Komputerowa analiza szeregów czasowych

Raport: **2**

Temat sprawozdania **Wykorzystanie poznanych metod służących do analizy zależności liniowej dla wybranych danych rzeczywistych.**
Nazwisko i Imię prowadzącego kurs **Inż. Justyna Witulska**

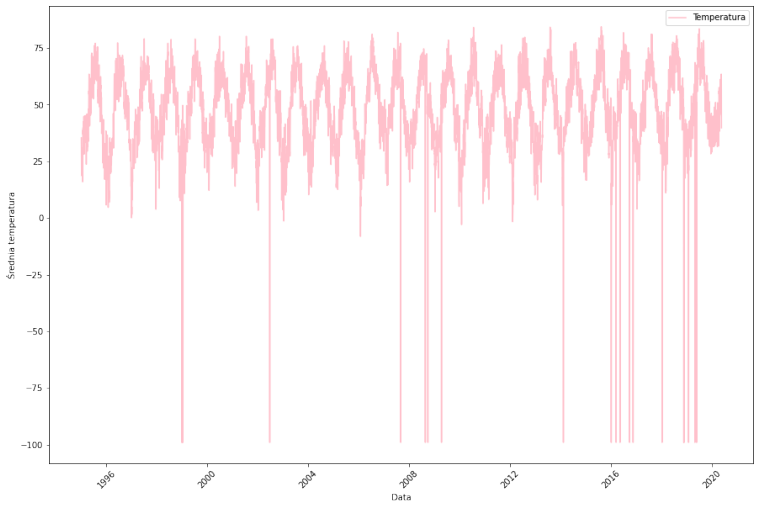
Wykonawca:	
Imię i Nazwisko, nr indeksu	Adrianna Ziobroniewicz, 262227 Anna Zadka, 262226
Wydział	Wydział matematyki, W13
Termin zajęć:	Środa, 7^{30}
Numer grupy ćwiczeniowej	T00-79c
Data oddania sprawozdania:	8 lutego 2023
Ocena końcowa	

Spis treści

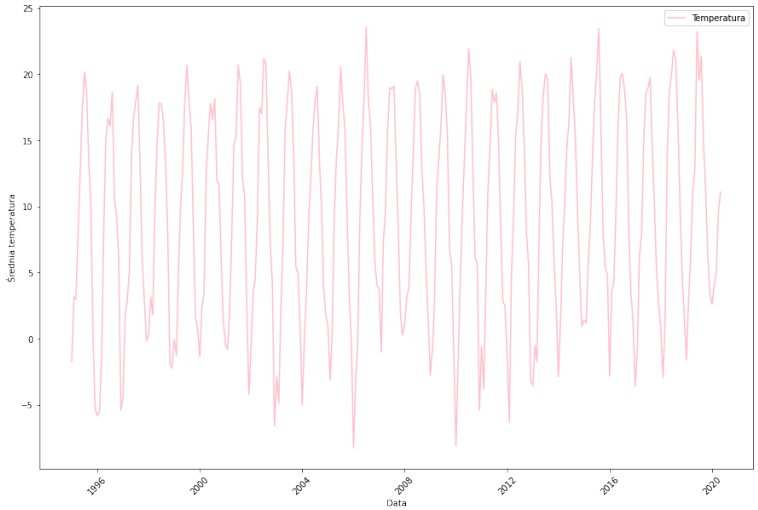
1. Wstęp

Wizualizacja danych oraz ich opis

Do analizy wzięliśmy dane ze strony:
$https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities.$
Analizowane przez będą temperatury w latach 1992-2020. Bierzemy pod uwagę 9266 obserwacji. Początkowo nasze dane są w Fahrenheitach, później zajmujemy się zamianą na stopnie Celsjusza.



