

Prognozowanie sezonowe - porównanie metod

Teoria Algorytmów i Obliczeń

Piotr Widomski

02.01.2022

1 Cel projektu

Celem projektu jest porównanie wydajności prognozowania sezonowych szeregów czasowych za pomocą maszyny **SARIMA** (**Seasonal ARIMA**) oraz liniowej kombinacji maszyn **ARIMA** działających na co n -tym elemencie szeregu.

2 Użyte dane

Do porównania metod użyte zostały dane przedstawiające średnią miesięczną temperaturę powietrza mierzoną na dublińskim lotnisku od roku 1941. Dane zostały pobrane z portalu inicjatywy Open Data prowadzonej przez rząd irlandzki. Pomiary wykonane zostały przez Met Éireann - irlandzki narodowy serwis meteorologiczny i udostępnione na licencji CC Attribution 4.0.

Na potrzeby projektu użyty został wycinek danych zawierających pełne lata, od 1942 do 2019 roku, czyli 78 lat pomiarów. Dane te zostały podzielone na treningowe oraz testowe w stosunku około 80%, gdzie dane treningowe zawierają 62 lata, od 1942 do 2003 roku, a dane testowe - 15 lat, od 2004 do 2019 roku. Rok 2021 został pominięty ze względu na niekompletność danych, natomiast 2020 - z powodu brakujących danych z marca.

3 Opis metod

3.1 ARIMA

ARIMA jest klasą modeli, które opisują szereg na podstawie poprzednich wartości. Model ARIMA składa się z trzech procesów:

- AR (autoregresyjny) - każda wartość jest liniową kombinacją pewnej liczby poprzednich wartości. Liczbę poprzednich wartości użytych przy obliczeniu następnej oznaczamy przez p , a proces autoregresyjny z rzędem regresji p - AR(p) i przedstawiamy następująco:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$$

gdzie y_{t-i} jest i -tą poprzednią wartością ciągu w chwili t , ϕ_i jest współczynnikiem i -tej poprzedniej wartości, ϵ_t zaburzeniem w momencie t , a α - wartością podstawową.

- MA (średniej ruchomej) - każda wartość jest zależna od zaburzenia w chwili obecnej oraz wcześniejszych. Liczbę poprzednich zaburzeń uwzględnionych przy obliczeniu następnej wartości oznaczamy przez q , a proces średniej ruchomej rzędu q - MA(q) i przedstawiamy:

$$y_t = \alpha + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

gdzie ϵ_t jest zaburzeniem w momencie t , θ_i - współczynnikiem i -tego poprzedniego zaburzenia, a α - wartością podstawową.

- I (integracja) - w celu modelowania ciągów niestacjonarnych, model ARIMA może wykonywać różnicowanie pewnego stopnia, w celu operowania na bardziej stacjonarnym ciągu. W takim wypadku, zamiast następnej wartości, model prognozuje różnicę wartości. Jednak dzięki znajomości poprzedniej wartości możemy wyliczyć na podstawie różnicy następną wartość. Stopień różnicowania oznacza się przez d .

Łącząc powyższe procesy, otrzymujemy ogólny wzór modelu ARIMA:

$$y_t^* = \alpha + \sum_{i=1}^p \phi_i y_{t-i}^* + \sum_{i=1}^q \theta_i \epsilon_{t-i} + \epsilon_t$$
$$y_t^* = \Delta^d y_t$$

Poszczególne modele ARIMA definiuje się za pomocą trzech zmiennych całkowitych p, d, q i zapisuje ARIMA(p, d, q).

3.2 SARIMA

SARIMA, lub **Seasonal ARIMA**, jest rozszerzeniem modelu ARIMA wspierającym ciągi czasowe wykazujące sezonowość, czyli takie, które posiadają powtarzające się z pewną określaną częstością zmiany wartości. Częstość tę oznaczmy przez S .

