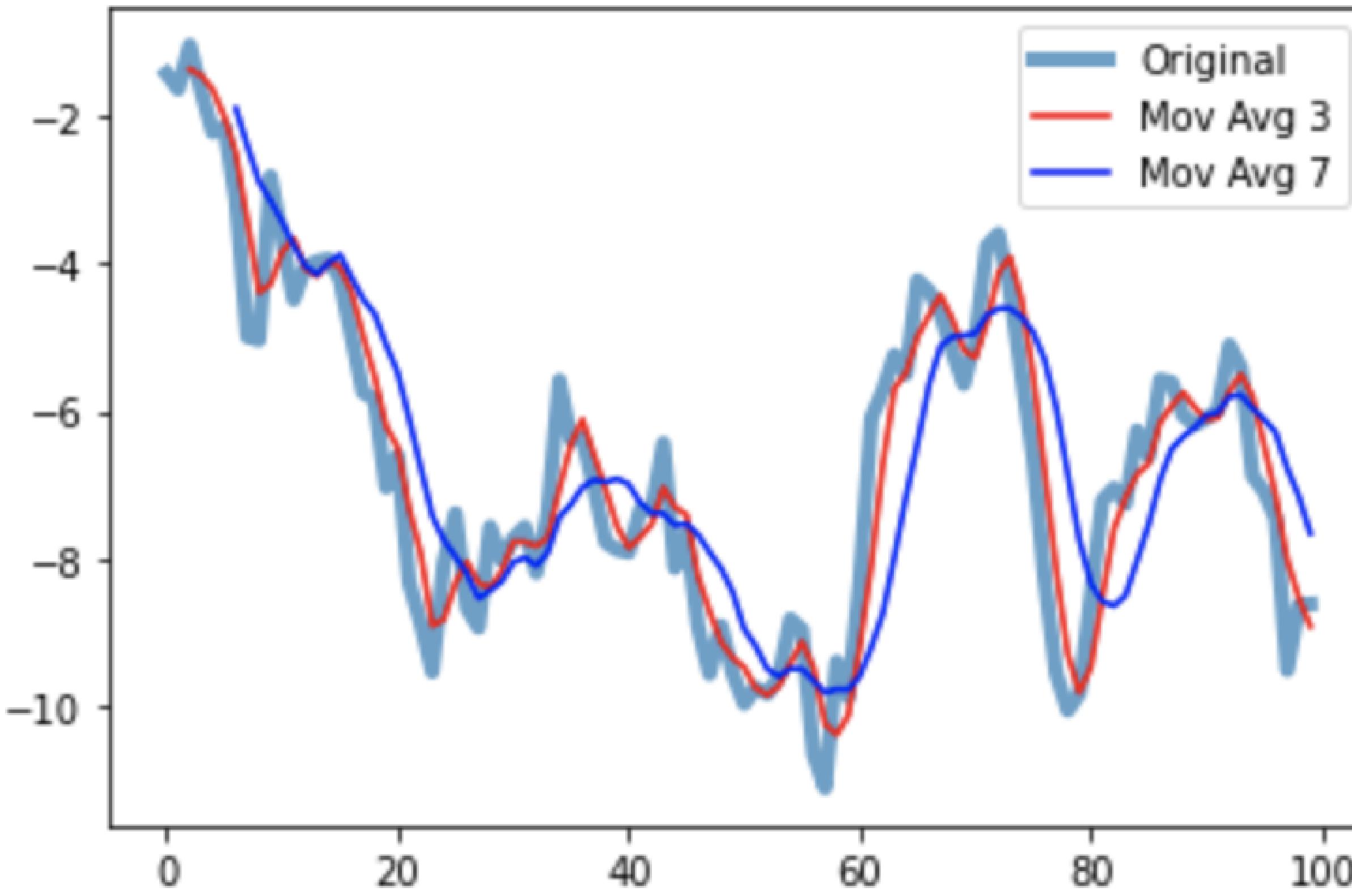
- Miesiąc
- Dzień tygodnia
- Czy to weekend?
- Czy to dzień wolny od pracy?
- Czy to początek miesiąca / kwartału

