Analiza minimalnej, dziennej temperatury w Melbourne przy pomocy modelu ARMA

Magdalena Szymkowiak, Adrian Sobczak

19 stycznia 2024

Spis treści

1 Wstęp

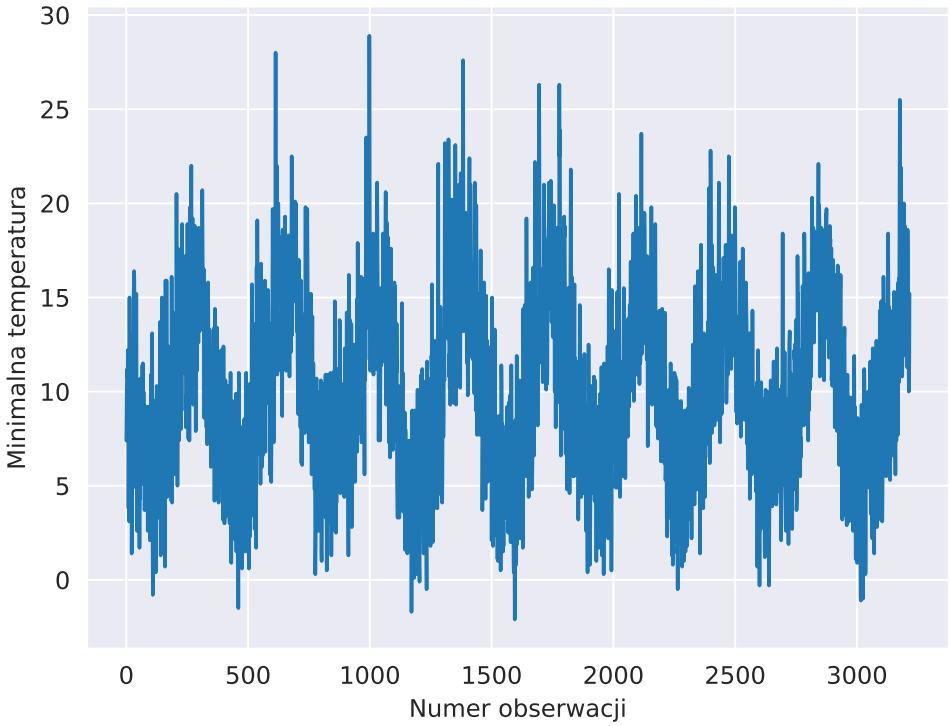
W niniejszym raporcie zajmiemy się analizą wartości minimalnej, dziennej temperatury w Melbourne od września 1943 roku do lutego 2023 roku. Wskaźnik temperatury, wyznaczany co okres jednego dnia na terenie miasta, intuicyjnie jest silnie zależny od swojej wartości w dniach poprzedzających. Bazując na intuicji wychodzimy z propozycją, że owy szereg czasowy wartości temperatury, da się opisać modelem ARMA. Zaczniemy od sprawdzenia czy dla surowych danych spełnione są założenia modelu ARMA i w zależności od werdyktu użyjemy przekształconych lub nieprzekształconych danych do dalszej analizy. Następnie z pomocą kryteriów informacyjnych wyestymujemy parametry modelu ARMA i przejdziemy do oceny jego dopasowania do danych rzeczywistych. Ostatnim etapem analizy będzie weryfikacja założeń dotyczących szumu. Celem całości jest ocena czy mamy podstawy zasadnie modelować minimalną, dzienną temperaturę w Melbourne modelem ARMA.

Dane, które będziemy analizować pobraliśmy z australijskiej, rządowej strony biura meteorologicznego [1]. Cały zbiór danych posiada 29257 rekordów.