Rozszerzenie to dodaje trzy nowe, analogiczne do modelu ARIMA procesy:

- Sezonową autokorelację - stopień oznaczamy przez P
- Sezonową średnią ruchomą - stopień oznaczamy przez Q
- Sezonową integrację - stopień oznaczamy przez D

Procesy te łączą się z procesami modelu ARIMA w sposób multiplikatywny. Z tego powodu, w celu łatwiejszego zdefiniowania wzoru modelu SARIMA, zdefiniujmy operator poprzednika B :

$$By_t = y_{t-1}$$

$$B^i y_t = y_{t-i}$$

Używając powyższego operatora możemy przekształcić model ARIMA do poniższej postaci:

$$\underbrace{\left(1 - \sum_{i=1}^p \phi_i B^i\right)}_{\text{AR}(p)} \cdot \underbrace{(1 - B)^d}_{\text{I}(d)} y_t = \underbrace{\left(1 + \sum_{i=1}^q \theta_i B^i\right)}_{\text{MA}(q)} \epsilon_t$$

Uwzględniając procesy sezonowe, o postaci analogicznej do postaci procesów nie sezonowych, otrzymujemy ogólny wzór modelu SARIMA:

$$\underbrace{\left(1 - \sum_{i=1}^p \phi_i B^i\right)}_{\text{AR}(p)} \cdot \underbrace{\left(1 - \sum_{i=1}^P \Phi_i B^{Si}\right)}_{\text{Sezonowe AR}(P)} \cdot \underbrace{(1 - B)^d}_{\text{I}(d)} \cdot \underbrace{(1 - B^S)^D}_{\text{Sezonowe I}(d)} y_t =$$

$$\underbrace{\left(1 + \sum_{i=1}^q \theta_i B^i\right)}_{\text{MA}(q)} \cdot \underbrace{\left(1 + \sum_{i=1}^Q \Theta_i B^{Si}\right)}_{\text{Sezonowe MA}(Q)} \epsilon_t$$

Poszczególne modele SARIMA definiuje się za pomocą zmiennych całkowitych niesezonowych p, d, q , zmiennych sezonowych P, D, Q oraz częstości sezonowej S i zapisuje SARIMA(p, d, q) x (P, D, Q) S .

3.3 Kombinacja liniowa maszyn ARIMA

Model ten ma za zadanie prognozować kolejne elementy szeregu używając grup maszyn ARIMA. Każda z grup składa się z unikatowej liczby maszyn n , gdzie $n|S$ oraz $n > 1$. i -ta maszyna w grupie n działa na co n -tym elemencie szeregu zaczynając od elementu i -tego. Dodatkowo w celu symulowania sezonowości maszyny te wykorzystują tylko ostatnie S elementów szeregu. Oznacza to że taka maszyna ma postać $\text{ARIMA}(\frac{S}{n}, 0, \frac{S}{n})$.

Podsumowując, każda z maszyn w grupie prognozuje i -ty następny element szeregu wykorzystując co i -ty z ostatnich S elementów szeregu. Łącząc wyniki maszyn w grupie otrzymujemy pełną prognozę.

Weźmy grupy maszyn spełniające powyższe warunki i ponumerujmy je $1, 2, \dots, n$. Oznaczmy przez $y_{i,t}$ element otrzymany z grupy i w momencie t . Wartość tą możemy uzyskać stosując wzór modelu ARIMA z maszyną z grupy i , której indeks wewnątrz grupy jest równy $t \bmod i$. Stosując kombinację liniową maszyn ARIMA otrzymujemy następujący wzór na następny element:

$$y_t = \sum_{i=1}^n \alpha_i y_{i,t}$$

gdzie α_i jest współczynnikiem i -tej grupy. Model ten będzie również nazywany grupą maszyn ARIMA lub ARIMA.