5. **Różnicowanie**

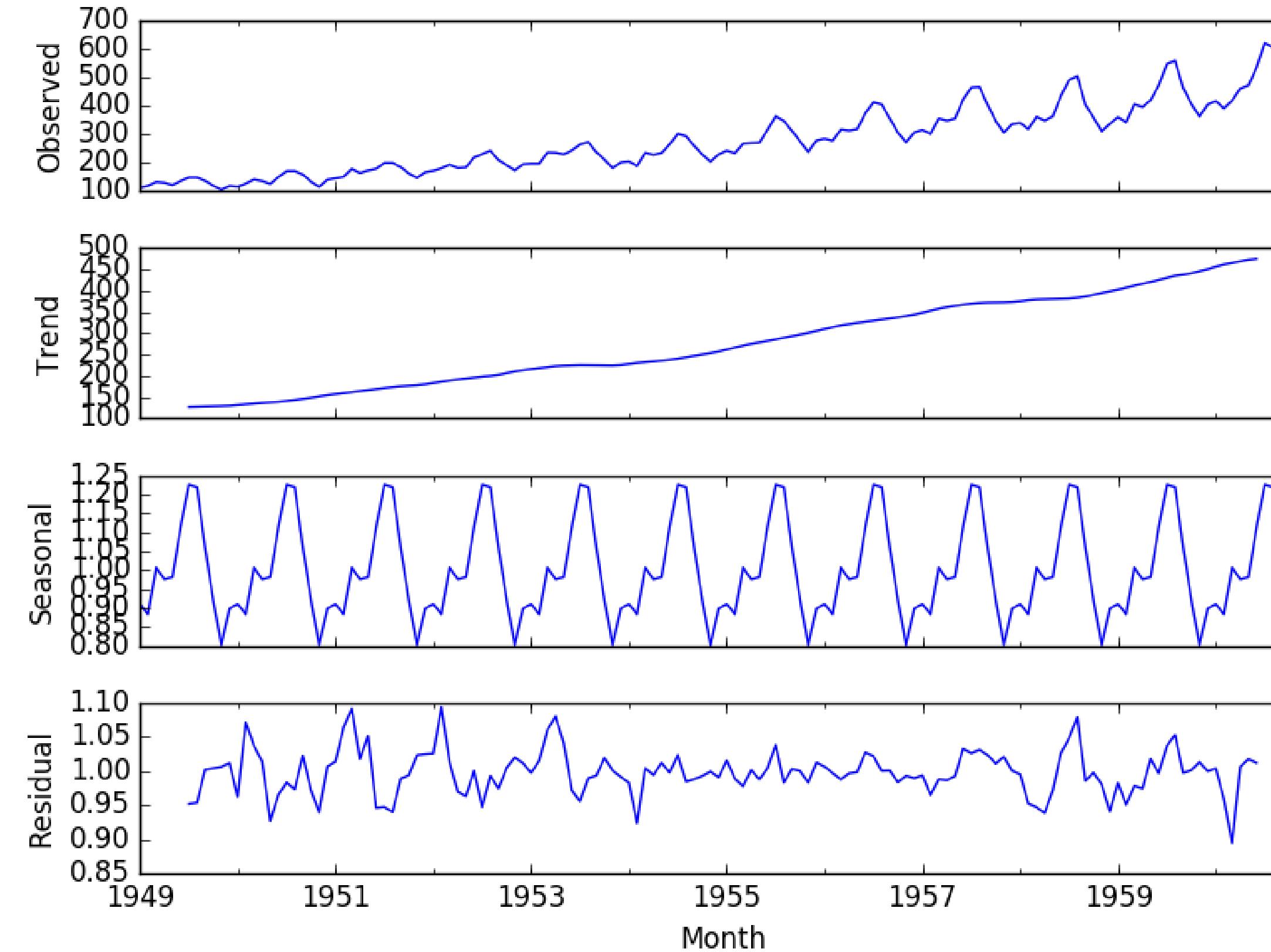
Wartości opóźnione



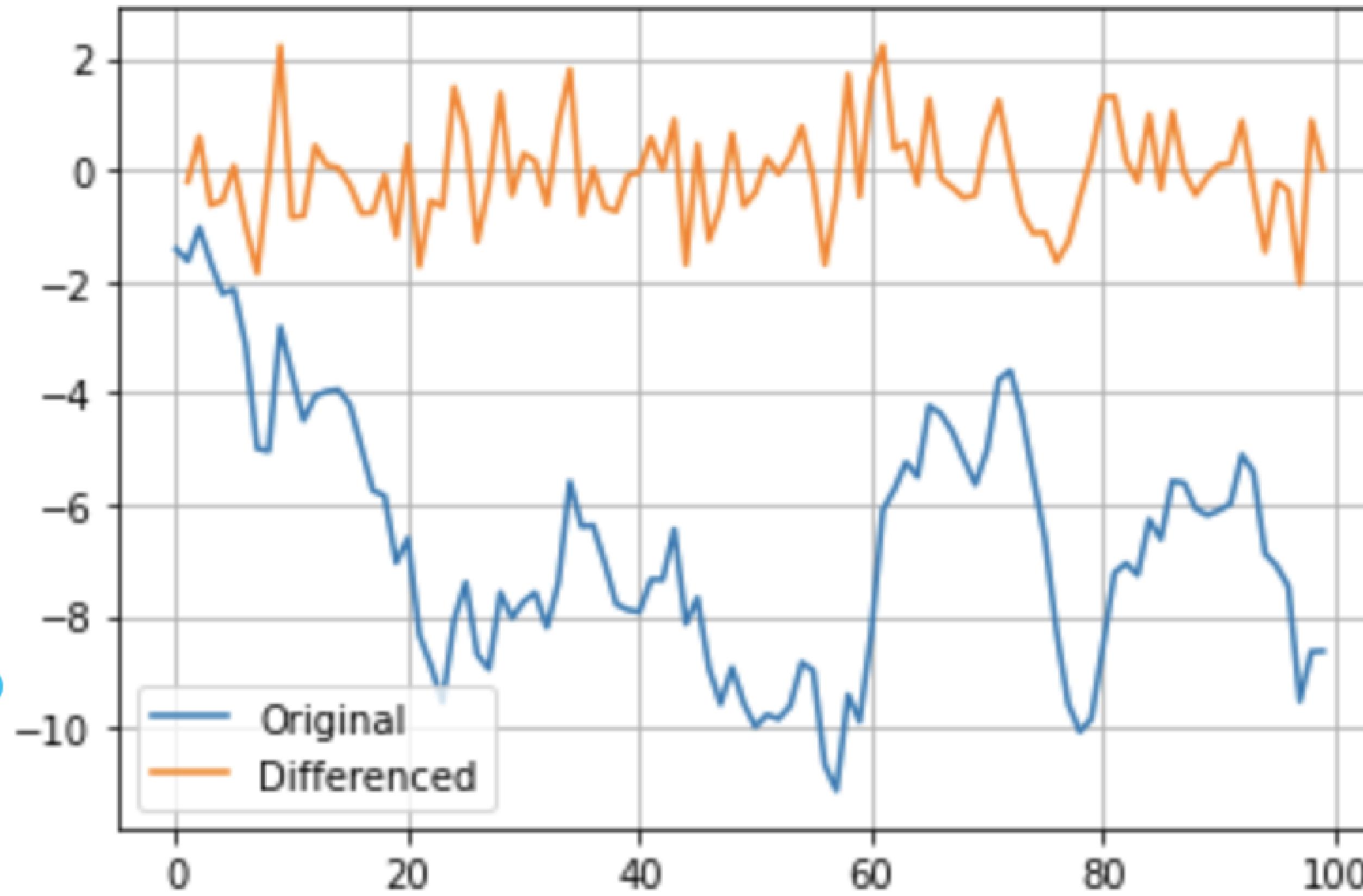
Kroczące statystyki



Rozkład na trend i sezonowość



Różnicowanie szeregu czasowego

