UNIWERSYTET
EKONOMICZNY
W POZNANIU

Michał Cukrowski

Prognozowanie nieliniowych szeregów czasowych za pomocą wybranych modeli ekonometrycznych oraz uczenia maszynowego

Nonlinear time series forecasting with selected econometric and machine learning models

Praca licencjacka

Promotor: dr A. Rutkowska

Pracę przyjęto dnia: .....

Kierunek: Informatyka i Ekonometria

Poznań 2022

Spis treści

Wstęp

Prognozowanie szeregów czasowych jest jedną z najważniejszych dyscyplin, którą zajmuje się ekonometria. Przez wiele lat opracowane zostały metody statystyczne w celu dokładnej i powtarzalnej estymacji wartości przyszłych, wśród których jednymi z najpopularniejszych są modele takie jak Autoregressive Moving Average (ARMA), Vector Autoregressive (VAR) oraz modele korzystające z kointegracji Vector Error Correction Model (VECM). Ich najważniejszym założeniem jednak jest liniowość relacji pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi - a praktyka pokazuje, że założenie to często nie jest spełniane, na przykład w danych makroekonomicznych lub giełdowych. W odpowiedzi na to zostały opracowane także modele nieliniowe takie jak na przykład modele przełącznikowe Self-Exciting Threshold Autoregressive (SETAR), Threshold Vector Autoregressive (TVAR), Threshold Vector Error Correction Model (TVECM) lub modele Markova Markov Switching Autoregressive (MS-AR).

Należy jednak zwrócić uwagę, że w przypadku modeli liniowych procedury testowania i wdrażania są stosunkowo proste i powtarzalne, natomiast w przypadku wystąpienia relacji nieliniowych nie istnieje jeden uniwersalny algorytm, który byłby w stanie wyjaśnić wszystkie zależności. Prognozowanie takich szeregów czasowych jest zatem o wiele bardziej pracochłonne i narażone na ryzyko błędnej identyfikacji. Motywacją niniejszej pracy jest zatem empiryczne sprawdzenie, czy wybrane modele uczenia maszynowego, które nie zakładają rodzaju zależności a priori: KNN, lasu losowego oraz drzewa decyzyjnego są w stanie dostarczyć podobnej jakości prognozy w stosunku do modeli ekonometrycznych.

Rozdział pierwszy wyjaśnia prognozowanie w ujęciu historycznym i pierwsze próby szacowania parametrów równania liniowego, różnice pomiędzy prognozami ilościowymi i jakościowymi oraz opisuje problematykę związaną z cechami statystycznymi szeregów czasowych. Wytłumaczone zostały także przykładowe problemy empiryczne związane z prawidłową identyfikacją szeregu nieliniowego.

W rozdziale drugim opisane zostają wszystkie wykorzystane modele ekonometryczne, uczenia maszynowego a także wykorzystane testy statystyczne i metryki błędów. Wyjaśniona zostaje również metodyka badania empirycznego.

Rozdziały trzeci, czwarty i piąty skoncentrowane są na dokładnym opisaniu szeregów czasowych branych pod uwagę a analizie reszt z poszczególnych modeli.

Rozdział szósty poświęcony jest na przedstawienie wyników badania a także ich interpretacji. Zaprezentowane zostają wnioski i obserwacje.

1 Rodzaje prognoz, danych i modeli

1.1 Kontekst historyczny

Prognozowanie, czyli przewidywanie przyszłych zjawisk lub zdarzeń (PWN, 2022) nie jest dziedziną nową, jego początki sięgają już czasów starożytnych, a pierwsze historyczne przekazy na ten temat odwołują się do Imperium Chaldejskiego. Przez stulecia Chaldejczycy zamieszkujący tereny Asyrii i Babilonii zbierali duże ilości obserwacji astronomicznych, wydarzeń historycznych i wyników rynkowych mając na celu wykorzystanie korelacji do odkrycia regularności w tych danych i przewidzenia wartości przyszłych - takich jak poziom wody Eufratu, cen surowców rolnych oraz wydarzeń politycznych (Elliott i in., 2006-). Oprócz tego interesowali się krótkoterminowym przewidywaniem pogody - nie były to jednak metody oparte na skomplikowanych obliczeniach, z którymi kojarzy się dzisiejsza metodyka meteorologii. Zamiast tego próbowano znaleźć relację pomiędzy zmianą pogody a wyglądem chmur lub występowaniem zjawisk optycznych takich jak efekt halo (NASA, 2022).

