**Mateusz Godlewski  
Numer albumu: 85163  
  
  
  
  
  
Oprogramowanie dydaktyczne do testowania działania wybranych metod dla modeli szeregów czasowych na różnych zbiorach danych.  
  
Praca magisterska  
na studiach II-go stopnia  
na kierunku Informatyka**

Opiekun pracy dyplomowej:  
dr inż. Andrzej Kułakowski  
Katedra Systemów Informatycznych

**Oprogramowanie dydaktyczne do testowania działania wybranych metod dla modeli szeregów czasowych na różnych zbiorach danych**

**Streszczenie**

Celem mojej pracy dyplomowej było zaprojektowanie oraz stworzenie oprogramowania dydaktycznego, umożliwiającego testowanie działania wybranych metod predykcji szeregów czasowych. Oprogramowanie to umożliwia wykonanie analizy statystycznej oraz eksploracyjnej na dostarczonym przez użytkownika szeregu czasowym.

W ninejszej pracy zawarty jest opis wybranego oprogramowania służącego do testowania działania metod matematycznych, charakterystyka zagadnień których dotyczy oprogramowanie oraz charakterystyka tworzonego oprogramowania.

Oprogramowanie zostało dokładnie udokumentowane w postaci instrukcji obsługi użytkownika. Natomiast projekt został zilustrowany w postaci wykresów i diagramów, a jego implementacja została przedstawiona poprzez fragmenty kodów źródłowych i zrzuty ekranu.

Słowa kluczowe: Szereg czasowy, statystyka, autoregresja, średnia ruchoma, Python, Flask.

**Didactic software for testing selected methods for time series models on various data sets**

**Summary**

Purpose of my thesis was to design and create didactic software that would enable user to test selected time series prediction methods. This software enables performing statistical and exploratory analysis on the time series given by user.

Present work contains description of selected software that is using for testing mathematical operations, the characteristic of issues related to the software and characteristic of developed program itself.

The software was documented in the form of user manual. Project of software has been detailed illustrated in the form of charts and diagrams, and its implementation has been presented through source code fragments and screenshots.

Keywords: Time series, statistics, autoregression, moving average, Python, Flask.

**SPIS TREŚCI**

1. WSTĘP11

2. CHARAKTERYSTYKA WYBRANEGO OPROGRAMOWANIA DYDAKTYCZNEGO DO TESTOWANIA DZIAŁANIA METOD MATEMATYCZNYCH13

2.1 Oprogramowania matematyczne13

2.2 Opis programu MATLAB00

3. OPRACOWANIE WYBRANYCH ZAGADNIEŃ DO KONSTRUOWANIA OPROGRAMOWANIA TESTUJĄCEGO00

3.1 Test00

3.2 Test00

3.2.1 Test00

4. CHARAKTERYSTYKA WYBRANEGO DO REALIZACJI ZADANIA OPROGRAMOWANIA00

4.1 Test00

4.2 Test00

4.2.1 Test00

5. PROJEKT OPROGRAMOWANIA DYDAKTYCZNEGO DO TESTOWANIA DZIAŁANIA WYBRANYCH METOD DLA MODELI SZEREGÓW CZASOWYCH00

6. IMPLEMENTACJA OPROGRAMOWANIA DYDAKTYCZNEGO00

6.1 Test00

7. TESTOWANIE DZIAŁANIA WYKONANEGO OPROGRAMOWANIA00

7.1 Test00

8. PODSUMOWANIE00

LITERATURA00

Spis rysunków00

Spis listingów00

**Wstęp**

Dynamiczny rozwój szeroko pojętej sztucznej inteligencji w ostatnim dziesięcioleciu sprawił, iż znalazła swoje zastosowanie w prawie każdej sferze życia codziennego. Uczenie maszynowe, sztuczne sieci neuronowe, czy analiza danych, wykorzystywane są w najbardziej popularnych serwisach internetowych czy aplikacjach mobilnych. Rozpoznawanie obrazów, rozpoznawanie mowy, eksploracja danych czy prognozowanie przyszłości, to jedynie kilka z wielu funkcjonalności dostarczanych przez algorytmy sztucznej inteligencji.