Rysunek 1. Wizualizacja danych z wartościami odstającymi - wyrażone w Fahrenheitach



Rysunek 2. Wizualizacja po oczyszczeniu danych wyrażone w Celsjuszach

2. Przygotowanie danych do analizy

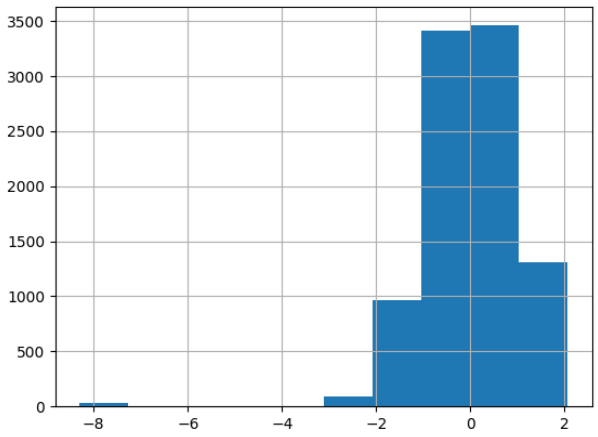
Badanie jakości danych

Jak wiadomo miasto, które wybraliśmy do analizy to Warszawa. Nasze dane dotyczące dat były rozdzielone, więc zajęliśmy się skonstruowaniem pełnej daty.Do analizy bierzemy średnią temperaturę na dany dzień. Temperatury są podane w stopniach Farenheita, aby były dla nas przyjazne to zamienimy je na stopnie Celsjusza i zaokrąglimy do jednego miejsca po przecinku.
Sprawdzamy czy w zbiorze znajdują się jakieś wartości nieznane (NaN), jednakże takich nie ma.
Kolejna weryfikacja to powtarzalność indeksów. Taką sytuację zauważamy w grudniu 2015 roku.

```
In [11]:  1 df['2015-12-30']

Out[11]: data
         2015-12-30    -4.1
         2015-12-30   -72.8
         Name: AvgTemperature, dtype: float64
```

Rysunek 3. 2 indeksy powtarzające się zanotowane 30 grudnia 2015 roku

Temperatura -72.8 stopnia jest bez dwóch zdań błędem pomiarowym, więc zajmiemy się jego usunięciem. Pozostała nam tylko jedna temperatura, ta poprawna. W ramach dalszej weryfikacji poprawności zbioru zbadamy obecność wartości odstających (outlierów) za pomocą metryki z-score. Określa ona jak daleko nasz wartości znajdują się od standardowych wartości w tym zbiorze. Im większy z-score tym bardziej można podejrzewać, że dana próbka może być outlierem.



Rysunek 4. Histogram dla wartości z-score

Większość próbek zawiera się w przedziałach od -3 do 2. Jest to standardowy zakres, który nie powinien budzić naszych obaw. Sprawdźmy, co to są za próbki, które mieszczą się w wartościach poniżej -4.

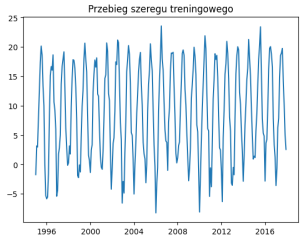
```
In [16]:  1 df[df_zscore < -4]

Out[16]: data
         1998-12-24   -72.8
         1998-12-25   -72.8
         1998-12-30   -72.8
         1998-12-31   -72.8
         1999-01-10   -72.8
         2002-06-18   -72.8
         2002-06-19   -72.8
         2002-06-20   -72.8
         2002-06-21   -72.8
         2007-08-28   -72.8
         2008-08-18   -72.8
         2008-09-24   -72.8
         2009-04-09   -72.8
         2014-02-06   -72.8
         2015-12-31   -72.8
         2016-03-10   -72.8
         2016-05-05   -72.8
         2016-09-21   -72.8
         2016-11-11   -72.8
         2018-01-04   -72.8
         2018-11-14   -72.8
```

Rysunek 5. Część próbek, które mieszczą się w wartościach poniżej -4.

