**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 MATLAB14

3. OPRACOWANIE WYBRANYCH ZAGADNIEŃ DO KONSTRUOWANIA OPROGRAMOWANIA TESTUJĄCEGO17

3.1 Szereg czasowy17

3.2 Analiza statyczna szeregów czasowych20

3.3 Metody predykcji szeregów czasowych23

3.3.1 Model autoregresyjny (AR)23

3.3.2 Średnia ruchoma (MA)26

3.3.3 Model autoregresyjny zintegrowany ze średnią ruchomą (ARIMA)27

4. CHARAKTERYSTYKA WYBRANEGO DO REALIZACJI ZADANIA OPROGRAMOWANIA29

4.1 Opis funkcjonalności oprogramowania29

4.4.1 Przesyłanie pliku30

4.4.2 Analiza statystyczna oraz wizualizacja szeregu czasowego 31

4.4.3 Prognozowanie szeregów czasowych 33

4.2 Wykorzystane technologie35

5. PROJEKT OPROGRAMOWANIA DYDAKTYCZNEGO DO TESTOWANIA DZIAŁANIA WYBRANYCH METOD DLA MODELI SZEREGÓW CZASOWYCH36

6. IMPLEMENTACJA OPROGRAMOWANIA DYDAKTYCZNEGO00

7. TESTOWANIE DZIAŁANIA WYKONANEGO OPROGRAMOWANIA00

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 dydaktycznego 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 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 omawianym oprogramowaniu. W rozdziale numer cztery "Charakterystyka wybranego do realizacji zadania oprogramowania" Zostały zaprezentowane narzędzia oraz technologie wykorzystane do implementacji systemu, jak i opis funkcjonalności oprogramowania. W kolejnym rozdziale pt. „Projekt oprogramowania dydaktycznego do testowania działania wybranych metod dla modeli szeregów czasowych” zawarty 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. Rozdział siódmy "Testowanie działania wykonanego oprogramowania" poświęcony został opisowi testów stworzonego oprogramowania. Ostatni, ósmy rozdział zawiera podsumowanie pracy.

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

Matematyka, jako przysłowiowa „królowa nauk”, bez wątpienia w ogromnej mierze przyczyniła się do postępu cywilizacyjnego, a zwłaszcza postępu technologicznego, z którym mamy obecnie do czynienia. Znajduje ona swoje zastosowanie w medycynie, sporcie, inżynierii, ekonomii, muzyce, sztuce oraz w wielu aspektach codziennego życia. Obecnie w celach naukowych lub dydaktycznych, wykorzystuje się tzw. „oprogramowania matematyczne”.

**2.1 Oprogramowania matematyczne**

Oprogramowanie matematyczne, to oprogramowanie wykorzystywane do modelowania, analizy oraz obliczeń numerycznych, symbolicznych lub geometrycznych. Jako przykład najprostszego z nich, można przedstawić oprogramowanie prostego kalkulatora [1]. W tym rozdziale skupię się na bardziej zaawansowanych oprogramowaniach, które umożliwiają modelowanie oraz wykonywanie bardziej zaawansowanych operacji. Wśród najpopularniejszych rozwiązań, dostępnych na rynku wyróżniają się:

* MATLAB;
* Magnolia;
* Maple;
* Scilab.

Na potrzeby pracy, zostanie omówione najbardziej popularne oprogramowanie matematyczne, jakim jest MATLAB oraz zostaną przedstawione przykłady zastosowania tego rozwiązania na praktycznych przykładach.



*Rysunek 2.1 Logo programu MATLAB []*

**2.2 Opis programu MATLAB**

MATLAB (*ang. MATrix LABoratory*) to interaktywne programistyczne środowisko przeznaczone do rozwijania algorytmów, wizualizacji i analizy danych. Rozwiązanie to pozwala również na prowadzenie obliczeń naukowych oraz inżynierskich. MATLAB oferuje możliwość szybszego rozwiązywania problemów, niż przy wykorzystaniu języków programowania [2].