Celem mojej pracy dyplomowej było stworzenie oprogramowania w formie aplikacji internetowej, umożliwiającego testowanie wybranych metod prognozy szeregów czasowych, na różnych zbiorach danych. Użytkownik ma możliwość wykonania analizy statystycznej, na dostarczonym przez niego zbiorze danych oraz wizualizacji zawartych w nim danych w postaci wykresów. Główną funkcjonalnością stworzonego oprogramowania, jest możliwość testowania metod predykcji przyszłych wartości szeregów czasowych, za pomocą Autoregresji lub Autoregresji ze zintegrowaną średnią ruchomą.

Dokument ten składa się z ośmiu rozdziałów przedstawiających najważniejsze aspekty pracy. Na wstępie przedstawiony został cel pracy oraz krótki opis poszczególnych rozdziałów. W rozdziale drugim pt. "Charakterystyka wybranego oprogramowania dydaktycznego do testowania działania metod matematycznych." przedstawiona została charakterystyka najpopularniejszego oprogramowania które jest wykorzystywane przy testowaniu metod matematycznych. Następny rozdział zatytułowany "Opracowanie wybranych zagadnień do konstruowania oprogramowania testującego." zawiera zagadnienia oraz definicje pojęć, które najczęściej występują w stworzonym oprogramowaniu. Rozdział numer cztery "Charakterystyka wybranego do realizacji zadania oprogramowania." Zostały zaprezentowane narzędzia oraz technologie wykorzystane do implementacji systemu. W kolejnym rozdziale pt. „Projekt oprogramowania dydaktycznego do testowania działania wybranych metod dla modeli szeregów czasowych.” opisany jest projekt architektury stworzonego oprogramowania w postaci wykresów i diagramów. Następnie w rozdziale szóstym „Implementacja oprogramowania dydaktycznego.” znajduje się przedstawienie zaimplementowanych funkcjonalności systemu, wraz z kodami źródłowymi oraz zrzutami ekranu stworzonego rozwiązania. Rozdział siódmy "Testowanie działania wykonanego oprogramowania" poświęcony został opisowi testów stworzonego systemu. Ostatni, ósmy rozdział zawiera podsumowanie pracy.

**2. CHARAKTERYSTYKA WYBRANEGO OPROGRAMOWANIA DYDAKTYCZNEGO DO TESTOWANIA DZIAŁANIA METOD MATEMATYCZNYCH**

**2.1 Oprogramowania matematyczne**

Wsród najpopularniejszych rozwiązań, dostępnych na rynku są:

* MATLAB
* Magnolia
* Maple
* Scilab

**2.2 Opis programu MATLAB**

…

**2.3 Przykłady wykorzystania programu MATLAB**

…

**3. OPRACOWANIE WYBRANYCH ZAGADNIEŃ DO KONSTRUOWANIA OPROGRAMOWANIA TESTUJĄCEGO**

W stworzonym oprogramowaniu pojawiają się pojęcia z dziedzin matematyki oraz statystyki. Aby w pełni zrozumieć logikę aplikacji, oraz operacje, które są wykonywane przez oprogramowanie należy zapoznać się z definicjami jak:

* Szereg czasowy
* Analiza statystyczna szeregów czasowych
* Metody predykcji szeregów czasowych

**3.1 Szereg czasowy**

Szereg czasowy to seria punktów danych zindeksowanych w porządku czasowym. Najczęściej jest on sekwencją wykonywaną w kolejnych, równych odstępach czasu. Zatem jest to sekwencja danych w czasie dyskretnym. Przykładami szeregów czasowych są wartości akcji na dziełdzie, temperatura powietrza, monitorowanie ciśnienia krwi itd. Szeregi czasowe są bardzo często przedstawiane za pomocą czasowego wykresu liniowego. Szeregi czasowe są najczęściej wykorzystywane w statystyce, przetwarzaniu sygnałów, rozpoznawaniu wzorców, ekonometrii, finansach matematycznych, prognozowaniu pogody, przewidywaniu trzęsień ziemi, inżynierii sterowania [1].