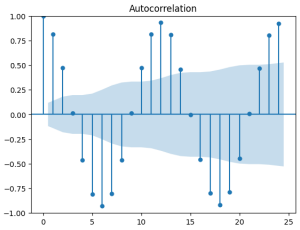
Widzimy wyraźnie, że jest to ta sama wartość, którą usuwaliśmy już wyżej. Jest to zatem błąd w pomiarach. Tym samym każdą z tych wartości usuniemy, a puste pola uzupełnimy wartości 'ffill" czyli wartością poprzednią. Tym samym jeśli informacja z np. 15 maja zostanie usunięta to za pomocą frontfill zostanie ona uzupełniona wartością z 14 maja. Jest to bezpieczny sposób uzupełniania informacja, gdyż można zakładać, że temperatura w sąsiadujących dniach jest zbliżona do siebie i dokonując takiego uzupełnienia nie przekłamujemy znacząco rzeczywistej informacji.
Nasz zbiór został sprawdzony pod wzlędem braków i outlierów. Możemy przejść do dalszej analizy. Na początek znacząco zmniejszymy nasz zbiór robiąc resampling miesięczny, czyli dane dzienne zamienimy na miesięczne w ten sposób, że weźmiemy średnią temperaturę ze stycznia, następnie średnią temperaturę z lutego itd.

Następnie dokonamy podziału zbioru na dwa: treningowy oraz testowy. Do treningowego użyte zostaną lata od 1995 do 2017, a na zbiór testowy przeznaczymy ostatnie dwa i pół roku

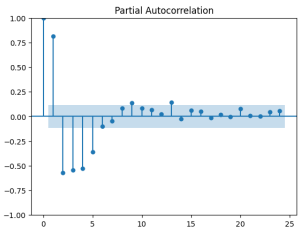
Wykresy ACF oraz PACF dla surowego szeregu



Rysunek 6. Przebieg szeregu treningowego



Rysunek 7. Wykres ACF dla surowych danych



Rysunek 8. Wykres PACF dla surowych danych

Wzór ACF, inaczej funkcja autokoleracji:

$$corr(X_t, X_s) = \frac{\gamma_x(t,s)}{\sqrt{\gamma_x(t,t)\gamma_x(s,s)}} = \rho_x(X_t, X_s).$$

Przy przesunięciu k jest to korelacja między wartościami szeregu oddalonymi o k przedziałów od siebie.

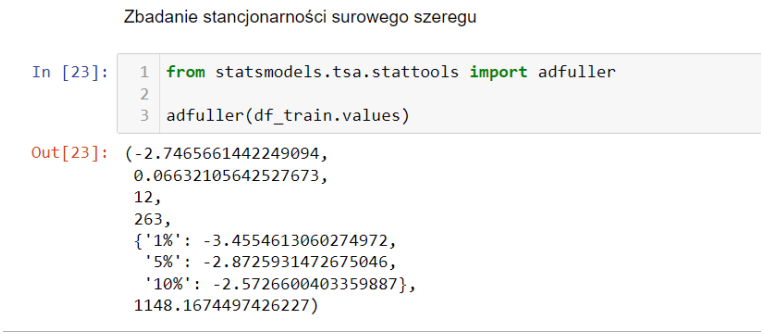
Wzór PACF, inaczej cząstkowa funkcja autokoleracji:

$$X_t = \phi_0 + \phi_1 X_t - 1 + ... + \phi_k X_t - k + Z_t,$$

gdzie $Z_t \sim WN(0, \sigma^2)$.
Przy przesunięciu k jest to korelacja między wartościami szeregu oddalonymi o k przedziałów od siebie, z jednoczesną rejestracją wartości z przedziałów znajdujących się pomiędzy.

Test ADF weryfikujący hipotezę o niestacjonarności dla surowych danych - Augmented Dickey-Fuller Test

Test ADF jest zasadniczo testem istotności statystycznej. Oznacza to, że istnieje testowanie hipotez, które jest związane z hipotezą zerową i alternatywną, w wyniku czego obliczana jest statystyka testowa i podawane są wartości p - z tej statystyki można wywnioskować czy dany szereg jest stacjonarny.
ADF jest to rozszerzona wersja Dickeya-Fullera. Rozszerza równanie testy o proces regresji wysokiego rzędu w modelu. Hipoteza zerowa jest jednak taka sama jak przy teście Dickeya-Fullera. Czyli $H_0 = \alpha = 1$. Hipoteza ta zakłada obecność pierwiastka jednostkowego($alpha = 1$), otrzymana wartość p powinna być mniejsza niż poziom istotności, aby móc odrzucić hipotezę zerową - tym samym wnioskują, że szereg jest stacjonarny.