2 Przygotowanie danych do analizy

2.1 Zbadanie jakości danych

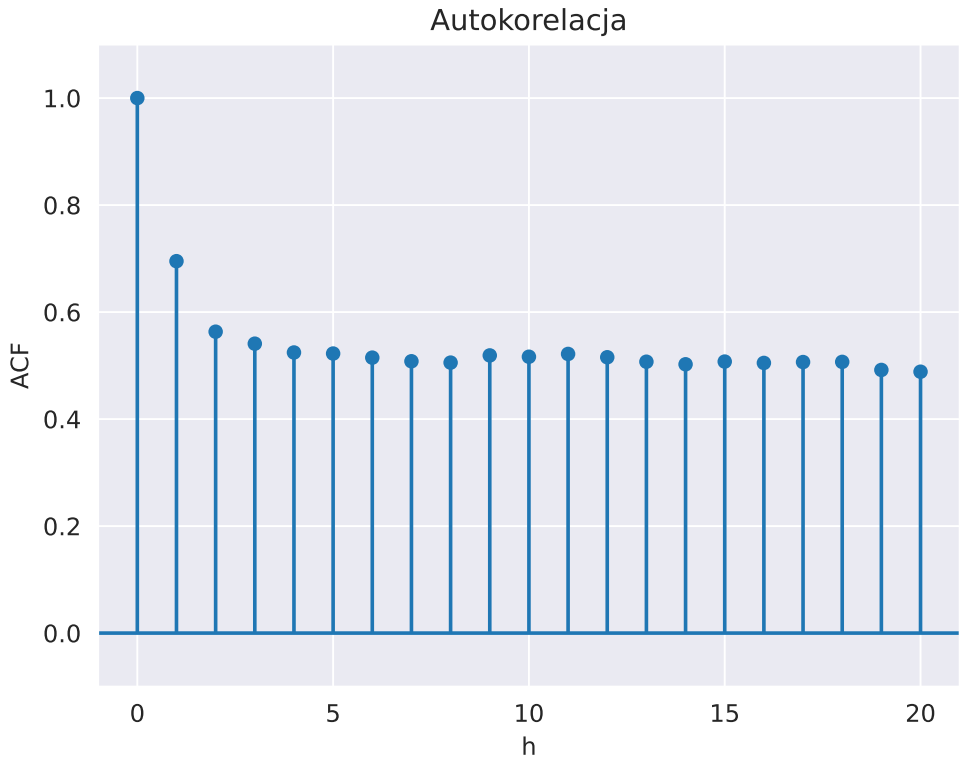
Pierwszym krokiem jaki podjęliśmy było usunięcie wszystkich wierszy, w których brakowało pomiaru najniżej temperatury. Następnie, wyodrębniliśmy 3217 najnowszych obserwacji, ponieważ przy tak ogromnej ilości danych trudno było zauważyć pewne zależności co mogło skutkować pogorszeniem jakości naszej analizy. Dzięki temu krokowi poprawiliśmy także czytelność wizualizacji danych. Wyodrębnione dane zawierają prawie 9 lat danych, tj. od kwietnia 2014 r. do początku lutego 2023 r.



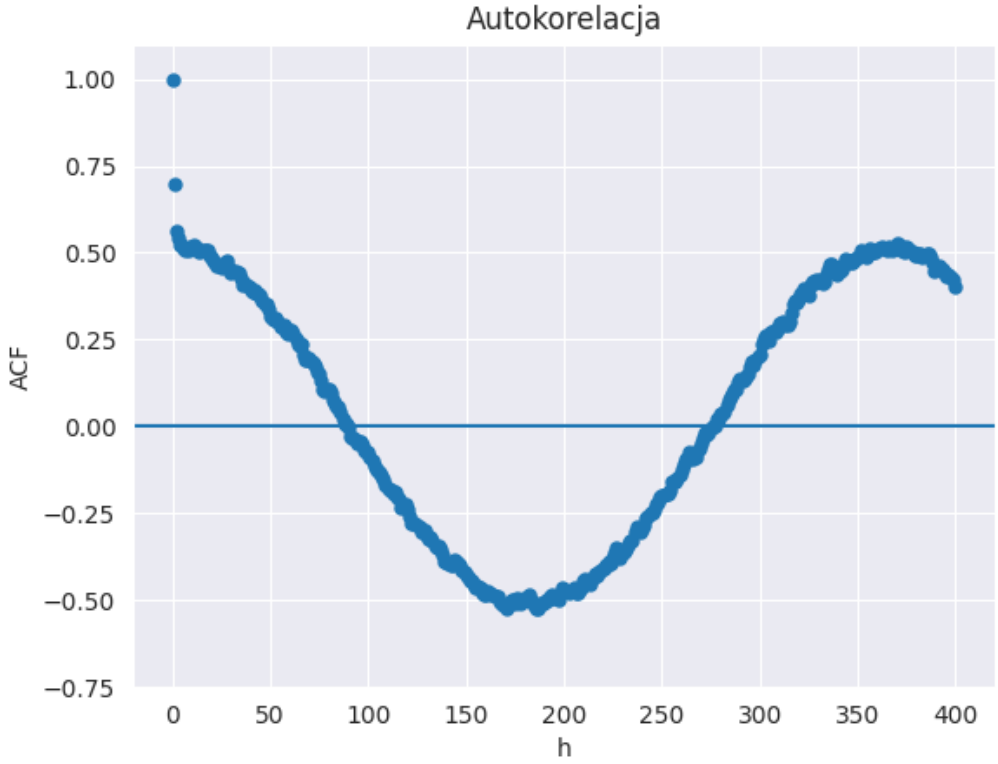
Rysunek 1: Wykres danych dziennej minimalnej temperatury w Melbourne.

2.2 Weryfikacja stacjonarności szeregu

Do sprawdzenia czy nasze dane są stacjonarnym szeregiem czasowym zaczniemy od przedstawienia funkcji autokorelacji oraz częściowej autokorelacji dla surowych danych.