****  
*Rysunek 3.X Wartości akcji firmy Amazon w latach 1997 – 2020*

Do oznaczania ciągów czasowych stosowane są różne notacje. Najczęściej ciąg czasowy **X** indeksowany liczbami naturalnymi jest zapisywany jako [1]:

**X = {X1, X2, …}**

Szeregi czasowe mogą (lecz nie muszą) posiadać własności takie jak [2]:

* **Tendencja rozwojowa (trend)** - czyli długookresowe, systematyczne i jednokierunkowe zmiany wielkości zjawiska. Są wywołane przez stałe przyczyny działające w całym okresie.

  
*Rysunek 3.X Trend szeregu czasowego (niebieska linia)[A]*

* **Wahania okresowe (wahania sezonowe)** - są to rytmiczne, powtarzające się zmiany poziomu zjawiska, które są wywoływane przez czynniki, np. przyrodnicze. Długość cyklu wahań sezonowych, to liczba okresów składających się na pełen cykl wahań okresowych



*Rysunek 3.X Szereg czasowy z widoczną sezonowością [B]*

* **Wahania cyklu koniunkturalnego (wahania cykliczne)** - to oscylacje, które wymagają wieloletnich obserwacji, posiadające nieregularny charakter i różną długość cyklu. Wyróżnia się okresy recesji, depresji, ożywienia i dobrobyt. Brakuje ugruntowanej statystycznej metody ich analizy oraz ich przebieg jest trudny do przewidzenia. Są to to systemowe, falowe wahania rozwoju gospodarki obserwowane w okresach dłuższych od roku i ich analiza wymaga wieloletnich obserwacji.



*Rysunek 3.X Szereg czasowy z widocznymi wahaniami cyklicznymi [C]*

* **Wahania przypadkowe** - czyli nieregularne odchylenia wielkości zjawiska od poziomu, którego spodziewalibyśmy się na podstawie oddziaływania innych czynników (trendu, wahań sezonowych i cyklicznych). Ich natura jest trudna do zdefiniowania. Mogą być wywoływane przez przypadkowe różnokierunkowe działania czynników, które nie są uwzględnione w analizie.



*Rysunek 3.X Szereg czasowy z widocznym losowym odchyleniem [D]*

**3.2 Analiza statystyczna szeregów czasowych**

Analiza statystyczna to proces zbierania i analizowania danych w celu zidentyfikowania wzorców i trendów. Jest to metoda wykorzystująca metody matematyczne w celu usunięcia wszelkich uprzedzeń podczas przeglądania informacji. Można ją również traktować jako narzędzie, które może wpływać na podejmowanie decyzji [3].

W większości przypadków (zestawów danych), podczas przeprowadzania analizy statystycznej można wyodrębnić pięć poniższych kroków [3]:

1. Opisanie charakteru analizowanych danych.
2. Zbadanie relacji danych do bazowej populacji.
3. Utworzenie modelu, aby podsumować relacje między danymi a populacją bazową.
4. Udowodnienie (lub obalenie) słuszność modelu.
5. Wykorzystanie wyników do analiz predykcyjnych, aby przewidywać przyszłe trendy.

W przypadku szeregów czasowych, niemożliwe jest zatosowanie typowej analizy statystycznej, albowiem nie możliwe jest wykonanie zbadanie relacji pomiędzy danymi. Szereg czasowy posiada tylko jeden rodzaj zmiennych, których wartość jest zmienna w dziedzinie czasu.