Autorzy MATLAB, bardzo duży nacisk postawili na automatyzację rutynowych procesów, na każdym etapie rozwiązywania danego zadania. Poczynając od gromadzenia danych, poprzez ich analizę i wizualizację, aż do generacji końcowego wyniku w postaci raportu. Rozpoczynając pracę z programem, w bardzo prosty i intuicyjny sposób można zaimportować dane z pliku, bazy danych czy specyficznego sprzętu pomiarowego. Uprzednio zaimportowane dane można zwizualizować za pomocą wielu dostępnych funkcji graficznych. W analizie danych mogą pomóc liczne gotowe procedury obliczeniowe. Na zakończenie pracy, środowisko pracy generuje raport z obliczeń [2].



*Rysunek 2.2 Schemat pracy w programie MATLAB []*

Środowisko MATLAB posiada wiele zalet, do najważniejszych udogodnień tego rozwiązania zalicza się [2]:

* zintegrowany język programowania wysokiego poziomu;
* narzędzia do importu danych;
* możliwość integracji własnego kodu języka C oraz Java;
* interaktywne narzędzia do wizualizacji i eksploracji danych;
* duża ilość wbudowanych funkcji obliczeniowych.

MATLAB posiada swój własny język programowania, umożliwiający pisanie w pełni funkcjonalnych programów. Korzystając z tego języka można w szybszy sposób tworzyć algorytmy, niż w przypadku innych języków programowania. Dzięki temu użytkownik nie jest obciążony wykonywaniem niskopoziomowych operacji, takich jak deklaracja zmiennych ani określania ich typów. Język ten udostępnia możliwość operacji na macierzach, co eliminuje w wielu przypadkach konieczność tworzenia zagnieżdżonych pętli „for”.

columns = 10000;

rows = 1;

bins = columns/100;

rng(now);

list = 100\*rand(rows,columns);

histogram(list,bins)

*Listing 1. Przykładowy fragment kodu w języku MATLAB*

Powyższy fragment kodu tworzy histogram stu losowo wygenerowanych wartości. Rezultaty uruchomienia powyższego kodu prezentują się następująco.



*Rysunek 2.3 Histogram wygenerowany za pomocą programu MATLAB*

Środowisko MATLAB posiada kilka bardzo przydatnych, wbudowanych narzędzi, które usprawniają pracę z oprogramowaniem, takich jak:

* MATLAB Editor – służy do edycji i debugowania kodu.
* Code Analyzer – sprawdza kod pod kątem występujących błędów i rekomenduje zmiany w kodzie, mające na celu uzyskanie najwyższej jakości.
* MATLAB Profiler – oblicza czas, jaki jest wymagany na wykonanie poszczególnych fragmentów kodu.

Omawiane rozwiązanie umożliwia wyeksportowanie wyników pracy w postaci pojedynczych wykresów, jak i w postaci kompletnego raportu z przeprowadzonych obliczeń. Program obsługuje eksport wyników w formatach takich jak: HTML, Word, LaTEX czy PDF.

MATLAB jest oprogramowaniem posiadającym szerokie zastosowanie. Oto niektóre z przykładowych obszarów zastosowania:

* Przetwarzanie sygnałów;
* Przetwarzanie obrazów;
* Telekomunikacja;
* Projektowanie układów sterowania;
* Matematyka finansowa.



*Rysunek 2.4 Interfejs 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 giełdzie, 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 [3].

****  
*Rysunek 3.1 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 [3]:

**X = {X1, X2, …}**

Szeregi czasowe mogą (lecz nie muszą) posiadać własności takie jak [4]:

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

  
*Rysunek 3.2 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 3.3 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 3.4 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 3.5 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 [5].

W większości przypadków (zestawów danych), podczas przeprowadzania analizy statystycznej można wyodrębnić pięć poniższych kroków [5]:

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 zastowanie 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 komplementarności 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 „pośrodku”, 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 [6]. Współczynnik autokorelacji można obliczyć posługując się wzorem [7]:

- 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 korelacji, lub jest ona niewielka.

****

*Rysunek 3.6 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;
* Ukryte modele Markova.

Jednak w ramach stworzonego oprogramowania, do predykcji przyszłych wartości zostały wykorzystane zostały modele statystyczne takie jak model autoregresyjny, oraz model 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 [8]:

gdzie:

* - wartość szeregu czasowego dla obserwacji *t*
* - 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) [8].

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. [9].