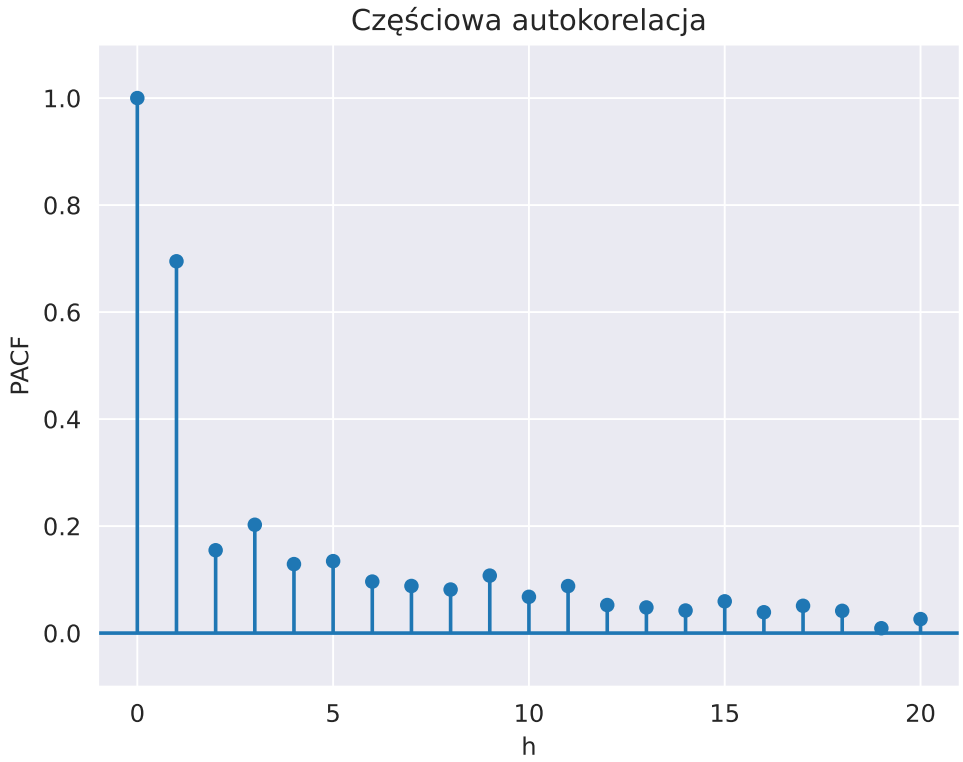


Rysunek 2: Wykres funkcji autokorelacji surowych danych, do 20. przesunięcia.

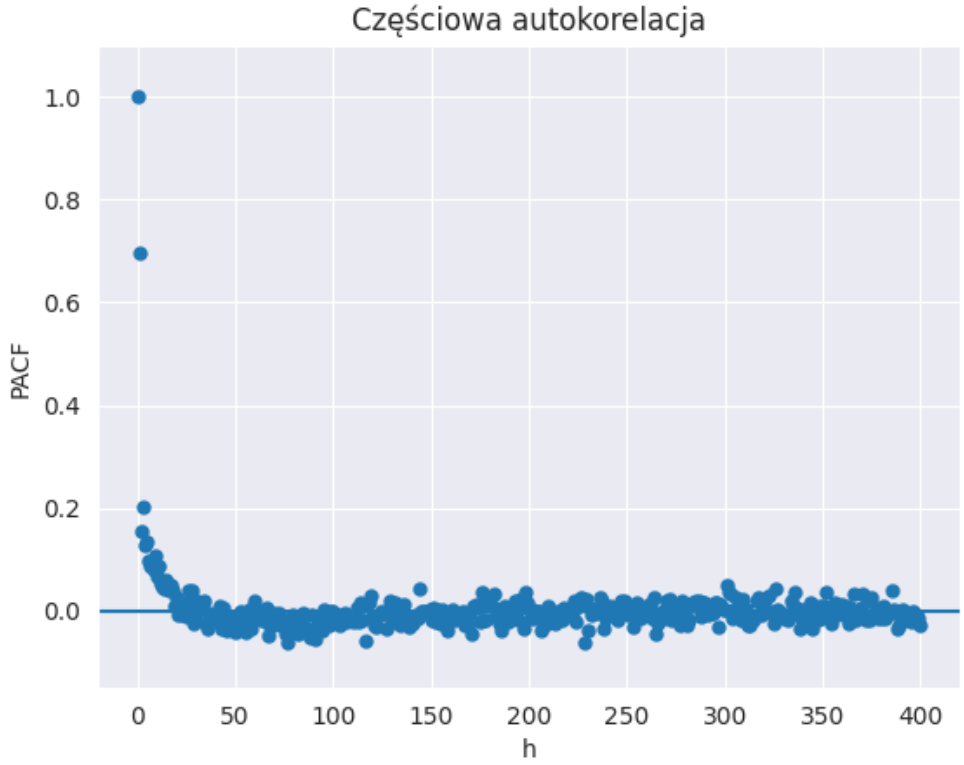


Rysunek 3: Wykres funkcji autokorelacji surowych danych, do 400. przesunięcia.

Wizualizacja funkcji autokorelacji przedstawiona na wykresach 2 i 3 jednoznacznie zaprzecza stacjonarności szeregu. Sinusoidale zachowanie dla dużych przesunięć jest odbiciem wyraźnej sezonowości danych którą widać na wykresie 1.



Rysunek 4: Wykres funkcji częściowej autokorelacji surowych danych, do 20. przesunięcia.



Rysunek 5: Wykres funkcji częściowej autokorelacji surowych danych, do 400. przesunięcia.