Na szeregach czasowych, w ramach analizy statystycznej, można wykonać takie operacje jak:

* Preprocessing danych

Proces polegający na wstępnej obróbce danych. W tym procesie, na szeregu czasowym, wykonywane są metody:

* + Walidacji danych - sprawdzenie poprawności typów danych. Najczęściej polega na sprawdzeniu czy wartości szeregu czasowego to liczby całkowite, lub zmiennoprzecinkowe.
  + Sprawdzenie komplementarości danych – sprawdzenie, czy szereg nie posiada brakujących wartości. Sprawdzenie to tyczy się zarówno wartości szeregów, jak i znaczników czasowych w dziedzinie czasu. Jeśli jedna z próbek (pomiarów) nie posiada wartości, jako wartość tej próbki przyjmowana jest uśredniona wartość sąsiednich pomiarów. Jeśli natomiast brakuje znacznika czasowego w próbce, jest ona usuwana z szeregu czasowego.
* Statystyka opisowa

Polega na wykonaniu różnego rodzaju wielkości obliczane na podstawie uzyskanych danych. Interpretacja wartości tych miar dostarcza informacji na temat charakteru rozkładu cechy. Wykonywane są przykładowo takie obliczenia jak:

* Wyznaczenie minimalnej i maksymalnej wielkości szeregu czasowego.
* Obliczenie wartości średniej arytmetycznej – obliczenie średniej arytmetycznej wszystkich wartości szeregu czasowego zgodnie ze wzorem:
* Obliczenie wartości odchylenia standardowego - obliczenie odchylenia standardowego wartości szeregu czasowego zgodnie ze wzorem:
* Wyznaczenie mediany – aby obliczyć medianę, należy posortować wszystkie wartości szeregu czasowego w sposób niemalejący. Jeśli liczba obserwacji jest liczbą nieparzystą, mediana jest wartością znajdującą się „po środku”, czyli próbka o numerze **,** natomiast w przeciwnym wypadku, gdy liczba próbek jest liczbą parzystą, mediana jest średnią arytmetyczną między dwoma środkowymi obserwacjami, czyli obserwacją oraz .
* Wyznaczenie kwantyla – wyznaczenie kwantyla rzędu **q (0 < q < 1)**, który jest liczbą **Q**, której 100% elementów danej zbiorowości statystycznej ma wartość nie większą niż **Q**.
* Wyznaczenie interkwartyla – obliczenie różnicy między trzecim a pierwszym kwartylem. Pierwszy kwartyl (dolny kwartyl), to kwantyl rzędu , natomiast trzeci kwartyl, to kwantyl trzeciego rzędu (kwartyl górny) – kwantyl rzędu .
* Identyfikacja punktów oddalonych

Polega na wyznaczeniu poszczególnych punktów w zbiorze danych (a w tym przypadku szeregu czasowym), których wartość jest mniejsza niż różnica pierwszego kwartylu i ilorazu liczby 1,5 oraz rozstępu międzykwartylowego lub większa, niż suma trzeciego kwartylu i ilorazu liczby 1,5 oraz rozstępu międzykwartylowego.

* Autokorelacja

Współczynnik korelacji między dwiema wartościami szeregu czasowego nazywany jest współczynnikiem autokorelacji. Współczynnik ten dla szeregu czasowego , można przedstawić jako:

Wartość **k**, zwana opóźnieniem, przedstawia przerwę czasową. Autokorelacja z opóźnieniem (k) równym 1, jest korelacją pomiędzy wartościami, których różni jeden okres czasu (jedną jednostkę czasu w szeregu czasowym). Autokorelacja z opóźnieniem **k**, jest korelacją między wartościami, które są oddalone od siebie o **k** okresów czasu. Wyznaczenie współczynnika autokorelacji jest sposobem pomiaru liniowej zależności między obserwacjami w czasie **t**, a obserwacjami z poprzednich czasów [4]. Współczynnik autokorelacji można obliczyć posługując się wzorem [5]:

- to współczynnik korelacji, - opóźnienie, - liczba obserwacji,  
 - kolejne obserwacje, kolejne obserwacje opóźnione o k, - średnia obserwacji.