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 [10]:

gdzie:

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

W przypadku tego modelu wybiera się model, którego najmniejsza jest wartość [11]:

gdzie:

* - 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 [12]:

gdzie:

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

****

*Rysunek 3.7 Przykładowa prognoza szeregu czasowego z wykorzystaniem AR.*

*Pomarańczowa linia oznacza prawdziwe wartości, niebieska natomiast wartości estymowane przez model [E]*

**3.3.2 Średnia ruchoma (MA)**

Kolejną z metod, która jest ważnym elementem procesu predykcji 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*). Średnią ruchomą stosuje się głównie w celu wygładzania krótkoterminowych wahań. Wśród średnich ruchomych wyróżnia się takie odmiany jak:

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

Biblioteki wykorzystane w tworzeniu oprogramowania wykorzystują prostą średnią ruchomą. Prosta średnia ruchoma SMA (*ang. simple moving average*) to zwykła średnia arytmetyczna wartości z ostatnich **n** okresów. oznacza ostatnią wartość [13].

gdzie:

* – liczba obserwacji
* - poszczególna obserwacja



*Rysunek 3.8 Wygładzenie przebiegu z wykorzystaniem średniej ruchomej [E]*

**3.3.3 Model autoregresyjny zintegrowany ze średnią ruchomą (ARIMA)**

Model ARIMA (*ang. autoregressive integrated moving average*) to model statystyczny wykorzystując szereg czasowy dla lepszego zrozumienia danych, oraz prognozowania jego przyszłych wartości. Model ten jest połączeniem trzech elementów, takich jak [14]:

* Autoregresja (AR) – odnosi się do modelu, który pokazuje zmieniającą się zmienną, ulegającą regresji według własnych, opóźnionych (poprzednich) wartości.
* Zintegrowanie (I) – reprezentuje szeregi czasowe, które zostały zróżnicowane tak, aby stały się one stacjonarne.
* Średnia ruchoma (MA) – zawiera zależność między obserwacją a błędem modelu średniej ruchomej zastosowanego do opóźnionych obserwacji.

Każdy z elementów w modelu ARIMA funkcjonuje jako parametr ze standardową notacją. W przypadku tego modelu są to parametry p, q i d. Każdy z parametrów jest wielkością całkowitą, która oznacza [14]:

* **p** – to liczba określająca rząd predykcji w komponencie autoregresji. Wartość tego parametru definiuje ile poprzednich wartości w szeregu czasowym będzie brane pod uwagę w predykcji kolejnych wartości.
* **q** - określa stopień zróżnicowania, wskazuje, ile razy wartości szeregu czasowego zostają opóźnione (odjęte) w celu przekształcenia szeregu czasowy w stacjonarny.
* **d** - oznacza liczbę błędów prognozy w modelu i jest również określany jako rozmiar okna średniej ruchomej.

Jeżeli wartość któregoś z parametrów zostanie ustawiona na 0, będzie to oznaczało, iż komponent ten nie zostanie wykorzystany. W ten sposób, model ARIMA może zostać przekształcony w dowolną kombinację komponentów modelu.

Komponenty modelu ARIMA, takie jak autoregresja oraz średnia ruchoma zostały omówione w poprzednich podrozdziałach. Nie został natomiast wspomniany komponent integracji szeregów czasowych. Do prognozy szeregów czasowych wykorzystywane są również modele takie jak ARMA. Lecz ich zastosowanie ma jedynie miejsce w przypadku, gdy szereg czasowy jest stacjonarny. Model ARIMA, który jest rozszerzeniem modelu ARMA o dodatkowy komponent integracji, może zostać zastosowany w przypadku szeregów niestacjonarnych, pod warunkiem, iż szereg czasowy na którym operuje model, może zostać sprowadzony do postaci stacjonarnej.

Szereg czasowy jest wtedy stacjonarny, kiedy ma stały poziom w czasie (średnią) oraz odchylenie standardowe od średniego poziomu (wariancje). Stacjonarność szeregu można sprawdzić za pomocą wykresu autokorelacji: szeregi niestacjonarne mają około 6 opóźnień istotnie różnych od zera. Jedną z metod sprowadzenia szeregu czasowego do stacjonarności jest jego zróżnicowanie (*ang. differencing*). Zróżnicowanie polega na obliczeniu różnic pomiędzy kolejnymi wartościami. Pierwsze z nich można oznaczyć jako:

Różnice między tymi wartościami będą reprezentować spadki i wzrosty wartości w szeregu czasowym. Tak więc przykładowy szereg może mieć przebieg wykazujący trend, a po zróżnicowaniu otrzymane zostaną wartości, które w większości przypadków będą stacjonarne. Istnieją przypadki, gdzie zróżnicowanie pierwszego stopnia nie jest wystarczające do uzyskania przebiegu stacjonarnego. To którego stopnia różnice obliczane są w procesie zróżnicowania, definiuje wcześniej wspomniany parametr **d** w modelu   
ARIMA [15].