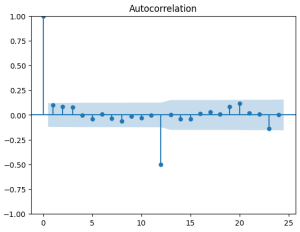
Rysunek 9. Wyniki - zbadanie stacjonarności surowego szeregu

Wartość p == 0.07 nie jest mniejsza niż 0.05 więc nie możemy odrzucić hipotezy zerowej. Tym samym stwierdzamy, że szereg jest niestacjonarny.

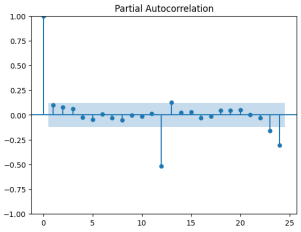
Wykresy ACF oraz PACF dla uzyskanego szeregu

Po zastosowaniu zróżnicowania udało nam się uzyskać szereg stacjonarny. Poniżej sprawdzamy jego stacjonarność pomocą testu ADF.

Rysunek 10. Przebieg szeregu zróżnicowanego treningowego



Rysunek 11. Wykres ACF dla zróżnicowanych danych



Rysunek 12. Wykres PACF dla zróżnicowanych danych

Test ADF weryfikujący hipotezę o niestacjonarności dla uzyskanego szeregu - Augmented Dickey-Fuller Test

Ponowne wykonanie testu na stacjonarność danych:

```
In [28]: adfuller(df_train_diff.values)

Out[28]: (-6.44977592207308,
          1.5337342029244693e-08,
          12,
          251,
          {'1%': -3.4566744514553016,
           '5%': -2.8731248767783426,
           '10%': -2.5729436702592023},
          1168.0219598430886)
```

Rysunek 13. Wykonanie testu na stacjonarność dla zróżnicowanych danych

Wartość p-value spadła do bardzo małej wartości. Oznacza to, że tym razem otrzymany szereg jest szeregiem stacjonarnym.

3. Modelowanie danych przy pomocy ARMA

Dobranie rzędu modelu - kryteria informacyjne oraz estymacja parametrów.

Pewnym rozwiązaniem powyższych problemów są kryteria informacyjne. Stosujemy je w następujący sposób. Dla ustalonego zbioru modeli obliczamy wartość kryterium, poczym wybieramy model, dla którego wartość kryterium jest najmniejsza. Dwa najstarsze i najbardziej znane to AIC i BIC.

$$AIC(\hat{\beta} = -2log(L(\hat{\beta})) + 2k$$
$$BIC(\hat{\beta} = -2log(L(\hat{\beta})) + 2klog(n),$$

gdzie k to ilość niezerowych współrzędnych wektora współczynników $\hat{\beta}$. Pierwszy składnik odpowiada za jakość estymacji, natomiast drugi ma za zadanie ograniczać ilość zmiennych w modelu.

Dalej zajmiemy się modelem ARMA używając zróżnicowanego zbioru. Postaramy się dobrać do niego parametry za pomocą kryteriów informacyjnych.

51]:

	p	q	AIC	BIC	HQIC
28	4	4	1289.792171	1325.551662	1304.161422
34	5	4	1292.453146	1331.788586	1308.259322
33	5	3	1298.490224	1334.249715	1312.859475
23	3	5	1298.844254	1334.603745	1313.213505
22	3	4	1299.196816	1331.380357	1312.129141

Rysunek 14. Wybieranie najlepszych parametrów p, q na podstawie wyników za pomocą kryteriów informacyjnych

Najlepszymi parametrami okazuje się p = 4 oraz q = 4. Tych wartości użyjemy do estymacji parametrów modelu.