Jeśli współczynnik autokorelacji jest bliski wartości 1, między obserwowanymi wartościami zachodzi silna korelacja dodatnia, jeśl bliski wartości -1, silna korelacja ujemna, natomiast gdy współczynnik ten jest bliski zeru, brak koleracji, lub jest ona niewielka

****

*Rysunek 3.X Wykres autokorelacji szeregu czasowego[E]*

**3.3 Metody predykcji szeregów czasowych**

Predykcja (prognoza) szeregów czasowych polega na przewidywaniu przyszłych wartości szeregów czasowych, na podstawie danych historycznych. Prognozowanie szeregów czasowych to czasami tylko analiza ekspertów badających dziedzinę i oferujących swoje prognozy. Jednak w wielu nowoczesnych aplikacjach prognozowanie szeregów czasowych wykorzystuje technologie komputerowe, w tym

* Uczenie maszynowe
* Sztuczne sieci neuronowe
* Maszyna wektorów nośnych
* Logika rozmyta
* Proces Gausowski
* Ukryte modele Markova

Jednak w ramach stworzonego oprogramowania, do predykcji przyszłych wartości zostały wykorzystane zostały modele statystyczne takie jak model autoregresyjny, oraz odel autoregresyjny zintegrowany ze średnią ruchomą.

**3.3.1 Model autoregresyjny (AR)**

Model autoregresyjny (*ang. autoregressive model, AR model*) to parametryczny model szeregu czasowego, który wykorzystywany jest do modelowania i predykcji zjawisk naturalnych, opisanych za pomocą szeregów czasowych. Model autoregresyjny jest jedną z formuł predykcji liniowej – takie formuły dokonują predykcji wyjścia układu w oparciu o wartości na wejściu modelu (danych z przeszłości).

Model autoregresyjny można przedstawić za pomocą notacji **AR(p)**, która wskazuje na to, iż chodzi o model autoregresyjny rzędu **p**. Notacja ta prezentuje się następująco [6]:

- parametry modelu  
 - stała (dla uproszczenia formuły, często pomijana)  
 – szum biały

Modele autoregresyjne są niezwykle elastyczne w obsłudze szerokiej gamy szeregów czasowych o różnych właściwościach. Model autoregresyjny określa, że ​​zmienna wyjściowa zależy liniowo od swoich wcześniejszych wartości i od składnika stochastycznego zatem model ma postać stochastycznego równania różnicowego (lub relacji rekurencyjnej, której nie należy mylić z równaniem różniczkowym) [6].

Najprostszym przykładem jest model autoregresyjny pierwszego rzędu **AR(1)** (inaczej nazywany liniowym procesem Markova), w którym wartość w danej chwili zależy wyłącznie od wartości w chwili poprzedniej i szumu. [7].

Podstawiając trzy kolejne wyrazy.

Bardzo ważną kwestią związaną z zastosowaniem modeli autoregresyjnych jest ustalenie odpowiedniego rzędu autoregresji. Do najbardziej popularnych metod ustalania rzędu autoregresji należą kryteria informacyjne takie jak:

* Kryterium Informacyjne Akaike’a (AIC)

Według tego kryterium, najskuteczniejszy jest model, którego najmniejsza jest wartość **AIC**, gdzie formuła **AIC** prezentuje się następująco [8]:

* - estymowane prawdopodobieństwo uzyskania takiej właśnie wartości obserwacji **j** jaka była naprawdę uzyskana
* - liczba parametrów modelu
* Bayesowskie kryterium informacyjne Schwarza (BIC)

W przypadku modelu Bayesa, wybiera się model, którego najmniejsza jest wartość [9]