Dotychczas zostały omówione wszystkie komponenty modelu ARIMA. Biorąc pod uwagę wszystkie elementy, model ten można opisać za pomocą następującej formuły [16].

* = zmienna objaśniana w dziedzinie czasu;
* *c* = stała (dla uproszczenia formuły, często pomijana);
* = współczynnik każdego parametru p;
* = współczynnik każdego parametru q;
* = szum biały.

**4. CHARAKTERYSTYKA WYBRANEGO DO REALIZACJI ZADANIA OPROGRAMOWANIA**

Głównym założeniem tworzonego oprogramowania było zapoznanie użytkownika z pojęciem szeregów czasowych oraz przedstawienie możliwości metod analitycznych oraz eksploracyjnych, jakie można na owych szeregach czasowych wykonywać. Stworzona aplikacja umożliwia przesłanie przez użytkownika własnego pliku w formacie CSV, zawierającego szereg czasowy. Jeśli plik przesłany przez użytkownika jest poprawnego formatu, dane w nim zawarte podlegają natychmiastowo analizie statystycznej oraz wizualizacji. Osoba korzystająca z aplikacji posiada możliwość metody prognozy przyszłych wartości szeregów czasowych oraz parametryzacji metody prognozującej. Jako rezultat wykonania metody predykcyjnej, wygenerowany zostaje prognozowany przebieg w postaci wykresu liniowego, wraz z współczynnikami błędu predykcji.

**4.1 Opis funkcjonalności oprogramowania**

Ten podrozdział jest poświęcony szczegółowym opisom poszczególnych funkcjonalności oprogramowania. Na podstawie owych charakterystyk zostanie zaprojektowana architektura systemu, oraz zaimplementowane zostaną poszczególne jego funkcjonalności. Z racji tego, iż oprogramowanie to poświęcone jest głównie szeregom czasowym, omówione zostaną takie funkcje programu jak:

* Przesyłanie pliku;
* Analiza statystyczna oraz wizualizacja szeregu czasowego;
* Prognozowanie szeregów czasowych.

**4.1.1 Przesyłanie pliku**

Jednym z udogodnień tworzonego oprogramowania jest możliwość przesłania przez użytkownika do aplikacji dowolnego pliku w formacie CSV, w którym znajduje się szereg czasowy. W przesłanym przez użytkownika pliku, powinny zawarte być dwie kolumny. W jednej z nich powinny znajdować się znaczniki czasowe w postaci ciągu znaków. Mogą one zawierać datę w dowolnym formacie, lub inny znacznik czasu poszczególnych obserwacji. Druga z kolumn natomiast, powinna przechowywać wartości zmiennej której dotyczy obserwacja. Notowania te, powinny być wartościami liczbowymi całkowitymi lub zmiennoprzecinkowymi. Kolejność kolumn nie ma znaczenia, ważne natomiast jest, żeby w pierwszym wierszu, w kolumnie przechowującej znaczniki czasu znajdował się ciąg znaków z wartością „date”, natomiast w kolumnie zawierającej wartości, ciąg znaków o wartości „value”. Kolumny te powinny być rozdzielone przecinkiem. Poniżej przedstawiony jest początek zawartosci przykładowego pliku CSV o poprawnym formacie.



*Rysunek 4.1 Przykładowa zawartość plku CSV zawierającego szereg czasowy*

Przesyłanie pliku powinno odbywać się na pomocą prostego formularza. Po naciśnięciu odpowiedniego przycisku przez użytkownika, powinno ukazać się okno eksploratora, z możliwościa wskazania przez niego pliku z dowolnej lokalizacji na dysku, do której ma dostęp. Oprogramowanie powinno być zabezpieczone przed możliwością przesłania przez użytkownika pliku o formacie innym niż CSV. Jednocześnie użytkownik zostanie poinformowany przed przesłaniem pliku, o wymaganiach co do jego formatu oraz zaleceniach w kwestii poprawności zawartych w nim danych.