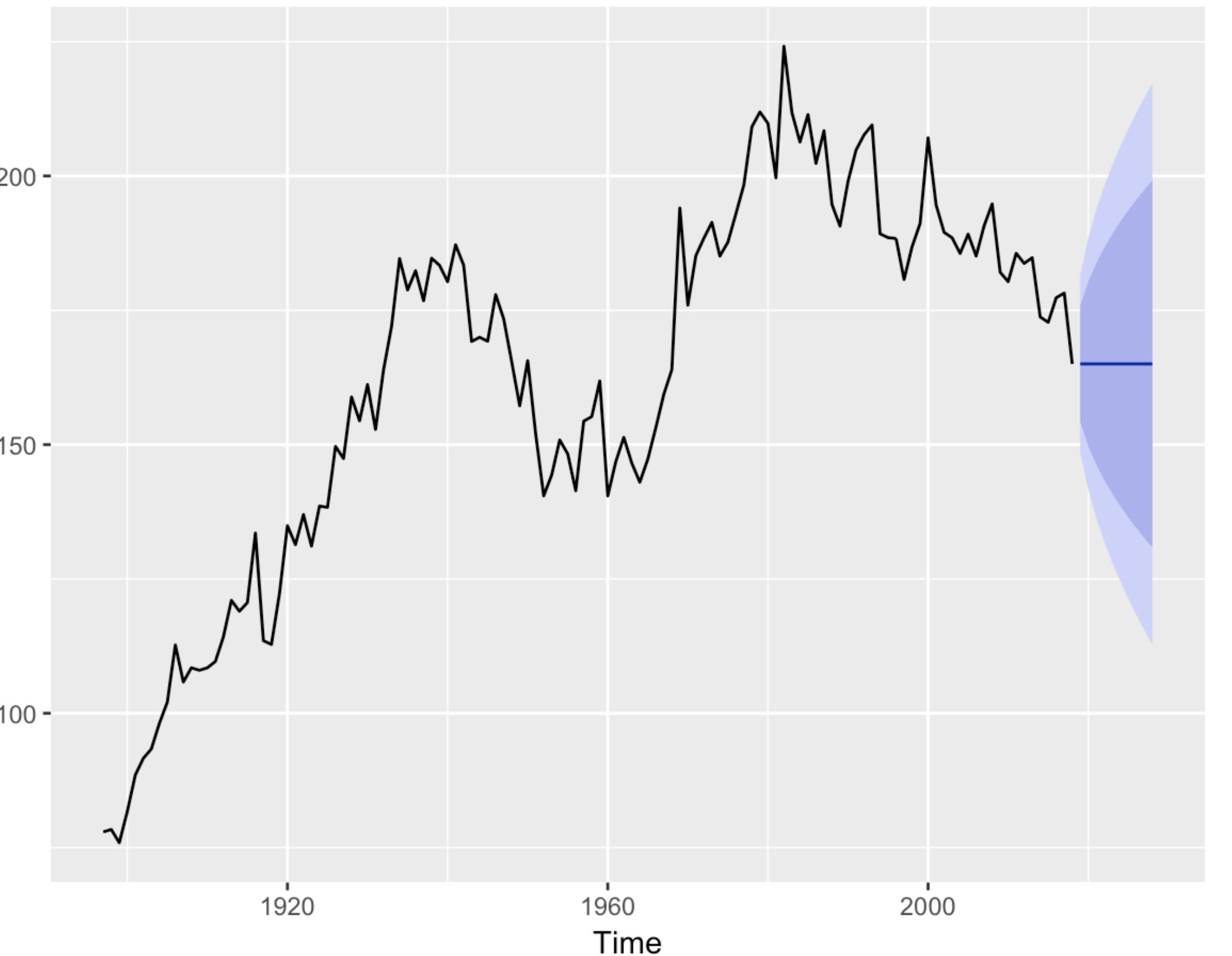


Feature Engineering - Przykład

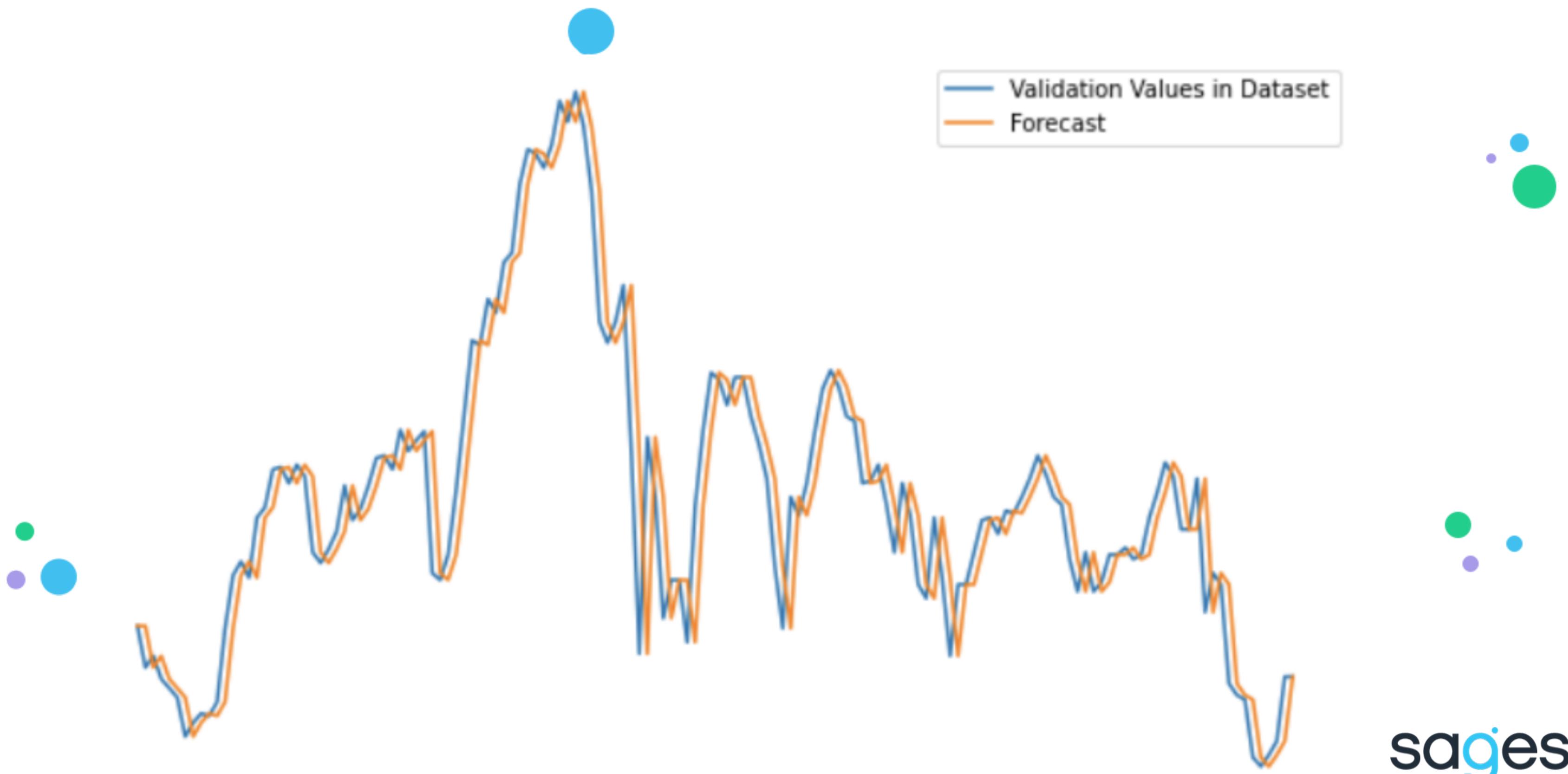
Data	Wartość	Lag_1	Lag_3	Srednia_3	Dzień tyg.	Weekend
2020-01-01	100	-	-	-	3	0
2020-01-02	200	100	-	-	4	0
2020-01-03	200	200	-	166.7	5	0
2020-01-04	120	200	100	173.3	6	1
2020-01-05	120	120	200	146.7	0	1
2020-01-06	120	120	200	120	1	0
2020-01-07	170	120	120	136.7	2	0

