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 modelów, 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 a proces autoregresyjny z rzędem regresji p - `AR(p)` i przedstawiamy następująco:

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

  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:

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

  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 oznaczymy przez S.

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

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

W celu zdefiniowania wzoru modelu `SARIMA` zdefiniujmy operator poprzednika B:

$$By_t = y_{t-1}$$
$$B^i y_t = y_{t-i}$$

Uwzględniając nowe procesy oraz operator B, otrzymujemy ogólny wzór modelu `SARIMA`:

$$\prod_{i=1}^{p}(1-\phi_i B^i)\prod_{i=1}^{P}(1-\Phi_i B^{Si})\prod_{i=1}^{d}(1-B^i)\prod_{i=1}^{D}(1-B^{Si})y_t = \prod_{i=1}^{q}(1+\theta_i B^i)\prod_{i=1}^{Q}(1+\Theta_i B^{Si})\epsilon_t$$

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 składa się z grup maszyn `ARIMA` działających na co n-tym elemencie ciągu z ostatnich S elementów, gdzie S jest częstością maszyny `SARIMA`. Każda z maszyn w grupie operuje na pewnym podciągu ciągu bazowego zawierającego co n-ty element ciągu. Podciągi te są parami rozłączne, oraz ich suma równa jest ciągowi bazowemu. Każdy podciągów w grupie jest równej długości. Z tego powodu narzucamy ograniczenie na n: $n|S$. Maszyny

wewnątrz grupy są uporządkowane według kolejności podciągów, czyli za pierwszą uznajemy maszynę, której podciąg zaczyna się od pierwszego elementu ciągu, drugą - od drugiego, i tak dalej.

Weźmy grupy maszyn spełniające powyższe warunki i ponumerujmy je $1, 2, \ldots, n$. 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 $tmodi$. 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