4 Implementacja

4.1 Biblioteka modeli

Do budowy modeli została użyta biblioteka `statsmodel`, znajdująca się w pakiecie języka Python o tej samej nazwie, zapewniająca implementacje modeli ARIMA oraz rozszerzenia SARIMA. Dodatkowo biblioteka ta umożliwia automatyczne dobranie odpowiednich współczynników na podstawie danych wejściowych oraz parametrów modelu oraz udostępnia metody umożliwiające prognozowanie kolejnych elementów ciągu. Wykorzystane zostały również metody ułatwiające analizę używanego ciągu czasowego, takie jak rozkład sezonowy, ACF oraz PACF, które opisane zostały w następnym rozdziale.

Dodatkowo, w celu dobrania optymalnych parametrów modeli, użyta została biblioteka `pmdarima`, która umożliwia wyszukanie optymalnych parametrów modelu na podstawie ustalonych kryteriów, po czym zwraca maszynę

utworzoną z dobranych przez siebie parametrów w implementacji kompatybilnej z tą, która znajduje się w bibliotece `statsmodels`.

4.2 Metody analizy

4.2.1 Rozkład sezonowy

Podczas analizy danych wykorzystany został rozkład sezonowy w modelu addytywnym. Rozkład ten polega na rozdzieleniu ciągu czasowego na trzy niezależne ciągi:

- trend - określa tendencje zmian wartości ciągu w danym momencie.
- element sezonowy - określa wartość elementu, która wynika z sezonowości ciągu. Jeżeli ciąg wykazuje sezonowość, to element sezonowy jest cyklem.
- element losowy - pozostałość wartości elementu po odjęciu trendu oraz elementu sezonowego. Odpowiada zaburzeniu, które w modelu **ARIMA** oznaczamy symbolem ϵ .

gdzie oryginalny ciąg możemy uzyskać poprzez ich sumę.

4.2.2 Funkcja autokorelacji

Funkcja autokorelacji (dalej ACF) opisuje w jakim stopniu element szeregu zależy od elementów poprzednich. Mając pomiary y_1, y_2, \dots, y_n wykonane w równych odstępach czasu wartość funkcji autokorelacji dla opóźnienia k zdefiniowana jest następująco:

$$r_k = \frac{\sum_{i=1}^{n-k} (y_i - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Funkcji ta została użyta w celu dobrania odpowiednich parametrów modelu. Przy konstrukcji wykresów użyte zostały kolejne opóźnienia od 0 do 30.

4.2.3 Funkcja częściowej autokorelacji

Funkcja częściowej autokorelacji (dalej PACF) wyznacza częściową korelację szeregu czasowego z jego własnymi opóźnionymi wartościami. Mając pomiary y_1, y_2, \dots, y_n wykonane w równych odstępach czasu wartość funkcji częściowej autokorelacji dla opóźnienia k zdefiniowana jest następująco:

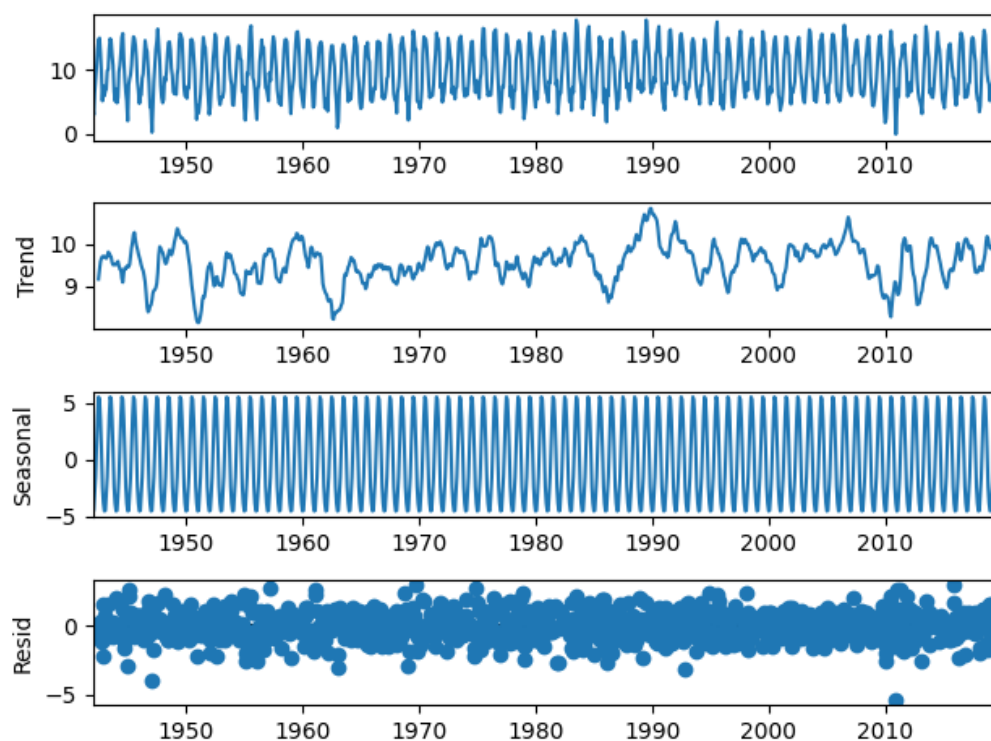
$$\rho_k = Corr(y_t, y_{t-k}) = \frac{\gamma_k}{\gamma_0}$$
$$\gamma_k = Cov(y_t, y_{t-k})$$