* - liczba obserwacji
* - liczba parametrów modelu
* - funkcja wiarygodności dla oszacowanego wektora parametrów
* Kryterium informacyjne Hannana-Quinna (HQC)

Według kryterium Hannana-Quinna, najbardziej precyzyjny jest model którego najniższa jest wartość [10]:

* - liczba parametrów
* - liczba obserwacji
* – suma kwadratów reszt wynikająca z przeprowadzonej regresji liniowej bądź też innego modelu

**3.3.2 Średnia ruchoma (MA)**

Kolejną z metod prognozowania przyszłych wartości szeregów czasowych, jest średnia ruchoma (*ang. moving average*). Jest to statystyczna metoda, za pomocą której można wyznaczyć wartości poszczególnych punktów poprzez obliczenie średnich z różnych podzbiorów pełnego zestawu danych. Nazywana jest również średnią kroczącą (*ang. moving mean*). Wśród średnich ruchomych wyróżnia się takie odmiary jak:

* Prosta średnia ruchoma
* Ważona średnia ruchoma
* Wykładnicza średnia ruchoma.
* Skumulowana średnia ruchoma
* Trójkątna średnia ruchoma

**LITERATURA**

**Bibliografia literatury**

[1] „Time series”, <https://en.wikipedia.org/wiki/Time_series> [przeglądane dnia 01.06.2021r]

[2] „Szereg czasowy”, <https://mfiles.pl/pl/index.php/Szereg_czasowy> [przeglądane dnia 01.06.2021r]

[3] „What Is Statistical Analysis?”, <https://www.businessnewsdaily.com/6000-statistical-analysis.html> [przeglądane dnia 01.06.2021r]

[4] „Autoregressive Models”, <https://online.stat.psu.edu/stat501/lesson/14/14.1> [przeglądane dnia 07.06.2021r]

[5] „Wzór na autokorelację”, <https://www.naukowiec.org/wzory/statystyka/autokorelacja_411.html> [przeglądane dnia 07.06.2021r]

[6] „Model AR”, [https://en.wikipedia.org/wiki/Autoregressive\_model](https://en.wikipedia.org/wiki/Autoregressive_model%20) [przeglądane dnia 07.06.2021r]

[7] „Model autoregresyjny”, <https://brain.fuw.edu.pl/edu/index.php/Model_autoregresyjny_(AR)> [przeglądane dnia 07.06.2021r]

[8] „Kryterium informacyjne Akaikego”, <https://pl.wikipedia.org/wiki/Kryterium_informacyjne_Akaikego> [przeglądane dnia 07.06.2021r]

[9] „Bayesian information criterion”, <https://en.wikipedia.org/wiki/Bayesian_information_criterion> [przeglądane dnia 07.06.2021r]

[10] „Kryterium informacyjne Hannana-Quinna”, <https://pl.wikipedia.org/wiki/Kryterium_informacyjne_Hannana-Quinna> [przeglądane dnia 07.06.2021r]

**Bibliografia rysunków**

[A] <https://i.stack.imgur.com/eFUp8.jpg> [przeglądane dnia 01.06.2021r]

[B] <https://robjhyndman.com/hyndsight/2011-12-14-cyclicts_files/figure-html/unnamed-chunk-2-1.png> [przeglądane dnia 01.06.2021r]

[C] <https://robjhyndman.com/hyndsight/2011-12-14-cyclicts_files/figure-html/unnamed-chunk-1-1.png> [przeglądane dnia 01.06.2021r]

[D] <https://support.minitab.com/en-us/minitab-express/1/tsplot_weight_with_outlier_single_y_simple_4col.xml_Graph_cmd1o1.png> [przeglądane dnia 01.06.2021r]

[E] <https://www.statsmodels.org/stable/plots/graphics_tsa_plot_acf.hires.png> [przeglądane dnia 07.06.2021r]