Sama funkcja częściowej autokorelacji przedstawiona na wykresach 4 i 5 nie wyklucza stacjonarności.

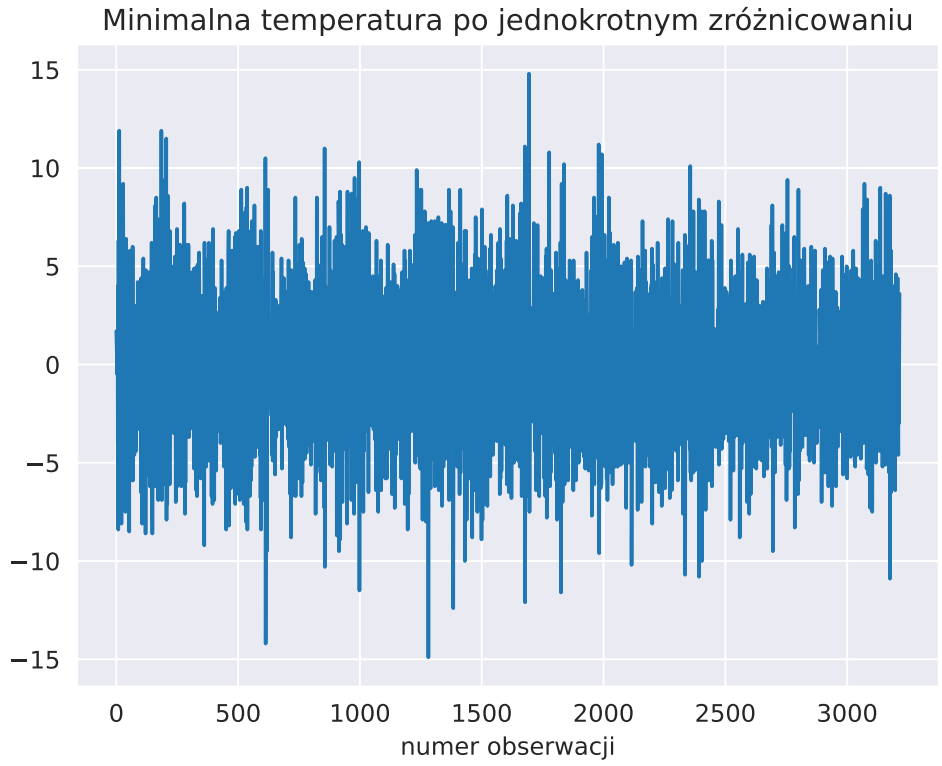
Zważając na brak zbieżności do zera funkcji autokorelacji wnioskujemy, że surowy szereg nie jest stacjonarny. Wynika to niejako z samej struktury danych temperaturowych, a dokładnie z sezonowości pór roku. Nas nie interesuje jednak jak temperatura zachowuje się na przestrzeni lat. Interesuje nas zależność od kilku ostatnich dni. Stąd zdecydujemy się na dekompozycję szeregu, aby pozbyć się informacji o sezonowości w większej skali, co pozwoli nam użyć modelu ARMA, zależnego tylko od kilku poprzedników.

2.3 Dekompozycja szeregu

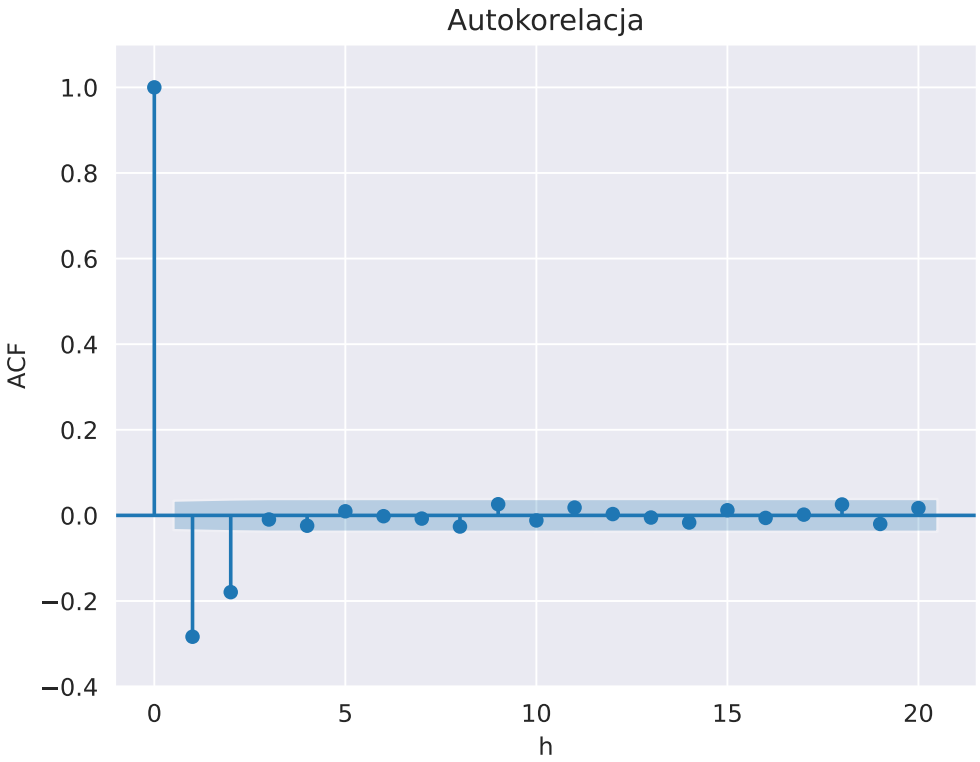
Jako metodę dekompozycji obraliśmy różnicowanie. Różnicowanie dla próbki $\{x_t\}_{t\in T}$, gdzie $T = \{1, \ldots, n\}$, tworzy nową próbkę, postaci