SARIMAX Results

Dep. Variable:	AvgTemperature	No. Observations:	264
Model:	ARIMA(4, 0, 4)	Log Likelihood	-634.896
Date:	Wed, 08 Feb 2023	AIC	1289.792
Time:	18:35:41	BIC	1325.552
Sample:	01-01-1996 - 12-01-2017	HQIC	1304.161
Covariance Type:	opg		

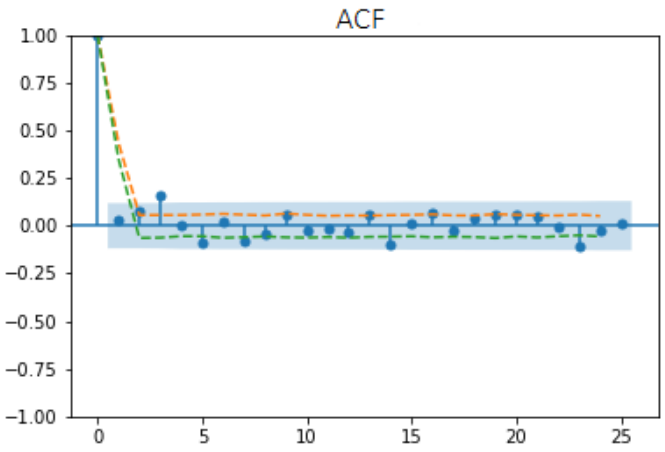
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0288	0.232	0.124	0.901	-0.425	0.483
ar.L1	-0.5973	0.073	-8.220	0.000	-0.740	-0.455
ar.L2	-0.1156	0.077	-1.510	0.131	-0.266	0.034
ar.L3	-0.4883	0.075	-6.494	0.000	-0.636	-0.341
ar.L4	-0.5902	0.074	-7.925	0.000	-0.736	-0.444
ma.L1	0.7031	0.498	1.412	0.158	-0.273	1.679
ma.L2	0.2535	0.152	1.673	0.094	-0.044	0.550
ma.L3	0.7374	1.231	0.599	0.549	-1.676	3.151
ma.L4	0.9649	0.977	0.988	0.323	-0.950	2.880
sigma2	6.9768	6.965	1.002	0.317	-6.675	20.629

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	5.73
Prob(Q):	0.92	Prob(JB):	0.06
Heteroskedasticity (H):	0.90	Skew:	0.03
Prob(H) (two-sided):	0.61	Kurtosis:	3.72

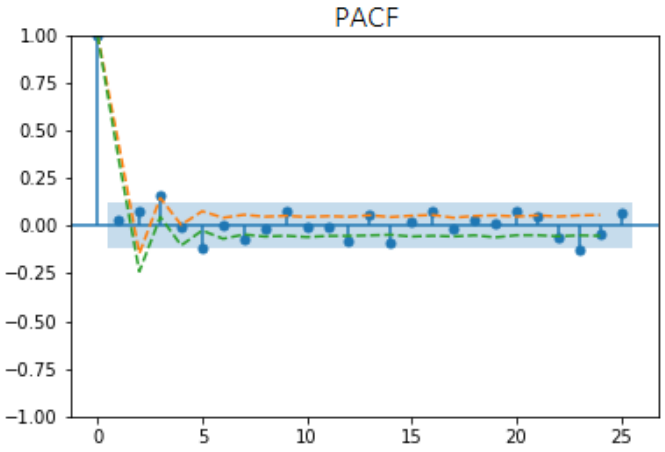
Rysunek 15. Cechy naszego modelu

4. Ocena dopasowania modelu

Chcielibyśmy dowiedzieć się, na ile nasz model jest wiarygodny.



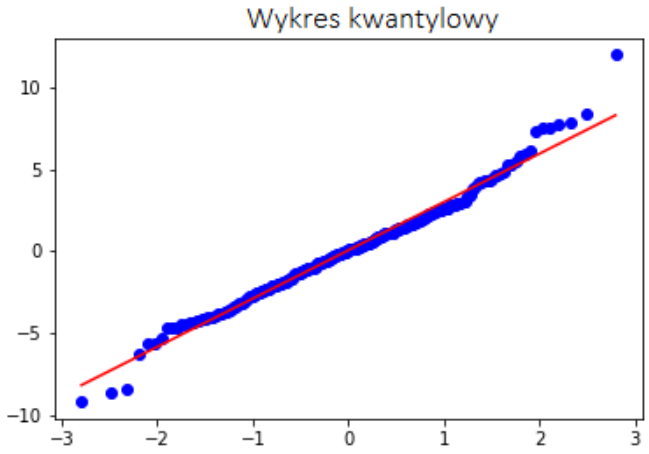
Rysunek 16. Wykres autokorelacji z przedziałami ufności.