Interesujące podejście do oceniania zdolności przewidywania przyszłości przejawiali później starożytni Grecy. Wraz z rozwojem ówczesnej debaty publicznej pojawiło się współzawodnictwo, którego obiektem zainteresowania poza sportem były między innymi teorie naukowe. Obejmowało ono umiejętność przepowiadania przyszłości na temat ludzkiego zdrowia, a rywalizowali między sobą lekarze oraz wróżbici. Warto zaznaczyć, że ci drudzy posiadali już solidnie ukształtowaną pozycję społeczną i nie odczuwali silnej potrzeby udowodnienia swoich racji w zawodach. Nie odcinali się także od metod ówczesnych medyków. Istnieją jednak przesłanki by sądzić, że lekarze utrzymywali jednak od wróżbitów pewien dystans. Niemniej jednak obie grupy stosowały częściowo pokrywający się zbiór metod. W kontekście współzawodnictwa to niestety te nowe nurty, takie jak prognozy lekarzy opierających swoje umiejętności o wiedzę Hipokratesa, musiały wywalczyć sobie klientów oraz reputację (Raphals, 2013).

Kiedy jednak zaczęto do prognozowania wykorzystywać narzędzia statystyczne? Do połowy osiemnastego wieku istniała przynajmniej jedna technika, która była często używana przez ówczesnych astronomów i nawigatorów. Była to zwykła średnia arytmetyczna. W obawie o błędy zwracano jednak dużą uwagę na dokładność pomiaru, a zatem by analizowane obserwacje zostały wykonane w takich samych warunkach, przez tego samego obserwatora i za pomocą tego samego ekwipunku. (Stigler, 1986)

Kiedy jednak pojawiły się pierwsze metody estymacji liniowej, które znamy dzisiaj jako równania regresji? Jedną z pierwszych metod szacowania parametrów równania liniowego była metoda najmniejszego odchylenia bezwzględnego opisana w 1757 roku przez Rogera Josepha Boscovicha. Opracowany przez niego algorytm był jednak metodą mającą swoje uzasadnienie w geometrii, formę analityczną opracował do niej dopiero 30 lat później Pierre Simon de Laplace(Dodge, 2010). Dla prostego równania regresji:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

należało znaleźć taką parę parametrów β, aby zminimalizować:

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i| \tag{2}$$

Był to jednak algorytm iteracyjny, opierający się na sprawdzaniu prostych poprowadzonych przez kolejne dwa punkty ze zbioru danych. Idea minimalizacji odchylenia bezwzględnego miała dwie poważne wady:

1. Nieunikalne rozwiązania – oznacza to, że istnieją przynajmniej dwa rozwiązania, dla których $\sum |e_i|$ przyjmuje optymalną wartość.

2. Degeneracja – oznacza to, że w rozwiązaniu przynajmniej jedna zmienna bazowa jest równa zero.

Prawdziwym przełomem było jednak opublikowanie w 1805 roku przez Legendre'a publikacji pod oryginalnym tytułem *"Nouvelles méthodes pour la détermination des orbites des comètes"*, czyli *"Nowe metody wyznaczania orbit komet"* oraz odpowiednich do niej uzupełnień, w której opisał on w szczegółach metodę najmniejszych kwadratów (Stigler, 1986). Analogicznie do metody Laplace'a, dla funkcji (1) należało znaleźć taką parę parametrów β, aby zminimalizować:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3}$$

W przeciwieństwie do metody najmniejszego odchylenia bezwzględnego, metoda ta nie posiada problemu z unikalnością rozwiązania oraz degeneracją, a jej niezaprzeczalną zaletą w tamtych czasach była łatwość i szybkość obliczeń.

1.2 Rodzaje prognoz

Rodzaje prognoz można podzielić na dwie najważniejsze kategorie:

1. ilościowa – stan danej zmiennej wyrażony jest liczbą, taka prognoza może być punktowa lub przedziałowa,

2. jakościowa – przewidywany jest stan zmiennej jakościowej.