$$\{\Delta x_t : \Delta x_t = x_t - x_{t-1}, t \in T \setminus \{1\}\}.$$

Zastosowanie tej operacji pozwala na usunięcie trendu liniowego oraz sezonowości z szeregu czasowego. Nie jest to prawdą dla każdego szeregu, jednak dla naszego przypadku, pojedyncze różnicowanie znacznie ograniczyło sezonowość co widać na wykresie 6.



Rysunek 6: Wykres danych po jednokrotnym zróżnicowaniu.

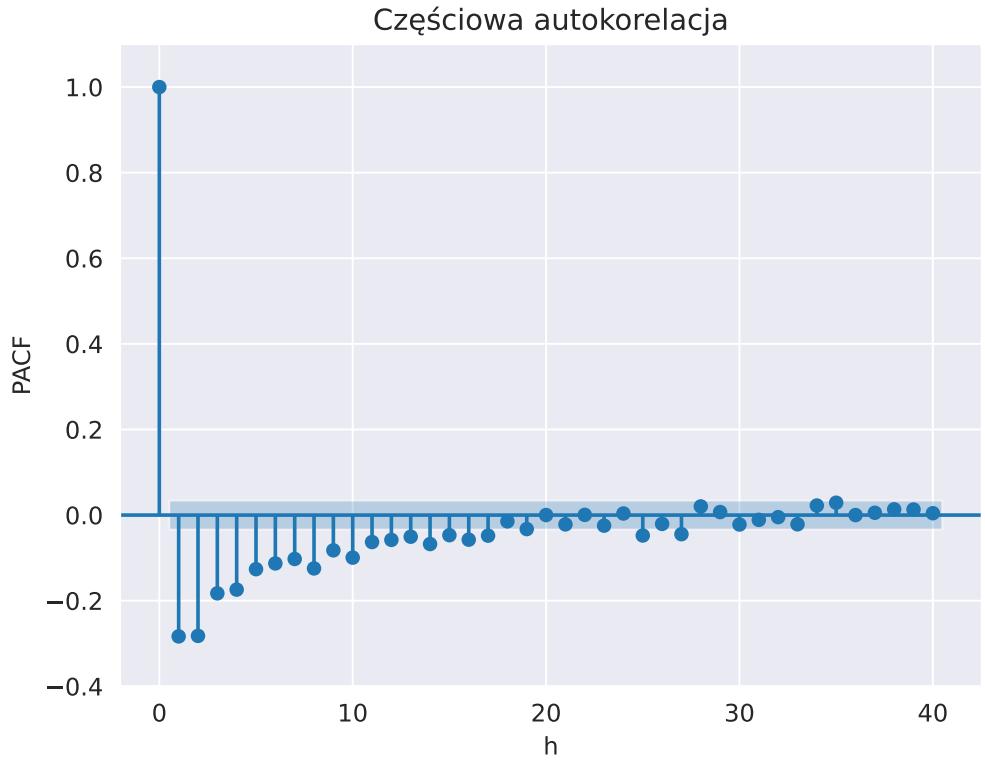


Rysunek 7: Wykres funkcji autokorelacji po zróżnicowaniu danych. Zakres do 20. przesunięcia.

Funkcja autokorelacji przedstawiona na wykresie 7 dla zróżnicowanego szeregu staje się bliska zeru już od trzeciego przesunięcia. Nie wykluczamy więc na jej podstawie stacjonarności.

Statystyka testowa	p-wartość
-16.248	0.000

Tabela 1: Wynik testu ADF dla zróżnicowanych danych.



Rysunek 8: Wykres funkcji częściowej autokorelacji po zróżnicowaniu danych. Zakres do 40. przesunięcia.

Funkcja częściowej autokorelacji przedstawiona na wykresie 8, podobnie jak funkcja autokorelacji, jest bliska zeru dla przesunięć większych od 20.

Zarówno funkcja autokorelacji oraz funkcja częściowej autokorelacji nie wykluczyły stacjonarności szeregu. Warto również wspomnieć o tym że nie mamy żadnych oczywistych przesłanek przeciwko stacjonarności w samym wykresie danych 6. Podejmiemy jednak w tym temacie jeszcze jeden krok, którym będzie test ADF. Wyniku testu są przedstawione w tabeli 1. P-wartość poniżej umownego poziomu istotności 0.05 nie daje nam powodu do odrzucenia hipotezy zerowej opowiadającej się za stacjonarnością szeregu.