**4.1.2 Analiza statystyczna oraz wizualizacja szeregu czasowego**

Przesłany przez użytkownika plik z szeregiem czasowym, który spełnia wspomniane w poprzednim podrozdziale wymagania zostaje natychmiast poddany metodom analizy statystycznej oraz wizualizacji. Jest to kolejna, z ważnych funkcjonalności oferowanych przez tworzone oprogramowanie. Metody statystyczne, którym podlegają dane dostarczone w pliku, są bardziej szczegółowo opisane w rozdziale 3.2 *Analiza statystyczna szeregów czasowych*.

Przed właściwą analizą danych zawartych w pliku, szereg czasowy zostaje zweryfikowany pod kątem:

* komplementarności danych – sprawdzenie czy w żadnej z kolumn, nie występuje brakująca wartość.
* walidacji danych – czy w kolumnach znajdują się dane poprawnego typu.
* unifikacji danych – dotyczy wyłącznie wartości szeregu czasowego, polega na zaokrągleniu wszystkich wartości szeregu do tej samej liczby miejsc po przecinku.

Po zweryfikowaniu danych oraz wykonaniu podstawowych metod statystycznych, jako rezultat opisywanej funkcjonalności, użytkownikowi wyświetli się tabela z wartościach takimi jak:

* Wartość minimalna;
* Wartość maksymalna;
* Średnia arytmetyczna wartości;
* Odchylenie standardowe;
* Interkwartyl.

Również w tym samym widoku, obok tabeli z danymi statystycznymi, użytkownik powinien ujrzeć wizualizację szeregu czasowego w postaci wykresów:

* Wykresu przebiegu szeregu czasowego – wykres przedstawiający cały cykl przebiegu wykresu czasowego. Na osi X opisane są tu znaczniki czasowe reprezentujące przedziały pomiaru, natomiast na osi Y opisane są wartości obserwacji. Przykładowy wykres znajduje się na rysunku 3.1.
* Wykresu autokorelacji – wykres przedstawiający wartości współczynników autokorelacji pomiędzy 100 poprzednimi wartościami. Wartości te zawierają się w przedziale od -1.0 to 1.0.



*Rysunek 4.2 Przykładowy wykres autokorelacji szeregu czasowego dla 100 poprzednich wartości*

* Histogramu rozkładu zmiennej – wykres przedstawiający liczebność obserwacji w zadanych przedziałach.



*Rysunek 4.3 Przykładowy histogram rozkładu zmiennej szeregu czasowego*

**4.1.3 Prognozowanie szeregów czasowych**

Główną funkcjonalnością tworzonego systemu, jest możliwość testowania wybranych metod predykcji przez użytkownika, na przesłanym przez niego szeregu czasowym. Użytkownik ma do wyboru dwa modele umożliwiające prognozowanie przyszłych wartości szeregów czasowych, model autoregresyjny, oraz zintegrowany model autoregresyjny ze średnią ruchomą. Modele te zostały bardziej szczegółowo opisane w rozdziale 3.3 *Metody predykcji szeregów czasowych.* Testowanie działania metod prognozujących polega na wytrenowaniu modelu za pomocą zadanej ilości obserwacji, a następnie zweryfikowaniu jego skuteczności poprzez porównanie wyników prognozy z rzeczywistym przebiegiem. Użytkownik ma możliwość parametryzacji obu tych metod. Poprzez odpowiedni dobór wartości parametrów, może optymalizować model w celu zwiększenia jego wydajności oraz zminimalizowania błędu predykcji. Parametryzacja odbywa się poprzez formularz w którym możliwe jest wprowadzenie wartości liczbowej parametru, lub wybór jednej z wielu możliwych opcji.