Naiwna Prognoza

Jutro będzie tak jak dzisiaj.
I pojutrze też ...

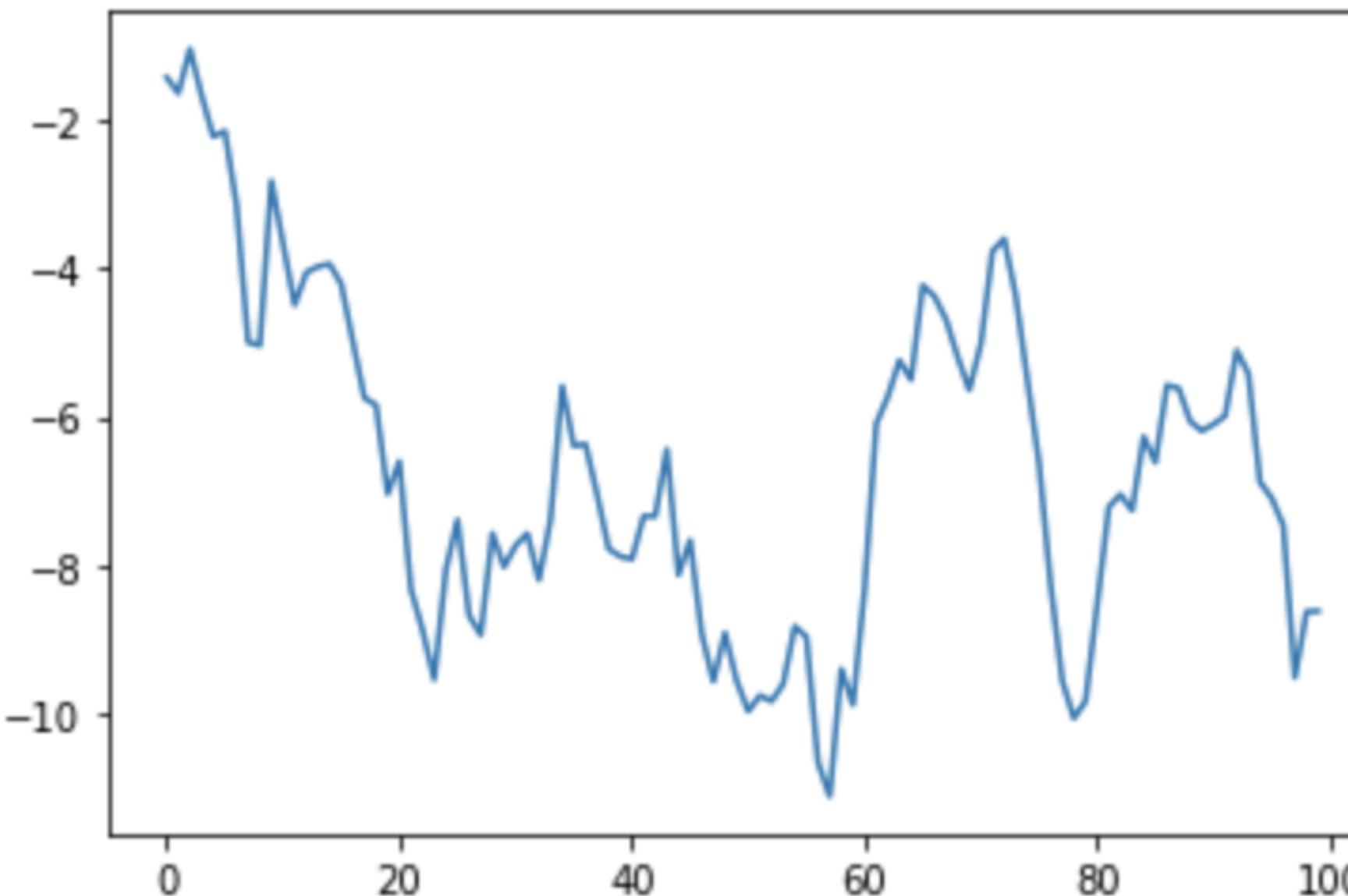


Naiwna prognoza - walidacja



Naiwna prognoza najlepszym rozwiązaniem

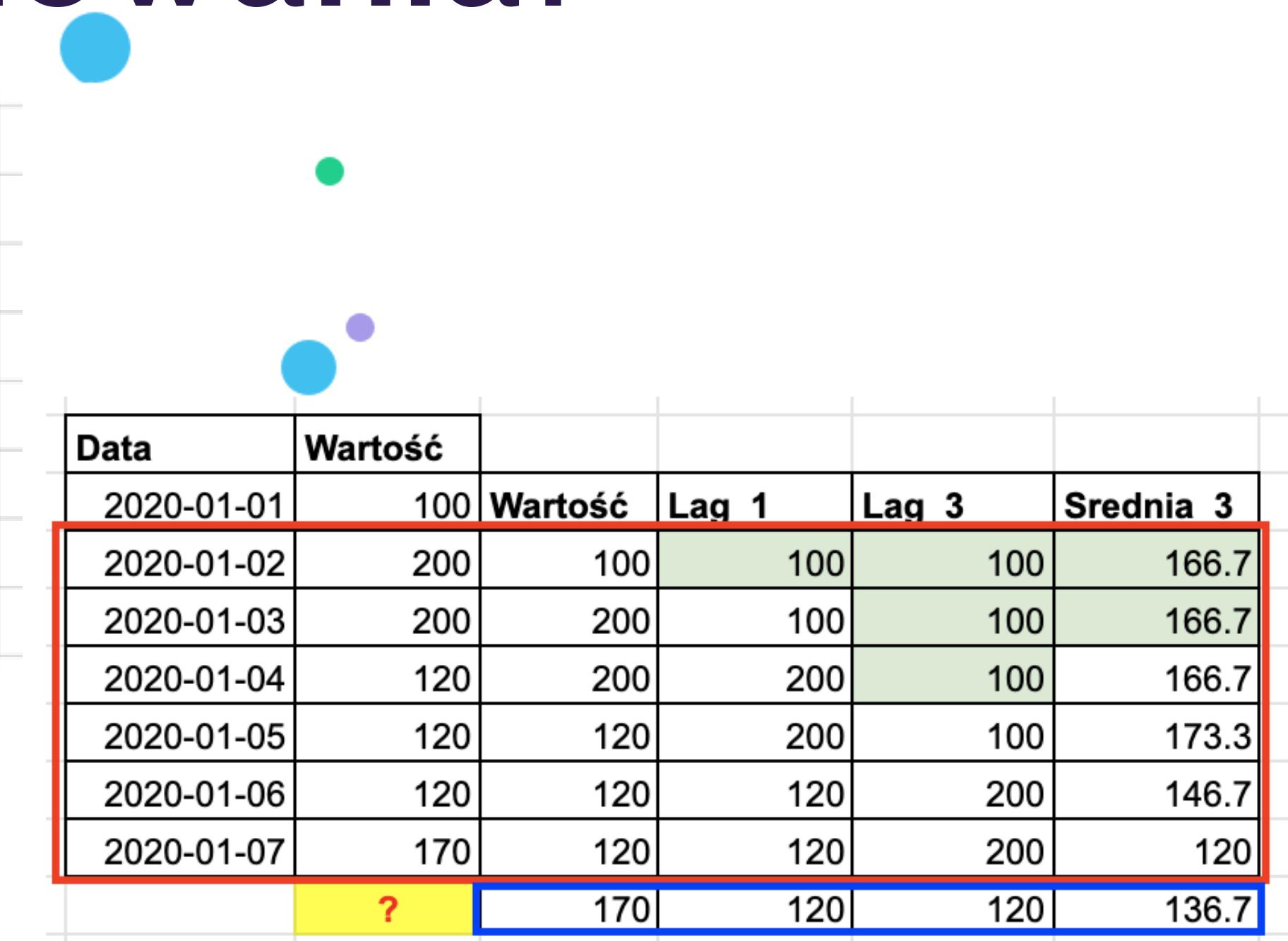
Naiwna prognoza będzie najlepszym rozwiązaniem, jeżeli szereg czasowy jest błędzeniem losowym (random walk).



Dickey-Fuller test - pozwala zweryfikować, czy proces jest błędzeniem losowym.

Jak przygotować dane do trenowania?

Data	Wartość	Lag_1	Lag_3	Srednia_3
2020-01-01	100	100	100	166.7
2020-01-02	200	100	100	166.7