Prognoza czysto jakościowa nie wymaga w ogóle obróbki danych, używany jest jedynie osąd prognosty. W ekstremalnym przypadku przewidywania specjalisty mogą być oparte na danych przetworzonych w jego myślach. Z drugiej strony, podejście czysto ilościowe nie wymaga osądu eksperta w ogóle i wytwarza wyniki czysto ilościowe (Hanke i Wichern, 2014a). Jest jednak istotnym, aby pamiętać, że nawet najlepszy model ilościowy może nie brać pod uwagę wszystkich informacji wynikających na przykład ze zmian strukturalnych lub anomalii statystycznych (które

nie zawsze są możliwe do wykrycia i wykorzystania do modyfikacji modelu w optymalnym momencie). Jeśli prognozowane są dane ekonomiczne, przykładem takich wydarzeń może być duży szok popytowy albo podażowy wynikający z wydarzeń politycznych, sankcji gospodarczych albo nagłego zerwania współpracy pomiędzy podmiotem będącym obiektem zainteresowania prognosty, a dostarczycielem dóbr produkcyjnych bądź dystrybutora produktu finalnego.

W ciągu ostatnich lat zainteresowanie metodami ilościowymi w estymacji interesujących nas informacji zdecydowanie wzrosło. Możliwości, jakie otwierają metody ekonometrii oraz uczenia maszynowego przyciągają uwagę zarówno naukowców jak i przedsiębiorców, ponieważ ich wykorzystanie jest bardzo praktyczne. Mają one za zadanie wspomagać procesy decyzyjne w sytuacjach wysokiego ryzyka lub wymagających odpowiedniej optymalizacji - takich jak na przykład planowanie produkcji w oparciu o dane sprzedaży, lub dostosowanie polityki banku centralnego do sytuacji makroekonomicznej.

Rysunek 1.1: Liczba wyszukań słowa ”Prognozowanie”w wyszukiwarce Google. Dane względne, 100 oznacza maksymalną ilość zapytań w badanym okresie. Źródło: Google Trends

Na przestrzeni lat, w związku ze wzrostem zainteresowania, metody statystyczne i uczenia maszynowego zostały w dużej mierze zautomatyzowane i są dostępne dla niemal każdego. Wraz z rozpowszechnieniem komputerów osobistych o dużej mocy obliczeniowej i zaawansowanych pakietów statystycznych, wygenerowanie prognozy nie jest tak skomplikowane jak w przeszłości, jednak ta łatwość obliczeń nie może zastąpić logicznego myślenia. Brak nadzoru i niewłaściwe stosowanie technik prognostycznych może doprowadzić do kosztownych błędów. (Hanke i Wichern, 2014b)

1.3 Rodzaje danych i modeli

Istnieją dwa podstawowe rodzaje danych, które analizować można za pomocą metod statystyki matematycznej:

1. dane przekrojowe – są to dane zgromadzone przez obserwację wielu podmiotów w jednym okresie (Wikipedia, 2022a),
2. dane czasowe – są to dane zindeksowane w porządku chronologicznym w wielu okresach (Wikipedia, 2022b).

Dane przekrojowe

W pewnym badaniu na temat wartości Y zebrane zostały dane X będące wektorem $X = (x_1, x_2, ..., x_k)$. Poniższy model opisuje zależność pomiędzy Y i X:

$$Y = f(X, \beta) + \varepsilon \tag{4}$$

gdzie β jest wektorem k parametrów, a ε jest czynnikiem losowym. Mając to na uwadze, modele można podzielić na 3 podstawowe rodzaje:

1. Parametryczne – modelami parametrycznymi są modele, w których wektor parametrów β z przyjętego wcześniej równania jest wektorem określonym w skończonej p-wymiarowej przestrzeni (p może być zarówno mniejsze jak i większe od k). Do estymacji tego typu modelu niezbędne jest oszacowanie parametrów β, badacz musi także założyć jaka jest jego budowa i założenia. Przykładem takiego modelu może być model regresji liniowej:

$$Y = \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{5}$$

   Dla tego modelu istnieją cztery założenia (PennState, 2022), które powinny być spełnione:

   (a) Liniowość – zależność pomiędzy Y i X jest liniowa,
   (b) Homoskedastyczność – wariancja ε jest stała,
   (c) Niezależność – Y są niezależne od ε,
   (d) Normalność – rozkład ε jest rozkładem normalnym

   Modele parametryczne mogą przyjmować także formę nieliniową, przykładem może być model linearyzowany:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 e^{\beta_3 x_2} + \varepsilon \tag{6}$$

2. Nieparametryczne – modele nieparametryczne różnią się od parametrycznych tym, że forma modelu nie zostaje określona a priori, ale jest określona na podstawie podstawionych danych. W ich przypadku zbiór estymowanych parametrów jest podzbiorem wektora o nieograniczonej liczbie wymiarów. Przykładem modelu nieparametrycznego mogą być algorytmy uczenia maszynowego takie, jak KNN, drzewo decyzyjne oraz las losowy.