Funkcji ta została użyta w celu dobrania odpowiednich parametrów modelu. Przy konstrukcji wykresów użyte zostały kolejne opóźnienia od 0 do 30.

4.3 SARIMA

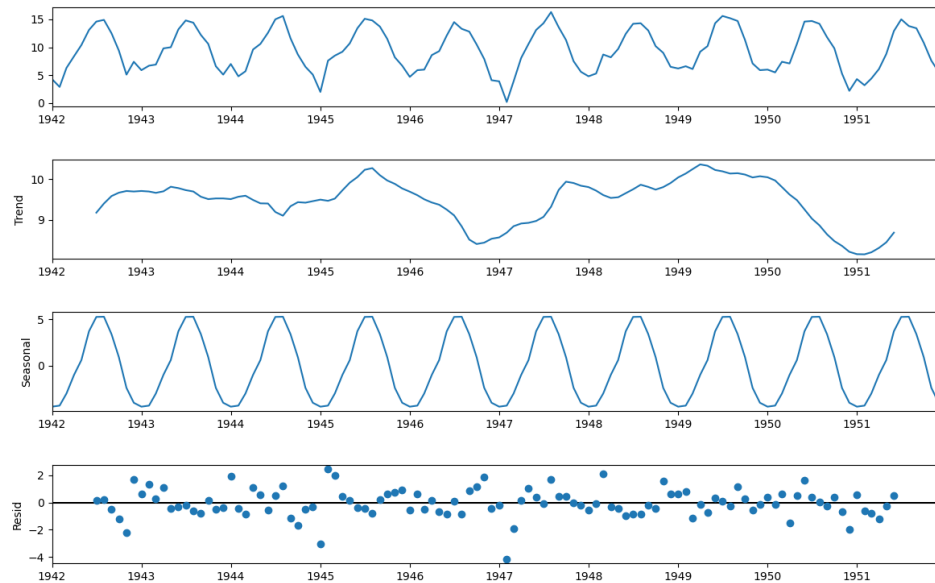
Przed konstrukcją modelu wykonana została analiza danych za pomocą algorytmu rozkładu sezonowego zaimplementowanego w bibliotece udostępnionej przez bibliotekę `statsmodels`. Parametry modelu dobrane zostały zgodnie ze wskazówkami przedstawionymi przez Kotas Hatalis w [1].

Wynik rozkładu prezentuje się następująco:



Rysunek 1: Rozkład sezonowy średniej miesięcznej temperatury w latach od 1942 do 2019. Przetawia kolejno: wartość, trend, element sezonowy oraz element losowy.