2020-01-03	200	200	100	166.7
2020-01-04	120	200	100	173.3
2020-01-05	120	120	200	146.7
2020-01-06	120	120	200	120
2020-01-07	170	120	120	136.7



Data	Wartość	Wartość	Lag_1	Lag_3	Srednia_3	
2020-01-01	100	100	100	100	166.7	
2020-01-02	200	200	100	100	166.7	
2020-01-03	200	200	200	100	166.7	
2020-01-04	120	200	200	100	166.7	
2020-01-05	120	120	200	200	100	173.3
2020-01-06	120	120	120	200	200	146.7
2020-01-07	170	120	120	120	200	120
	?	170	120	120	120	136.7

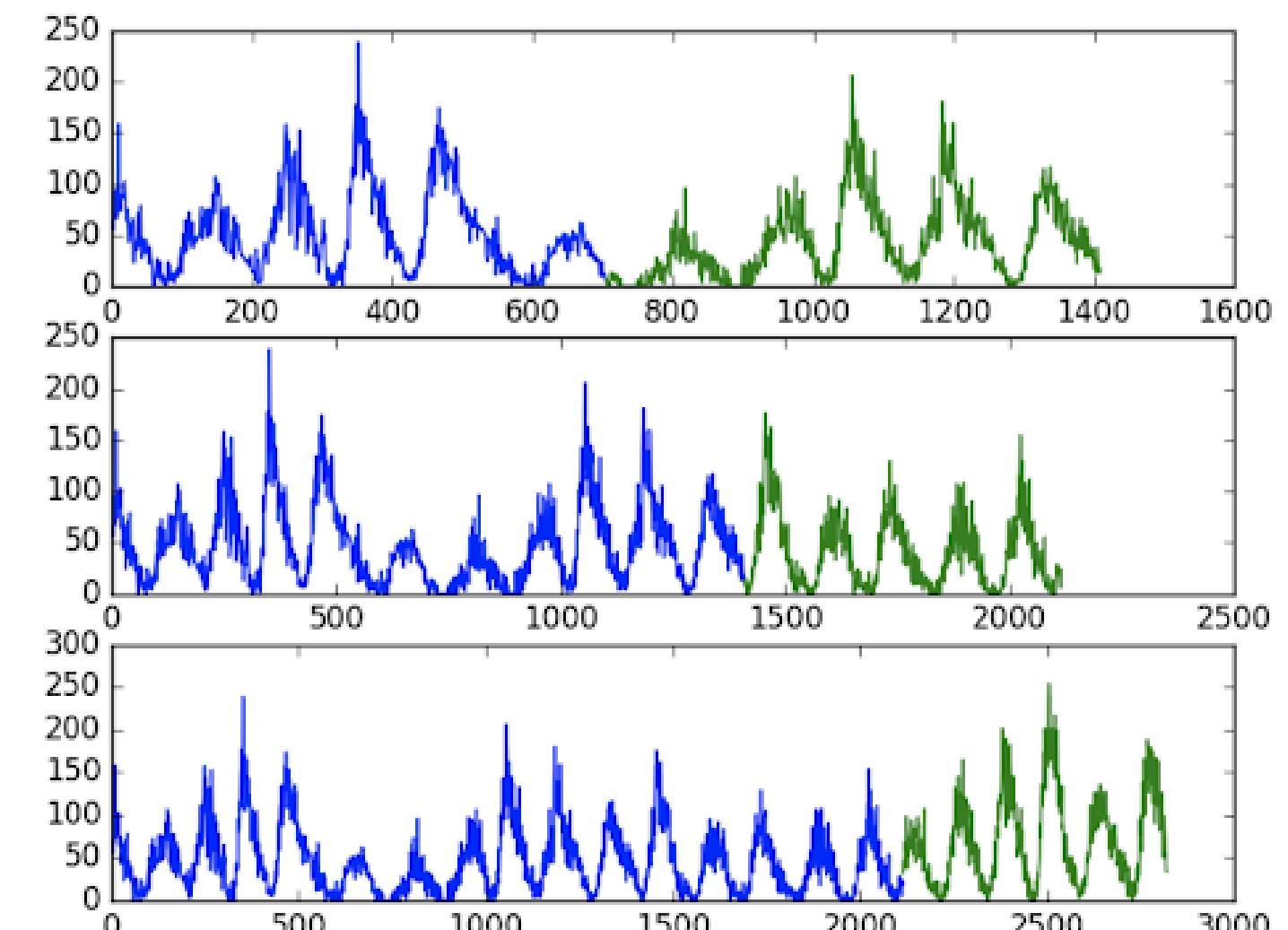
Dlaczego walidacja krzyżowa nie działa?