Na podstawie powyższych rozważań postulujemy, że zróżnicowany szereg jest

AIC	p	q
16407.658941	1	2
16408.266272	0	3
16408.664459	0	4
16408.845527	2	2
16408.869681	2	1
16409.155815	3	2
16409.278379	1	4
16409.377718	2	3
16409.381358	4	1
16410.033712	4	4

Tabela 2: Wartości krytertium informacyjnego AIC, posortowane od najmniejszego, względem wartości parametrów p i q.

stacjonarny. Pozwala nam to zastosować modelowanie z użyciem modelu ARMA którym zajmiemy się w następnym rozdziale.

3 Modelowanie danych przy pomocy ARMA

3.1 Wyznaczenie rzędu modelu

Pierwszym krokiem w dobraniu modelu ARMA będzie wyznaczenie rzędu modelu. Aby to osiągnąć posłużymy się kryteriami informacyjnymi AIC, BIC oraz HQIC. Kryteria informacyjne wyznaczymy za pomocą modelu dostępnego w bibliotece Pythona `statsmodels.tsa.arima.model.ARIMA`, dla par (p, q) na zbiorze $A \times A$, gdzie $A = \{1, 2, 3, 4, 5\}$.

Patrząc na wartości kryteriów przedstawione w tabelach 2, 3 oraz 4, szukamy pary (p, q) dla której kryteria będą minimalizowane. Bezkonkurencyjnym faworytem jest tutaj para $(1, 2)$, która minimalizuje dwa kryteria AIC i HQIC, a względem kryterium BIC jest na drugim miejscu. Bazując na tym rankingu przyjmujemy, że nasz zróżnicowany ciąg temperatur będziemy modelować za pomocą modelu ARMA(1,2).

3.2 Estymacja parametrów modelu

Ostatnim krokiem dobierania modelu będzie wyestymowanie parametrów, dla znanengo rzędu. Użyjemy w tym celu tego samego modelu co do wyznaczenia kryteriów informacyjnych, tj. `statsmodels.tsa.arima.model.ARIMA` w Pythonie.

BIC	p	q
16435.558619	0	2
16438.038409	1	2
16438.645740	0	3
16439.249150	2	1
16445.119821	0	4
16445.142683	1	1
16445.300889	2	2
16446.662906	3	1
16450.702452	1	3
16451.687070	3	2

Tabela 3: Wartości krytertium informacyjnego BIC, posortowane od najmniejszego, względem wartości parametrów p i q.

HQIC	p	q
16418.547776	1	2
16419.155107	0	3
16419.758517	2	1
16419.966112	0	2
16421.731062	0	4
16421.912130	2	2
16423.274146	3	1
16424.400184	3	2
16424.522749	1	4
16424.622087	2	3

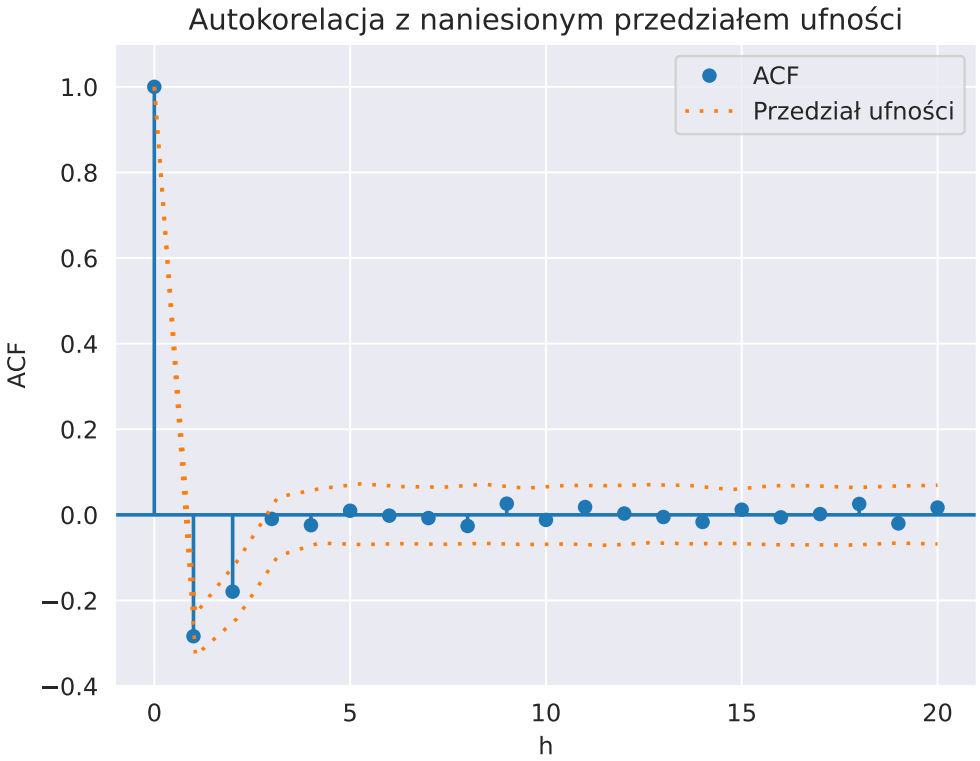
Tabela 4: Wartości krytertium informacyjnego HQIC, posortowane od najmniejszego, względem wartości parametrów p i q.