3. Semiparametryczne – modele semiparametryczne są hybrydą pomiędzy modelami parametrycznymi i nieparametrycznymi. Łączą one swoje najlepsze cechy, czyli mniejszą złożoność obliczeniową w estymacji niż modele nieparametryczne ale także zachowane są częściowo możliwości interpretacji. Ich struktura może się znacznie od siebie różnić w zależności od danych, jednak przykładowa forma modelu semiparametrycznego wyrażona może być równaniem:

$$Y = \beta_0 + \beta_1 x_1 + f(x_2) + \varepsilon \tag{7}$$

W takim przypadku relacja pomiędzy Y a x_1 jest liniowa o stałym współczynniku, natomiast nieznana jest relacja pomiędzy Y a x_2 (Mahmoud, 2019).

Szeregi czasowe

Na początku wyjaśnione zostaną pojęcia procesu stochastycznego oraz szeregu czasowego (SGH, 2022):

1. Proces stochastyczny Y_t w czasie dyskretnym stanowi ciąg zmiennych losowych, które są uszeregowane względem obserwacji w czasie, oznaczanych indeksem t.

2. Szereg czasowy y_t stanowi pojedynczą realizację procesu stochastycznego.

Procesy stochastyczne mogą być czysto losowe lub zależne. Procesem czysto losowym, nazywa się skokowy proces X_t, którego elementy są ciągiem niezależnych zmiennych losowych o tych samych rozkładach. Jego funkcja autokowariancji jest równa:

$$\gamma(k) = cov\,(X_t\,,\, X_{t-k}) = 0\; dla\; k \neq 0 \tag{8}$$

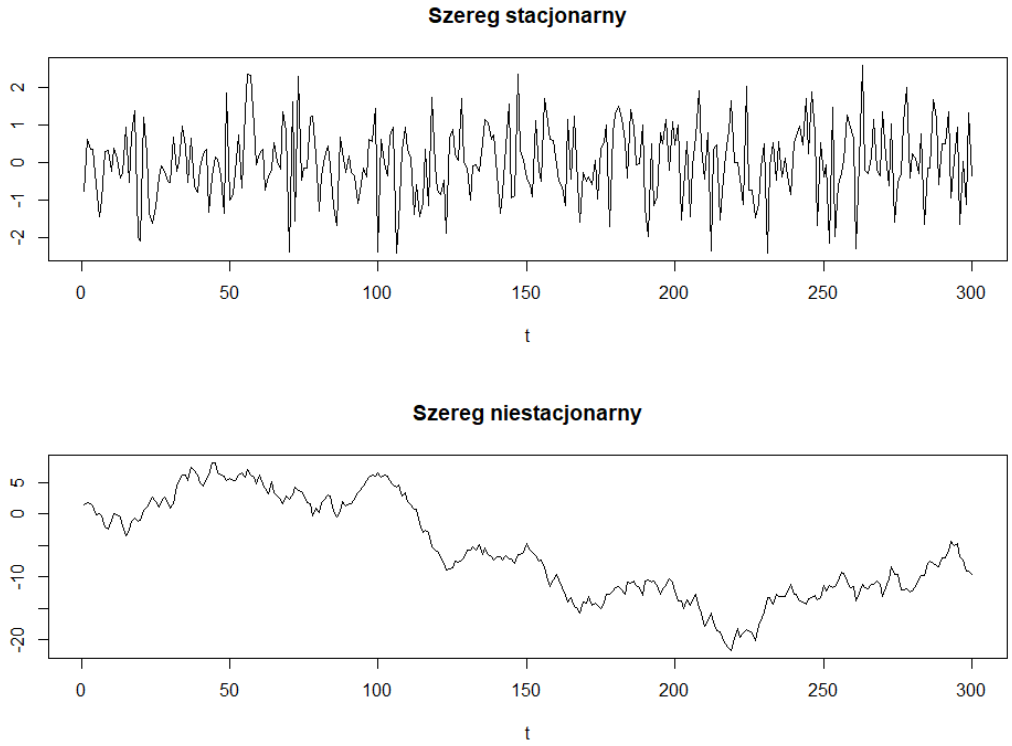
Proces taki nazywany jest również białym szumem. Istotnym rodzajem procesów stochastycznych są procesy stacjonarne. Dzielą się one na szeregi ściśle stacjonarne i słabo stacjonarne. (Maddala, 1994) Proces ściśle stacjonarny dla dowolnego rzędu posiada identyczny łączny rozkład dowolnego zbioru n obserwacji $X_{t_1}, X_{t_2}, \ldots, X_n$ z łącznym rozkładem $X_{t_1-k}, X_{t_2-k}, \ldots, X_{t_n-k}$, oraz posiada następujące właściwości:

1. Średnia $\mu = E(X_t)$

2. Wariancja $\sigma^2 = var(X_t)$

3. Funkcja autokowariancji $\gamma(t_1,t_2) = cov(X_{t_1,}, X_{t_2})$

Szeregi stacjonarne oraz niestacjonarne są w wielu przypadkach proste do rozróżnienia wizualnie:



Rysunek 1.2: Wykresy wygenerowanych procesów stacjonarnych i niestacjonarnych

Charakteryzują się one zatem stałą średnią, niezależną od czasu wariancją oraz stałą, zależną od opóźnienia funkcją autokowariancji. Założenie stałości wariancji jest jednak w praktyce bardzo restrykcyjne, dlatego wyróżniane są także procesy słabo stacjonarne, opisywane następującymi właściwościami:

1. Średnia $\mu = E(X_t)$
2. Funkcja autokowariancji $\gamma(t_1,t_2) = cov(X_{t_1,}, X_{t_2})$

Szeregi nieliniowe

Liniowy szereg czasowy można zapisać w formie ogólnej rzędu p jako sumę:

$$Y_t = \mu + \sum_{p=-\infty}^{\infty} \beta_p Y_{t-p} \tag{9}$$

Gdzie μ i β są liczbami rzeczywistymi, z $\beta_0 = 1$, $\sum_{p=-\infty}^{\infty} |\beta_p| < \infty$, a sam proces Y_t jest procesem idd o dobrze zdefiniowanej funkcji gęstości. Identyfikacja zależności w takim procesie przebiega

zazwyczaj poprzez zastosowanie testów ACF i PACF. Jeżeli szereg nie spełnia tych warunków, to powoduje to, że jest on nieliniowym szeregiem czasowym – dlatego świat zależności nieliniowych jest niezwykle obszerny i aby dalsza analiza miała sens, muszą zostać nałożone pewne ograniczenia. Powoduje to różne podejścia w badaniach na ich temat, a to skutkuje różnymi klasami modeli nieliniowych. (R. S. Tsay i Chen, 2019, str. 3). Przykładem modeli nieliniowych mogą być:

1. modele przełącznikowe,
2. modele dwuliniowe,
3. modele warunkowej heteroskedastyczności,
4. modele nieparametryczne,
5. modele z długą lub średnią pamięcią,