Metody walidacji modeli dla szeregów czasowych

1. Train-test split
2. Multiple train-test split.
3. Walkforwad validation

Od czego zależy?



Czy tradycyjne metody są już do niczego?

Spyros Markidakis - profesor
Uniwersytetu w Nikozji, Cypr, 2018

Algorytmy: 8 tradycyjnych + 10 ML

Zestawy danych: 1045

- jednowymiarowych szeregów czasowych, częstotliwość - od godzinowych do rocznych.



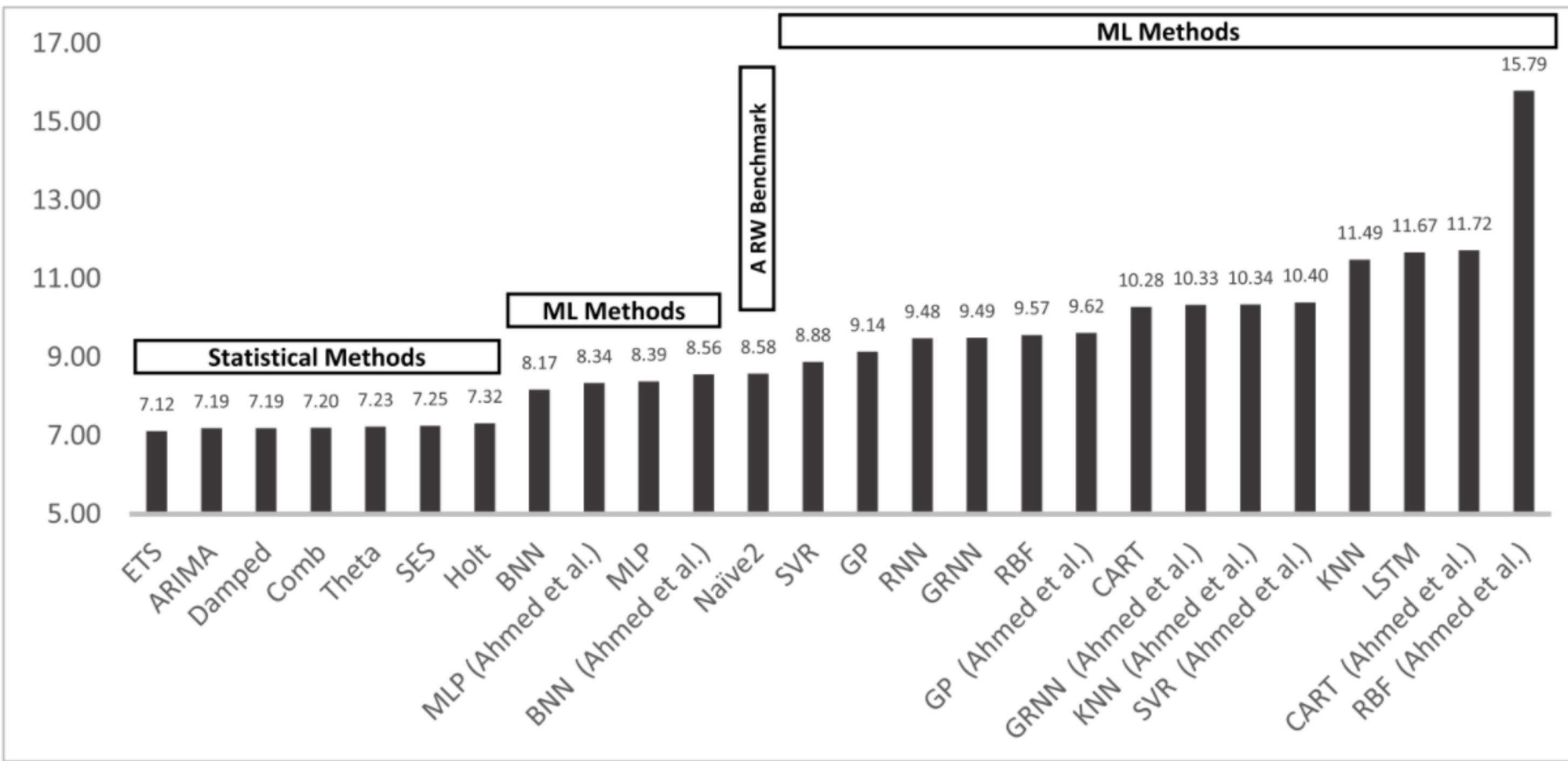
Przygotowanie danych: różne kombinacje

Walidacja: walk forward, 18 ostatnich obserwacji



Wyniki badań

- błędy prognozowania



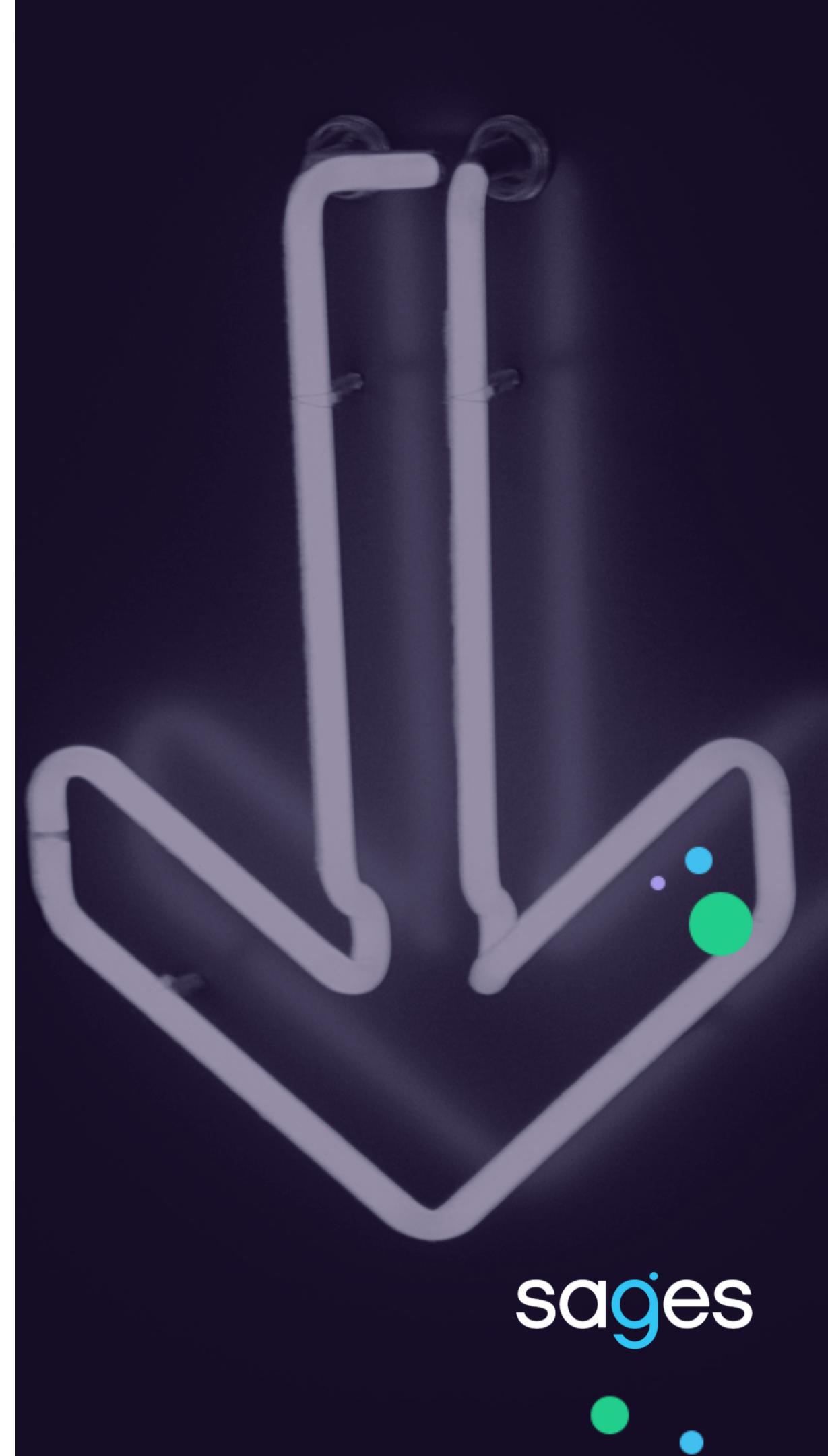
Co nie było uwzględnione w badaniu?

- 1. Złożone nieregularne struktury
- czasowe
- 2. Wartości brakujące
- 3. Mocno zaszumionie dane

Może właśnie tam ML działa lepiej?