	ϕ_1	θ_1	θ_2	σ^2
wartość parametru	0.1397	-0.6915	-0.1981	9.5878
odchylenie standardowe	0.056	0.056	0.047	0.222
przedziały ufności $\alpha = 0.05$	$[0.030, 0.250]$	$[-0.800, -0.583]$	$[-0.289, -0.107]$	$[9.153, 10.023]$

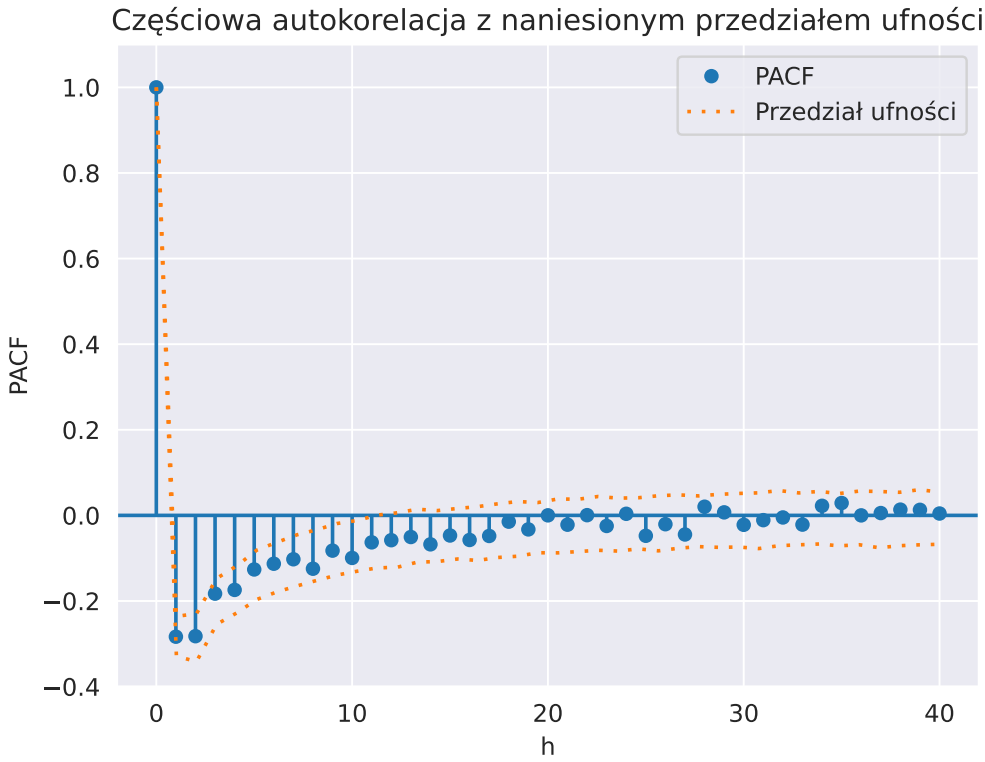
Tabela 5: Wyniki estymacji parametrów przy założeniu, że dane pochodzą z modelu ARMA(1,2).

4 Ocena dopasowania modelu

Do zbadania jakości dopasowania modelu wyznaczyliśmy przedziały ufności dla funkcji autokorelacji, otrzymane przez wysymulowanie wartości z modelu teoretycznego (Monte Carlo, 1000 powtórzeń) i wyznaczenie odpowiednich kwantyli dla poziomu ufności $\alpha = 0.05$.



Rysunek 9: Wykres funkcji autokorelacji po zróżnicowaniu danych wraz z przedziałem ufności na poziomie $\alpha = 0.05$. Zakres do 20. przesunięcia.



Rysunek 10: Wykres funkcji częściowej autokorelacji po zróżnicowaniu danych wraz z przedziałem ufności na poziomie $\alpha = 0.05$. Zakres do 40. przesunięcia.

Na wykresie 9 widzimy, że wartości funkcji autokorelacji mieszczą się w przedziałach ufności naszego wyestymowanego modelu. Podobną sytuację możemy zaobserwować dla funkcji częściowej autokorelacji na wykresie 10.

Uznajemy nasze wyniki za bardzo zadowalające. Fakt, że wszystkie wartości funkcji autokorelacji i autokorelacji próbkowej mieszczą się w przedziałach ufności, mówi nam że w języku tych dwóch metryk nasze dane istotnie wpasowują się w model ARMA(p, q).