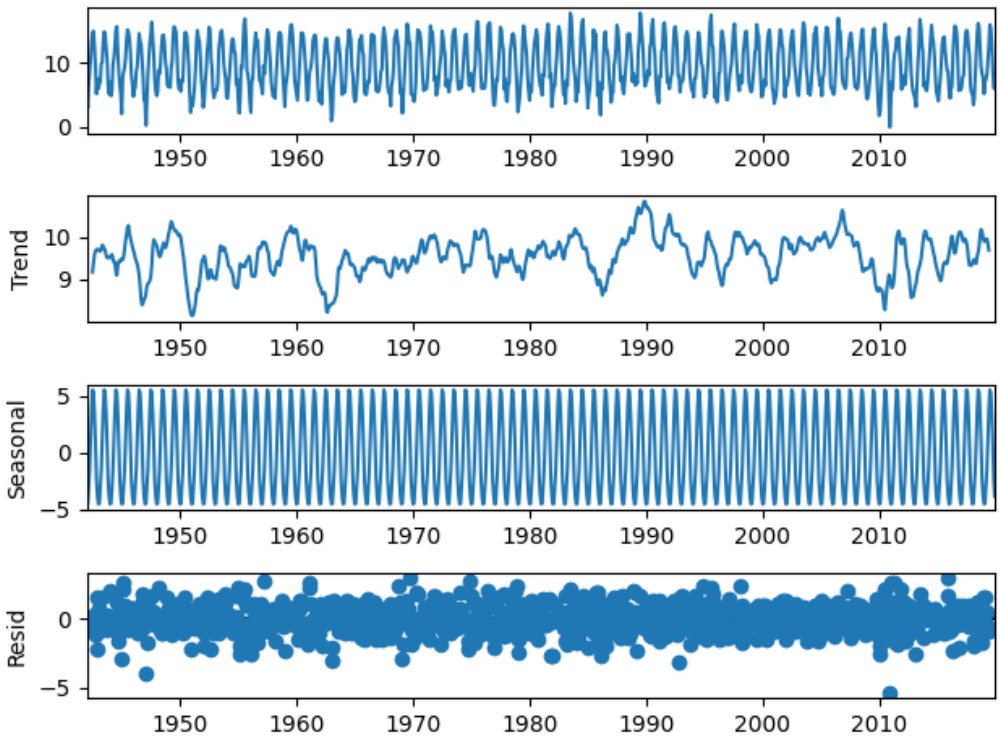
4.1 Biblioteka modeli

Do budowy modeli została użyta biblioteka `statsmodel` zapewniająca implementacje modeli `ARIMA` oraz rozszerzenia `SARIMA`. Dodatkowo biblioteka ta umożliwia automatyczne dobranie odpowiednich współczynników na podstawie danych wejściowych oraz parametrów modelu. Dodatkowo udostępnia metody umożliwiające prognozowanie kolejnych elementów ciągu.

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`.

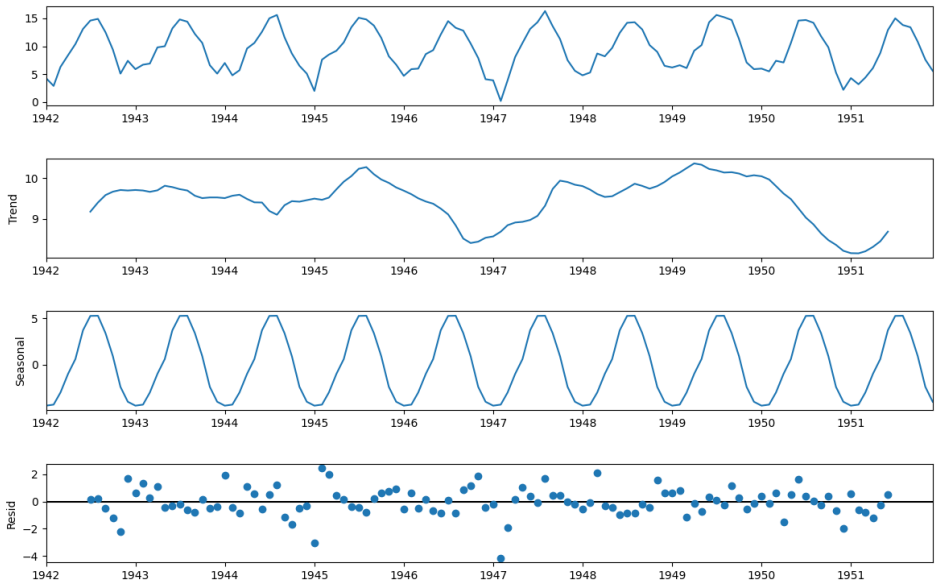
4.2 SARIMA

Przed konstrukcją modelu wykonana została analiza danych za pomocą algorytmu rozkładu sezonowego udostępnionego przez bibliotekę `statsmodels`. Parametry modelu dobrane zostały zgodnie ze wskazówkami przedstawionymi przez Kostas Hatalis Wynik rozkłądu prezentuje się następująco:



Rysunek 1: Rozkład sezonowy danych

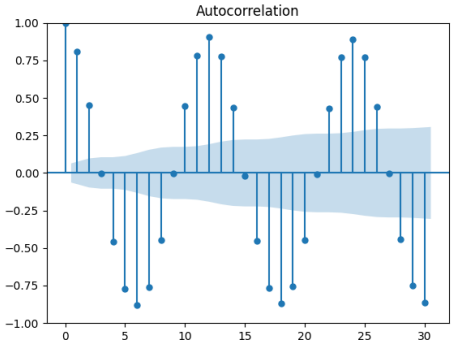
Wykres przedstawia kolejno: wartość, trend, komponent sezonowy oraz elemen losowy każdego elementu ciągu. W celu łatwiejszej analizy skupiono się na pierwszych dziesięciu latach:



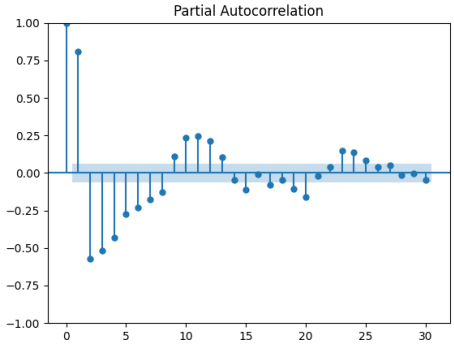
Rysunek 2: Rozkład sezonowy danych z pierwszych 10 lat

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 autokorelacji (ACF) oraz częściowej autokorelacji danych (PACF):



Rysunek 3: Wykres autokorekacji



Rysunek 4: Wykres częściowej autokorekacji

Wykres ACF posiada największą wartość dla wartości $12, 24, \ldots$, co potwierda hipotezę o sezonowej częstości wynoszącej 12 miesięcy. Wykres przyjmuje wartości ponad poziomem 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 poziomem istotności dla pierwszej wartości (pomijając zerową), co sprawia że jako parametr p użyta została wartość 1.

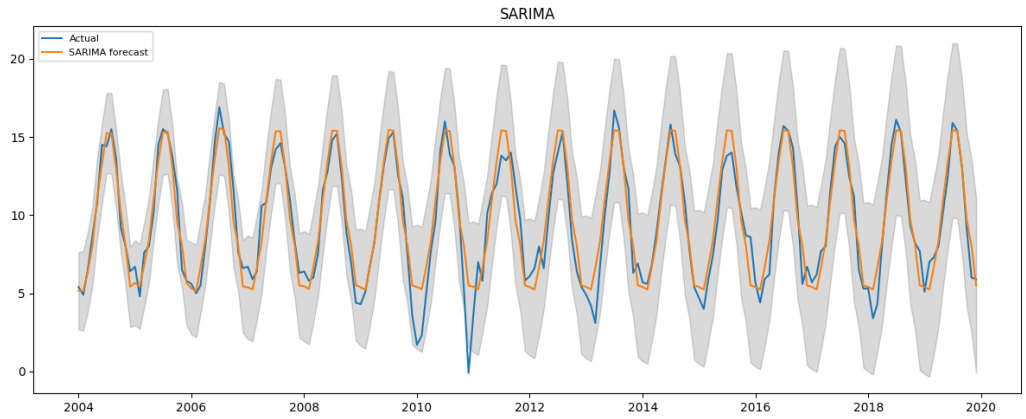
Na podstawie analizy danych model `SARIMA(1, 0, 2) x (P, 1, 0)12`. Dokładna wartość P została dobrana przy użyciu metody automatycznego dobierania parametrów modelu udostępnionej przez bibliotekę `pmdarima` i wyniosła 2. Zatem ostateczna postać modelu wyniosła `SARIMA(1, 0, 2) x (2, 1, 0)12`.

4.3 ARIMA

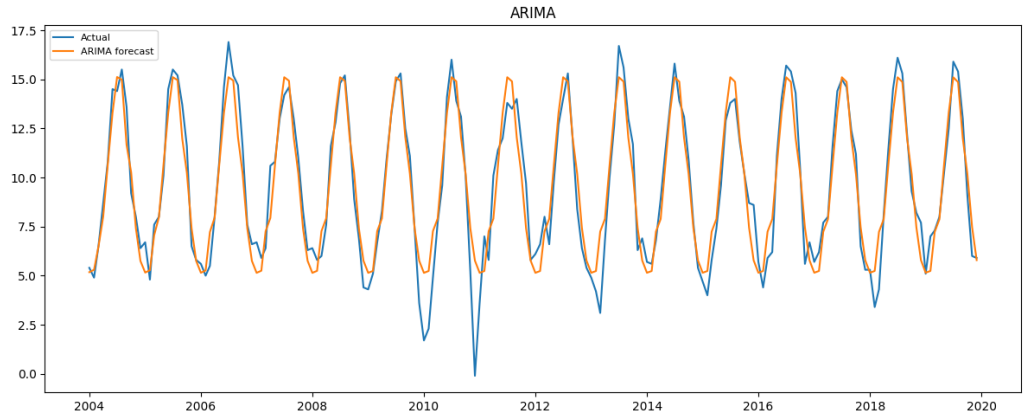
Grupa maszyn `ARIMA` została zaimplementowana zgodnie z opisem, z grupami o n równym odpowiednio $2, 3, 4$ i 6. Definicja grupy mówi, że maszyny działają na co n-tym elemencie z ostatnich S. Dlatego też każda maszyna ma postać `ARIMA(`$\frac{S}{n}$`, 0, `$\frac{S}{n}$`)`, gdzie $d = 0$ wynika z analizy 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