W przypadku modelu autoregresyjnego, możliwe jest ustawienie wartości takich parametrów jak:

* Stosunek podziału zbioru wartości – parametr ten jest wartością zmiennoprzecinkową definiującą w jakich proporcjach zbiór wartości szeregu czasowego ma zostać podzielony na podzbiór treningowy oraz testowy. Przykładowo jeśli użytkownik dostarczy szereg czasowy o liczebności obserwacji równej 300, oraz jako wartość tego parametru poda 0.6, to model zostanie wytrenowany na 180 obserwacji, natomiast pozostałe 120 posłuży przy testowaniu jego skuteczności. Jako domyślna wartość tego parametru zostanie przyjęta 0.8.
* Kryterium doboru parametru opóźnienia – parametr definiujący, według którego kryterium ma zostać dobrana wartości rzędu predykcji (ilość poprzednich obserwacji, na podstawie których zostanie wykonania prognoza). Dla autoregresyjnych modeli dostępne są kryteria takie jak:
  + Kryterium Informacyjne Akaike’a (AIC);
  + Bayesowskie kryterium informacyjne Schwarza (BIC);
  + Kryterium informacyjne Hannana-Quinna (HQC).

Ich matematyczne formuły znajdują się w rozdziale 3.3.1 *Model autoregresyjny*.

W przypadku zintegrowanego modelu autoregresyjnego ze średnią ruchomą, użytkownik może dostosowywać parametry takie jak:

* Stosunek podziału zbioru wartości – definicja tego parametru jest taka sama, jak w przypadku parametru modelu autoregresyjnego, opisanego na poprzedniej stronie.
* Parametr Q – wartość tego parametru oznacza rząd predykcji. Domyślna wartość tego parametru to 10.
* Parametr I – stopień zróżnicowania szeregu czasowego. Domyślnie
* Parametr P – określa rozmiar okna średniej ruchomej.

Przeznaczenie oraz bardziej szczegółowe znaczenie powyższych parametrów znajduje się w rozdziale 3.3.3 *Model autoregresyjny zintegrowany ze średnią ruchomą (ARIMA).*

Jako wynik metody prognozującej, użytkownikowi zostanie wygenerowany widok, na którym znajdować się będzie zestawienie prognozowanych wartości wraz z realnym przebiegiem w postaci wykresu liniowego, ilość obserwacji którymi model był trenowany oraz rezultaty predykcji, zależne od zastosowanej metody.

Jeśli jako metoda prognozująca, wybrana została metoda autoregresji, wynikiem będzie:

* Rząd predykcji, który został wybrany na podstawie kryterium wskazanego przez użytkownika.
* Wartość błędu średniokwadratowego (RMSE) – określa skutecznośc modelu prognozującego. Im mniejsza wartość błędu RMSE, tym prognozy modelu były bardziej precyzyjne. Błąd RMSE określa się wzorem:

gdzie:

* - wielkość prognozowana;
* - wielkość rzeczywista.

Jeśli natomiast prognoza przeprowadzona była za pomocą modelu ARIMA, wartościami które określają błąd predykcji, będą wyniki formuł matematycznych kryteriów wspomnianych w przypadku omawiania parametrów modelu autoregresyjnego, czyli: Kryterium Informacyjnego Akaike’a (AIC), Bayesowskiego kryterium informacyjnego Schwarza (BIC) oraz Kryterium informacyjnego Hannana-Quinna (HQC).

**4.2 Wykorzystane technologie**

Oprogramowanie dydaktyczne, będące tematem pracy, zostało stworzone w formie aplikacji internetowej. Dostęp do aplikacji możliwy jest z każdego urządzenia posiadającego przeglądarkę internetową, oraz dostęp do internetu. Przy tworzeniu tego systemu zostały wykorzystane następujące języki programowania oraz narzędzia:

* Język Python w wersji 3.6.8 - interpretowany, zorientowany obiektowo język programowania. Dzięki wysokopoziomowym, wbudowanym strukturą danych oraz dynamicznemu typowaniu jest on idealnym językiem programowania do szybkiego i prostego tworzenia aplikacji. Język Python obsługuje wiele modułów, pakietów oraz bibliotek, co zachęca do tworzenia modułowego programu i ponownego wykorzystania tworzonego kodu. Python jak i jego standardowa biblioteka są dostępne bezpłatnie w postaci kodu źródłowego lub binarnego na wszystkich platformach [17]. Cała logika biznesowa tworzonego systemu została zaprogramowana właśnie w tym języku.
* Micro-framework Flask w wersji 1.1.2 – microframework języka Python, zapewniający narzędzia oraz biblioteki, które umożliwiają szybkie i proste budowanie aplikacji internetowych. Flask jest uważany za microframework, ponieważ nie zawiera on w sobie zależności do zewnętrznych bibliotek, co ogranicza jego bazową funkcjonalność, jednocześnie tworząc go lekkim i łatwym w instalacji. Rozwiązanie to posiada dużą ilość wtyczek dostarczających dowolną funkcjonalność do tworzonych aplikacji [18]. Biblioteka ta została wykorzystana w projekcie jako rdzeń aplikacji, w którym odbywały się wszystkie obliczenia, a ich wynikiem były renderowane strony internetowe.
* Biblioteka Matplotlib w wersji 3.3.2 – najbardziej popularna biblioteka w języku Python wykorzystywana do wizualizacji danych. Jest to międzyplatformowa biblioteka do tworzenia 2 lub 3 wymiarowych wykresów bazując na danych zawartych w tablicach. Główną zaletą tego rozwiązania jest możliwość wykorzystania jej możliwości na dowolnym systemie operacyjnym [19]. Biblioteka Matplotlib posłużyła do wizualizacji przebiegu szeregów czasowych oraz wykresów z nimi związanymi.
* Biblioteka Statsmodels w wersji 0.10.2 – to moduł języka Python, udostępniający klasy i funkcje do estymacji różnych modeli statystycznych, a także do przeprowadzania testów statystycznych i statystycznej eksploracji danych. Biblioteka ta została użyta w stworzonym systemie do tworzenia modelów predykcyjnych [20].

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

Test

**LITERATURA**

**Bibliografia literatury**

[1] „Mathematical software”, <https://en.wikipedia.org/wiki/Mathematical_software>, [przeglądane dnia 01.06.2021r]

[2] „MATLAB”, <http://www.ont.com.pl/produkty/lista-produktow/matlab>, [przeglądane dnia 01.06.2021r]

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

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

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

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

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

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

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

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

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

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

[13] „Średnia ruchoma”, <https://pl.wikipedia.org/wiki/%C5%9Arednia_ruchoma>, [przeglądane dnia 08.06.2021r]

[14] „Autoregressive integrated moving average”, <https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average>, [przeglądane dnia 08.06.2021r]

[15] „Co trzeba wiedzieć korzystając z modelu ARIMA i które parametry są kluczowe”, <https://support.predictivesolutions.pl/ekspress/download/EB54/mat/model_ARIMA_istotnosc_parametrow.pdf>, [przeglądane dnia 08.06.2021r]

[16] „Dow Jones Sustainability Index: Use of forecasting models to assist decision making”, <https://www.revistaespacios.com/a15v36n11/15361121.html>, [przeglądane dnia 08.06.2021r]

[17] „What is Python? Executive Summary”, <https://www.python.org/doc/essays/blurb/>, [przeglądane dnia 17.06.2021r]

[18] „What is flask?”, <https://pymbook.readthedocs.io/en/latest/flask.html>, [przeglądane dnia 17.06.2021r]

[19] „Visualization with Matplotlib”, <https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html>, [przeglądane dnia 17.06.2021r]

[20]

**Bibliografia rysunków**

[] <http://www.ont.com.pl/wp-content/uploads/2014/10/content_schemat_pracy.png>, [przeglądane dnia 01.06.2021r]

[] <https://upload.wikimedia.org/wikipedia/commons/thumb/2/21/Matlab_Logo.png/667px-Matlab_Logo.png>, [przeglądane dnia 01.06.2021r]

[] <http://www.ont.com.pl/wp-content/uploads/2014/10/content_matlab.png>, [przeglądane dnia 01.06.2021r]

[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]

[F] <https://pythondata.com/wp-content/uploads/2019/01/comparison_plt-1024x521.png>, [przeglądane dnia 08.06.2021r]

[G] <https://www.researchgate.net/profile/Denis-Butusov/publication/323860289/figure/fig3/AS:606644442517510@1521646714418/An-example-of-a-moving-average-with-the-smoothing-interval-of-1000-points-Time-scale-is.png>, [przeglądane dnia 08.06.2021r]

**Spis rysunków**

Rysunek 2.1 Logo programu MATLABs XX

Rysunek 2.2 Schemat pracy w programie MATLABs XX

Rysunek 2.3 Histogram wygenerowany za pomocą programu MATLAB s XX

Rysunek 2.4 Interfejs programu MATLABs XX

**Spis listingów**

Listing 1. Przykładowy fragment kodu w języku MATLABs XX