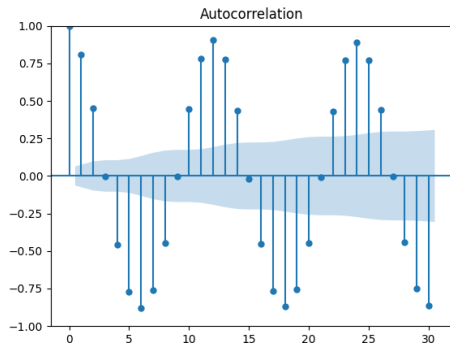
Wykres przedstawia kolejno: wartość, trend, element sezonowy oraz element losowy każdego elementu ciągu. Z uwagi na dużą ilość elementów, wykres ten jest trudny w analizie. Z tego powodu, podczas analizy, skupiono się na pierwszych dziesięciu latach pomiarów:



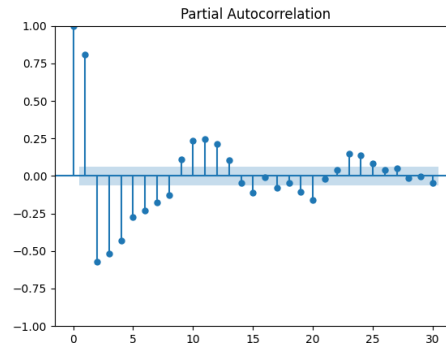
Rysunek 2: Rozkład sezonowy danych średniej miesięcznej temperatury w latach od 1942 do 1952. Przetawia kolejno: wartość, trend, element sezonowy oraz element losowy.

Z rozkładu danych zauważono, że zgodnie z oczekiwaniami dane posiadają stały element sezonowy o okresie wynoszącym dwanaście miesięcy. Element sezonowy jest stały, więc jako wartość parametru D użyto 1. Dodatkowo trend danych jest niejednoznaczny. Z tego powodu wartość parametru d ustalono jako 0.

Jako pomoc przy dobraniu pozostałych parametrów modelu SARIMA użyte zostały wykresy ACF oraz PACF:



Rysunek 3: Wykres autokorelacji



Rysunek 4: Wykres częściowej autokorelacji

Wykres ACF posiada największą wartość dla wartości 12, 24, ..., co potwierdza hipotezę o sezonowej częstotliwości wynoszącej 12 miesięcy. Wykres przyjmuje wartości ponad poziom istotności dla dwóch pierwszych wartości (pomijając zerową), co powoduje że jako parametr q użyta została wartość 2. Dodatkowo wartość dla elementu 12, czyli S , jest dodatnia. Z tego powodu parametr P będzie przyjmował wartość dodatnią, a parametr Q - wartość 0.

Wykres PACF przyjmuje wartości ponad poziom istotności dla pierwszej wartości (pomijając zerową), co sprawia że jako parametr p użyta została wartość 1.

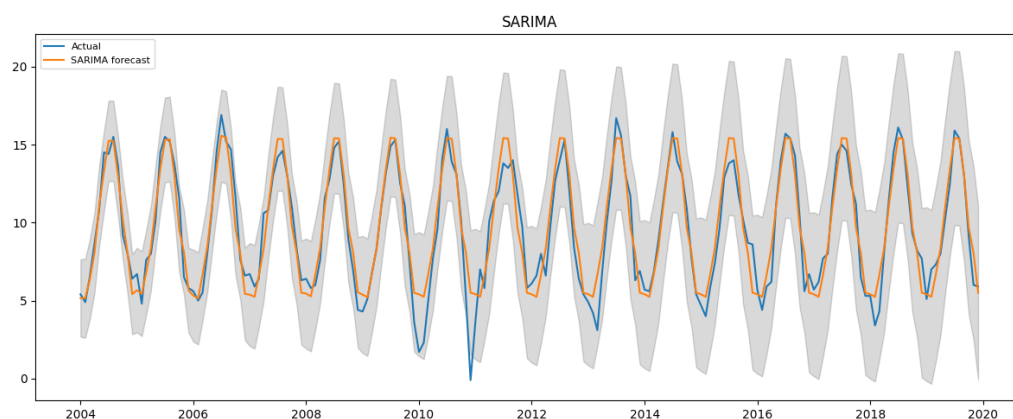
Na podstawie analizy danych model $\text{SARIMA}(1, 0, 2) \times (P, 1, 0)_{12}$. Dokładna wartość P została dobrana przy użyciu metody automatycznego dobierania parametrów modelu udostępnionej przez bibliotekę `pmdarima` i wyniosła 2. Zatem ostateczna postać modelu wyniosła $\text{SARIMA}(1, 0, 2) \times (2, 1, 0)_{12}$.