5 Weryfikacja założeń dotyczących szumu

Biały szum to szereg czasowy $\{Z\}_t$, taki że

$$\forall_t \, \mathrm{E}\left[Z_t\right] = 0 \quad \wedge \quad \forall_{t \neq s} \, \mathrm{Cov}\left[Z_t, Z_s\right] = 0.$$

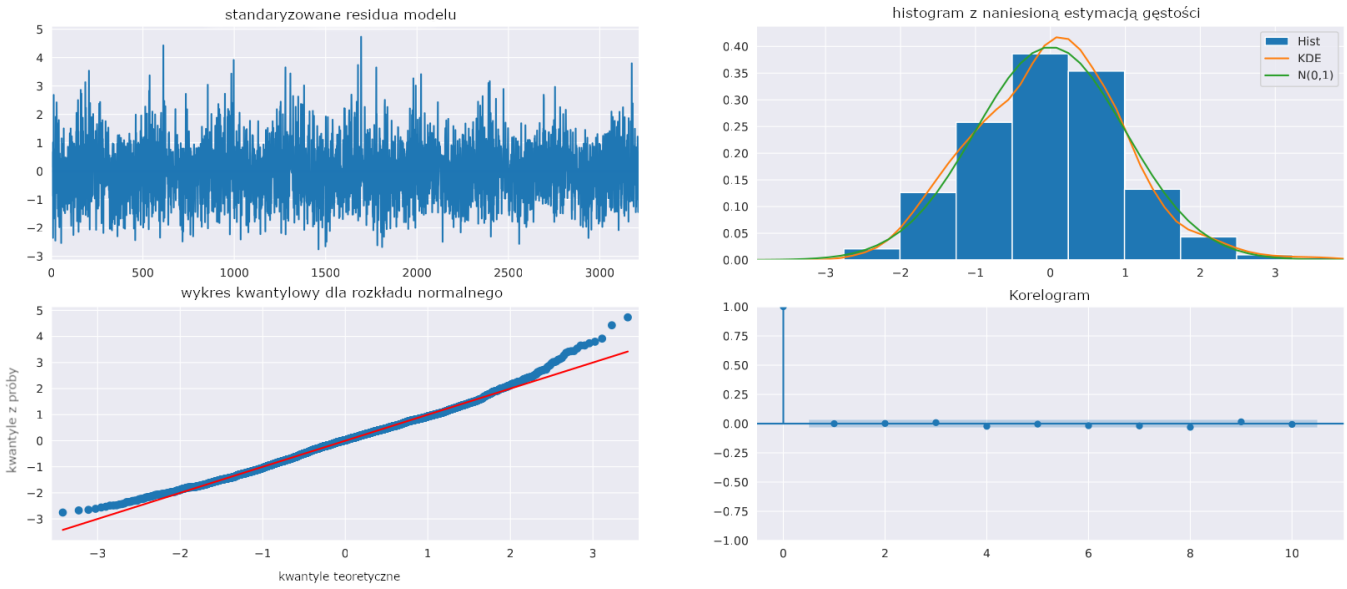
Istotnym założeniem modelu ARMA jest fakt, że szereg odpowiadający za część

Statystyka testowa	p-wartość
0.012	0.991

Tabela 6: Wynik t-testu dla residuów.

średniej ruchomej jest białym szumem. Na wykresie 11 przedstawiliśmy cztery wizualizacje tego szeregu dla naszych danych. Na histogramie oraz wykresie kwantylowym widzimy podobieństwo do rozkładu normalnego. Nie jest to jednak warunek konieczny białego szumu. Wartości korelogramu są bardzo bliskie zeru dla niezerowych przesunięć. Ten fakt opowiada się za nieskorelowaniem szeregu i dostarcza nam argumentów dzięki którym możemy wnioskować że residua to biały szum.

Estymator średniej zbliżony jest do zera. Dodatkowo, wykonaliśmy też t-test do potwierdzenia założenia dotyczącego zerowej średniej. Wyniki testu zamieszczone w tabeli 6, potwierdzają hipotezę zerową, dla której średnia residuów jest równa 0.



Rysunek 11: Wykres rozproszenia, histogram z naniesionymi gęstościami, wykres kwantylowy oraz korelogram dla residuów dobranego modelu.

6 Zakończenie

Podsumowując naszą analizę, szukając modelu ARMA który zamodeluje dzienną minimalną temperaturę w Melbourne, zróżnicowaliśmy dane i ostatecznie uzyska-

liśmy model ARMA(1,2) wyrażony wzorem

$$X_t - 0.1397 \cdot X_{t-1} = Z_t - 0.6915 \cdot Z_{t-1} - 0.1981 \cdot Z_{t-2},$$

gdzie $Z_t \sim \mathrm{WN}(0, 9.5875)$. Żadna z przeprowadzonych przez nas weryfikacji nie dała nam żadnych przesłanek do odrzucenia tego modelu. Stąd postulujemy, że nasz model istotnie modeluje jednokrotnie zrożnicowaną minimalną temperaturę w Melbourne.

Zauważmy też, że wybraliśmy model arma odgórnie. Fakt, że nie spotkaliśmy na drodze naszej analizy żadnej przesłanki o niezasadności naszego wyboru, świadczy o tym, że model ARMA może być rzeczywistym modelem, z którego pochodzą dane obserwacje. Potwierdzenie modelu teoretycznego w mierzalnym eksperymencie fizycznym jest istotną przesłanką za sensownością modelu ARMA.

Bibliografia

[1] Australijski rząd biura meteorologicznego. *Climate Data Online*. URL: `http://www.bom.gov.au/climate/data/`.