Rysunek 17. Wykres częściowej autokorelacji z przedziałami ufności.

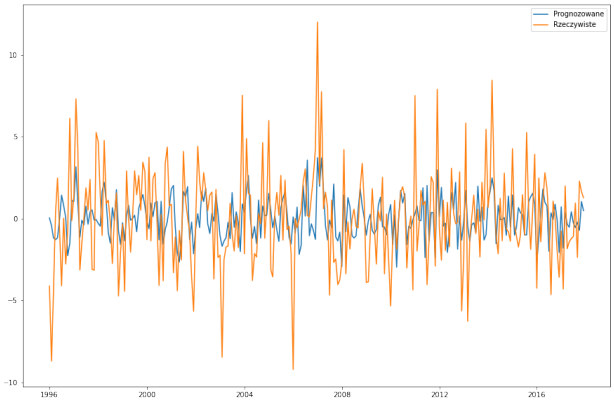
Widzimy, że wartości funkcji w większości wpadają w przedziały ufności. Pozwala nam to zaufać powstałemu modelowi i na tej podstawie przejdziemy do prognozy dla przyszłych obserwacji.

Możemy przeanalizować nienormalność rozkładu na podstawie zachowania ogona. Wykres poniżej porównuje kwantyle empirycznego rozkładu (niebieska linia) z kwantylami standardowego rozkładu normalnego (czerwona przerywana linia). W tym samym duchu informujemy o znacznym odchyleniu rozkładu empirycznego od kwantyli normalnych, dostarczając wyraźnych dowodów na ciężkie ogony i ujemną skośność.



Rysunek 18. Porównanie linii kwantylowych z trajektorią.

Symulujemy dane za pomocą modelu ARMA(2,4), tak aby były jak najbliższe prawdziwym. Aby zmierzyć dopasowanie modelu najpierw sprawdźmy jak dopasował się on do danych treningowych za pomocą wykresu.



Rysunek 19. Porównanie prognozy dla przyszłych obserwacji i rzeczywistych danych.

Można zauważyć, że wartość niebieskie (prognozowane) w pewnym zakresie pokrywają się z wartościami rzeczywistymi. Model bardzo słabo odwzorował wartości rzeczywiste. W pierwszych miesiącach starał się modelować różnice, jednak im dalsza była prognoza tym bardziej wartości predykowane dążyły do zera. Za pomocą metryki mean_absolute_error ocenimy dopasowanie modelu na zbiorze testowym oraz treningowym.

```
Błąd MAE prognozy dla zbioru treningowego: 2.103873468302429
Błąd MAE prognozy dla zbioru testowego: 2.5189992337196245
```