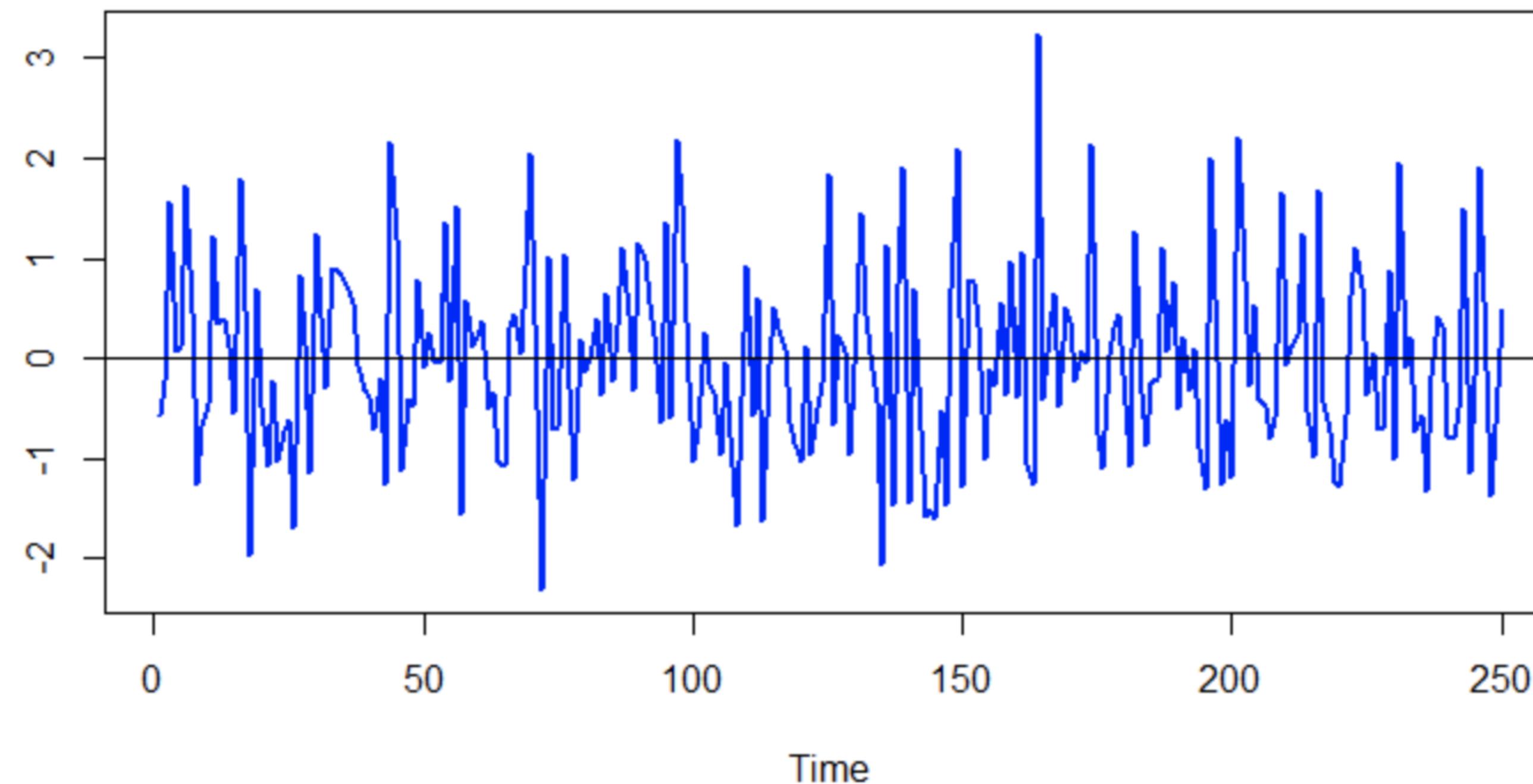
Algorytmy do spróbowania

- 1. Prognoza Naiwna (Naive, Naive 2)
- 2. Modele auto regresyjne (AR, ARIMA, SARIMA)
- 3. Wygładzanie wykładnicze (Holta, Wintersa)
- 4. ML - modele liniowe (Linear, Ridge, Lasso, ElasticNet)
- 5. ML - modele nieliniowe (KNN, SVR, drzewa decyzyjne)
- 6. ML - Ensemble Learning (lasy losowe, XGBoost)
- 7. Głębokie uczenie (MLP, CNN, LSTM, hybrydy)

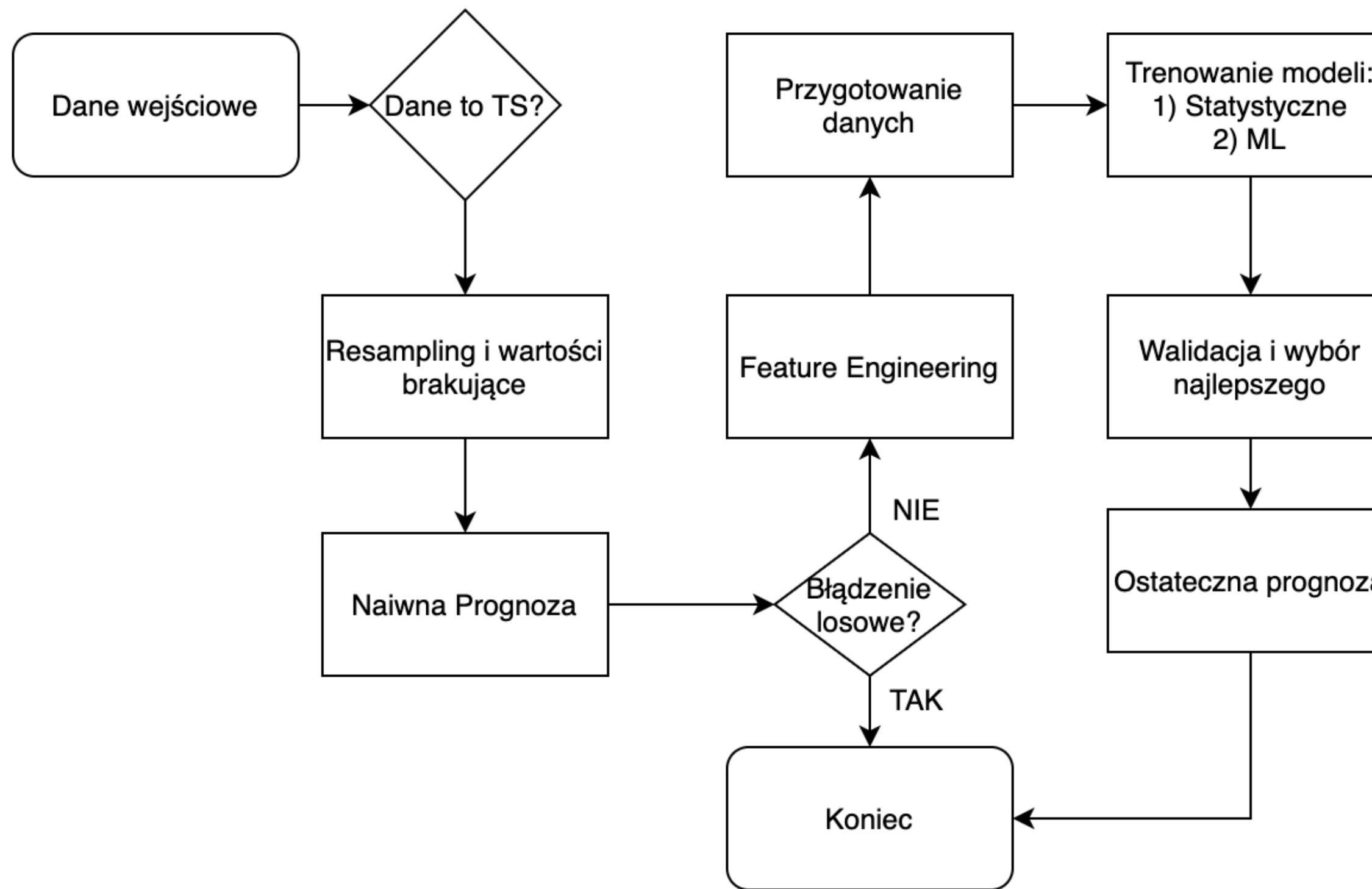


Kiedy lepiej się już nie da?

Błędy prognozowania są białym szumem.



Podsumowanie - Proces ML dla szeregów czasowych





Jakieś pytania?

sages

20% ZNIŻKI

OTWARTE SZKOLENIA SAGES Z
AUTORSKIEJ OFERTY:
[HTTPS://WWW.SAGES.PL/](https://www.sages.pl/)

PROMOCJA JEST WAŻNA DO 31 LIPCA
2020

KOD: WEBINAR201

sages