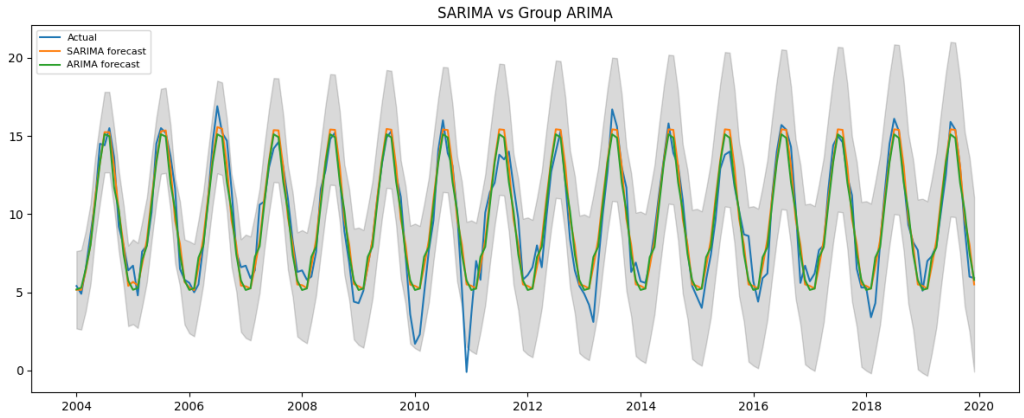
Wyniki prezentują się następująco:



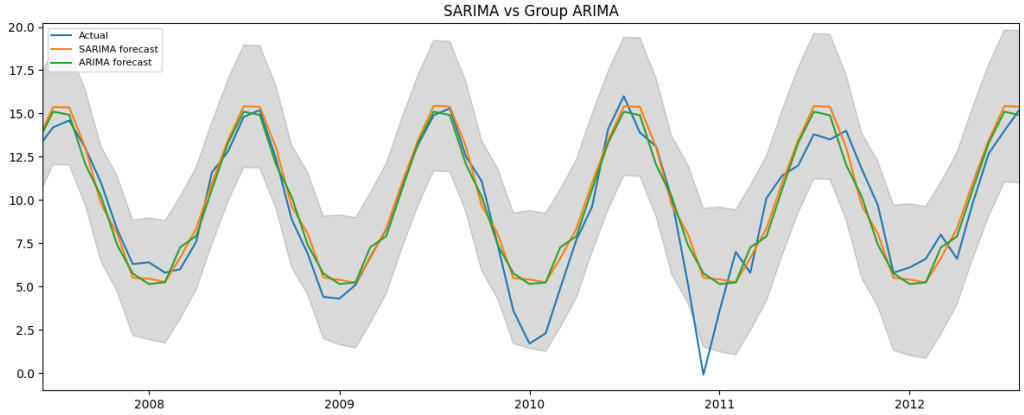
Rysunek 5: Wykres przewidywań modelu `SARIMA` z zaznaczoną pewnością 95% oraz prawdziwych wartości



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 efektywnośc 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 wiośnie. 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.

Dobrane dane charakteryzują się niejednolitym trendem oraz niską względną różnicą wartości. Może to być jedna z przyczym obserwowanego zachowania danych.