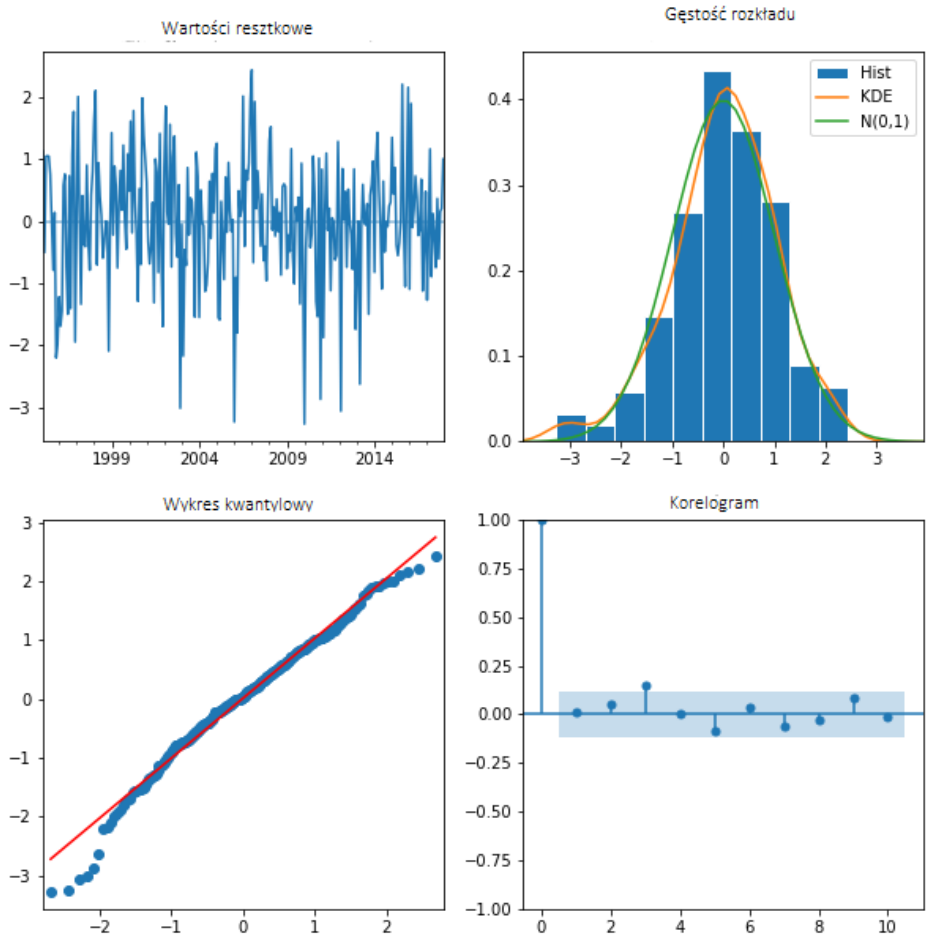
Rysunek 20. Błąd MAE

Błąd MAE w okolicy 2 oznacza, że nasze wartości różnią się średnio o 2 w stosunku do wartości rzeczywistych. Jest to więc całkiem duży błąd biorąc pod uwagę, że nasze różnice kształtują się na poziomie -6 do 6.

5. Weryfikacja założeń dotyczących szumu

Podczas tworzenia modelu ARMA oraz dalszych obliczeń zakładaliśmy warunki:

— założenie o średniej
— założenie o stałości wariancji
— założenie o niezależności
— założenie o normalności rozkładu



Rysunek 21. Analiza założeń

Wartości resztowe spełniają założenia rozkładu normalnego. Patrząc na histogram zauważyć możemy rozkład Gaussowski. Aby zbudowany model można uznać za poprawny wartość reszt

nie powinny być ze sobą skorelowane: jak widzimy na korelogramie warunek ten jest spełniony.

Test ARCH jest testem szumu, ale dla kwadratowych szeregów czasowych. Innymi słowy, badamy autokorelację wyższego rzędu(nieliniową). Sprawdzimy testem ARCH, czy zachowana jest wariancja w resztkach.

```
(17.093680448013323,
 0.07231660208098335,
 1.7513852981407754,
 0.07004057749996509)
```

p.value na poziomie 0.07. Jest to wartość lekko powyżej 0.05 co oznacza, że stałość wariancji nie jest zachowana.

6. Zakończenie

Podsumowując: Model dopasował się do danych i potrafił je odwzorować w niewielkim stopniu. Zaprognozowaliśmy różnicę w pogodzie na kolejne 2.5 roku. Oznacza to, że dzięki temu modelowi możemy przewidywać średnią pogodę na kolejne lata. Warto jednak zastrzeć, że prognozowanie takiej pogody raczej nie jest dokładne. Cechą modelów statystycznych, którą warto odnotować, jest stałość obliczeń. W przeciwieństwie do modeli uczenia głębokiego, czy niektórych modeli uczenia maszynowego (których również można używać do prognozowania), dopasowanie modelu po każdym treningu będzie takie samo.

7. Źródła

1. https://www.kaggle.com/datasets/sudalairajkumar
2. "Analiza szeregów czasowych-Time Series Analysis" część 2 Jerzy Stefanowski Politechnika Poznańska Projekt eksploracji danych Poznań 2020