4.4 ARIMA

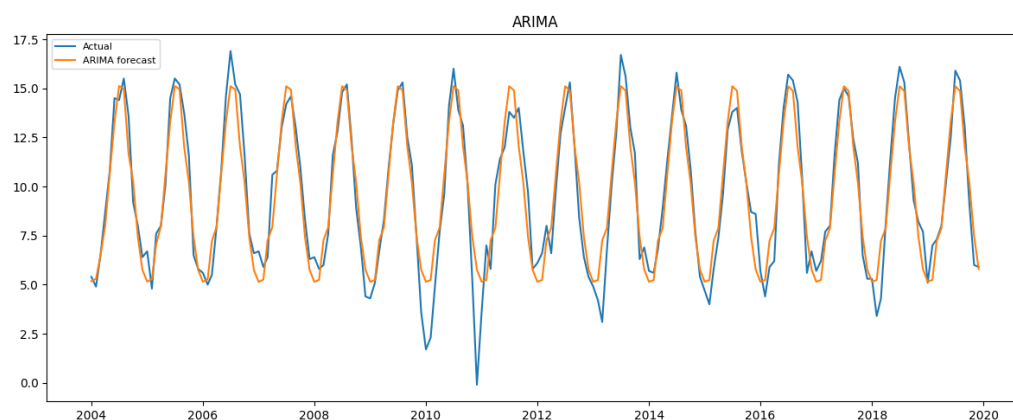
Grupa maszyn **ARIMA** została zaimplementowana zgodnie z opisem, z grupami o n równym odpowiednio 2, 3, 4 i 6. Definicja grupy mówi, że maszyny działają na co n -tym elemencie z ostatnich S . Dlatego też każda maszyna ma postać $\text{ARIMA}(\frac{S}{n}, 0, \frac{S}{n})$, gdzie $d = 0$ wynika z analizy przeprowadzonej podczas konstrukcji maszyny **SARIMA**. Podczas stosowania kombinacji liniowej wyników, z powodu bliskich wartości przewidywanych przez każdą z grup, jako współczynnik dla każdej grupy przyjęto wartość $\frac{1}{4}$.

5 Wyniki

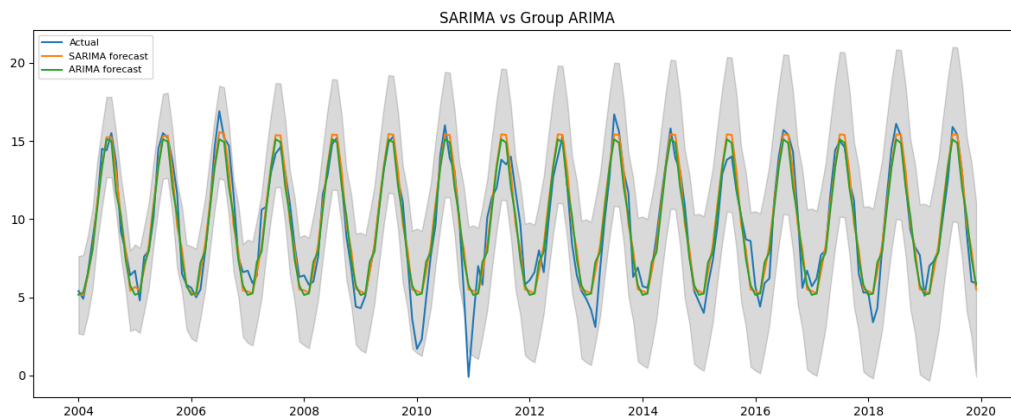
Na kolejnych wykresach szarym kolorem zaznaczona została pewność prognozy na poziomie 95%. Oznacza to prawdopodobieństwo 95%, że prawdziwa wartość będzie zawierać się w tym przedziale. Wyniki prezentują się następująco:



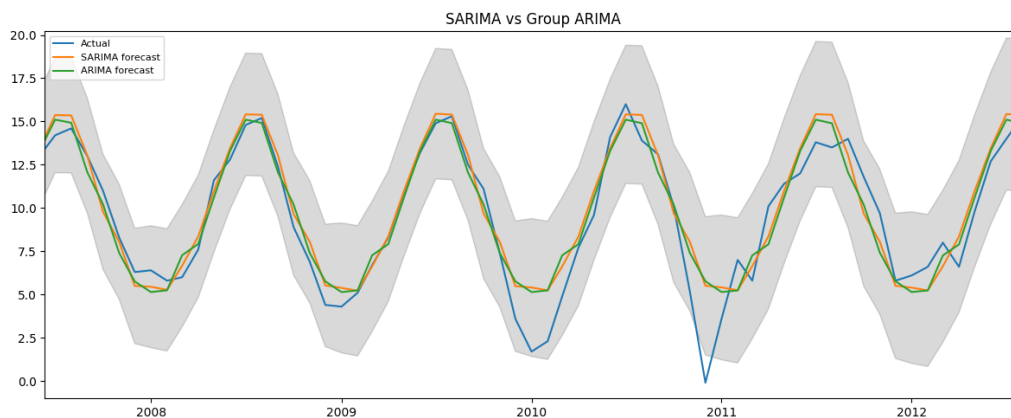
Rysunek 5: Wykres przewidywań modelu **SARIMA** z zaznaczoną pewnością 95% oraz prawdziwych wartości



Rysunek 6: Wykres przewidywań modelu kombinacji liniowej **ARIMA** oraz prawdziwych wartości



Rysunek 7: Wykres przewidywań obu modeli oraz prawdziwych wartości



Rysunek 8: Wykres przewidywań obu modeli oraz prawdziwych wartości przycięty do 5 lat

Model **SARIMA** osiągnął średni błąd kwadratowy równy 1,2023 oraz średni błąd procentowy równy 0,4212. Liniowa kombinacja maszyn **ARIMA** osiągnęła średni błąd kwadratowy równy 1,2148 oraz średni błąd procentowy równy 0,4346.

6 Wnioski

Oba modele osiągnęły porównywalną efektywność oraz nie wykazują opóźnienia w zmianie trendu, co świadczy o świadomości sezonowej obu modeli. Jednocześnie wykresy obu modeli przypominają cykl, który nie wykazuje reakcji na nietypowe wartości, jak w przypadku przełomu 2009/2010 roku.

Przyglądając się wykresom można zauważyć, że model **SARIMA** posiada kształt bliższy wartościom rzeczywistym w okresie zimowym, kiedy wartości temperatury są najniższe i wachają się przez kilka miesięcy, przed ponownym wzrostem we wczesnej wiosnie. Natomiast model **ARIMA** lepiej przystosował się do wartości w okresie po nowym roku, gdzie często widzimy nagły skok, po którym następuje lekkie wygładzenie oraz ponowny wzrost.

Dobre dane charakteryzują się niejednorodnym trendem oraz niską względną różnicą wartości. Może to być jedna z przyczyn obserwowanego zachowania danych.

Literatura

- [1] K. Hatalis, *Tutorial: Multistep Forecasting with Seasonal ARIMA in Python*, Data Science Central, 2018, <https://web.archive.org/web/20201205000439/https://www.datasciencecentral.com/profiles/blogs/tutorial-forecasting-with-seasonal-arima>.