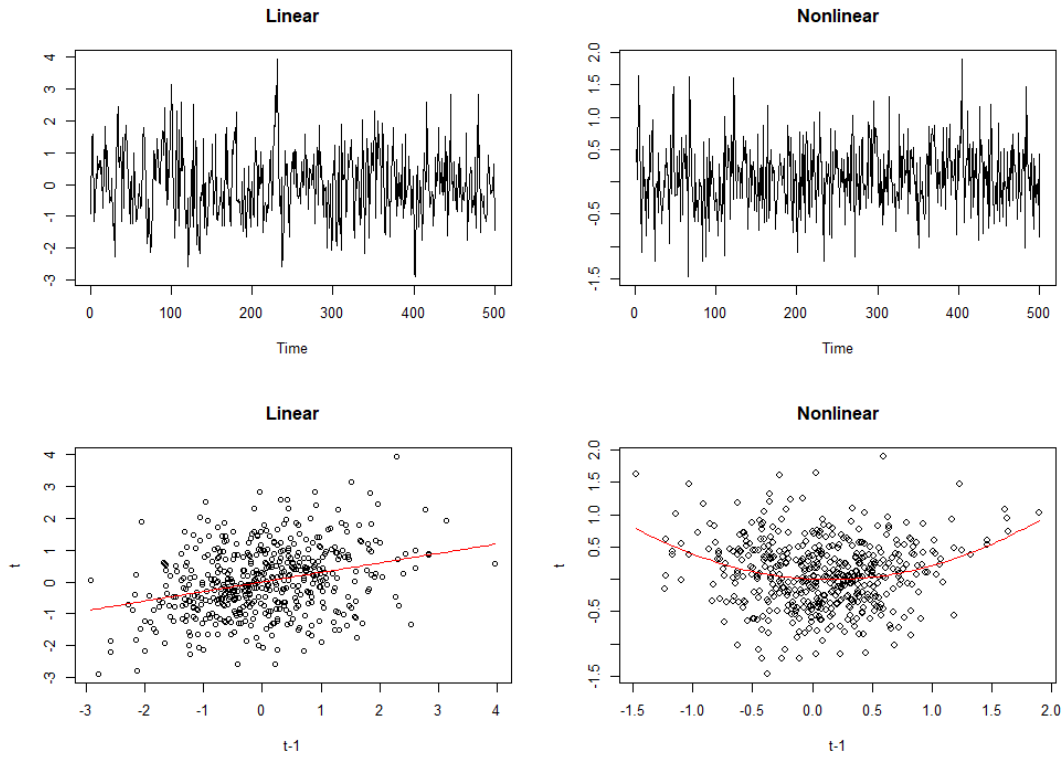
Jest to oczywiście niepełna lista stanowiąca jedynie pogląd dla czytelnika. Szeregi nieliniowe rzędu p brane pod uwagę w niniejszej pracy można zapisać jako:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}) + \varepsilon_t \tag{10}$$

Testy autokorelacji, które są zazwyczaj używane do klasycznej analizy szeregów czasowych mierzą jedynie zależności liniowe – nie są one dokładne, gdy zależność definiująca szereg jest nieliniowa (Henrik Madsen and Jan Holst, 2009). W przypadku, gdy nie zostanie wykazana autokorelacja liniowa, muszą być przyjęte inne kryteria oceny zależności. Jeżeli przykładowy szereg czasowy jest zależny wyłącznie od jednego opóźnienia, to może on przyjąć formy:

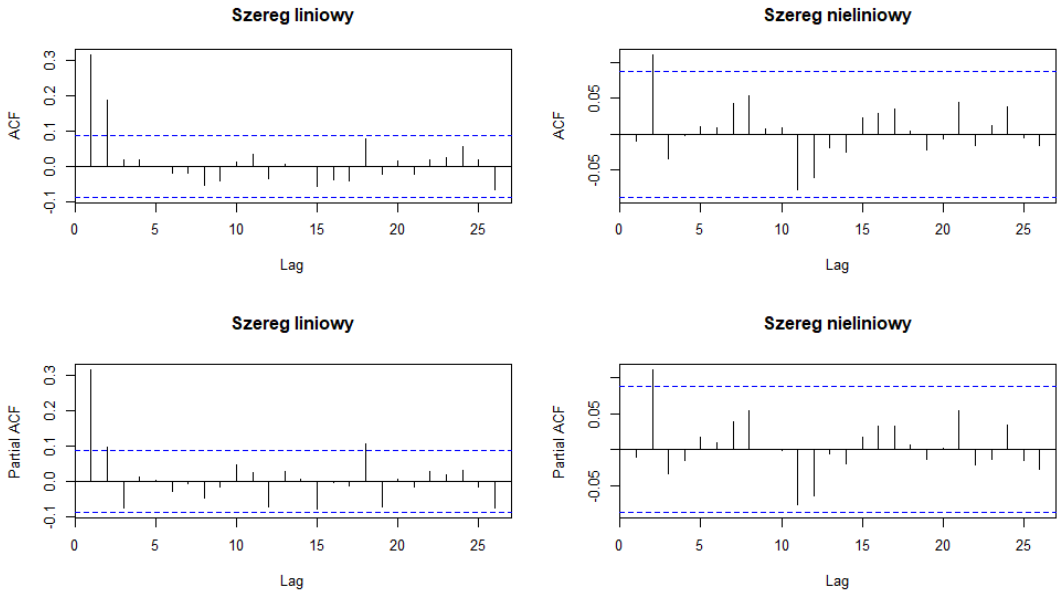
$$\begin{aligned} Y_t &= \beta Y_{t-1} + \varepsilon_t \\ Y_t &= f(Y_{t-1}) + \varepsilon_t \end{aligned} \tag{11}$$

Aby zademonstrować empiryczne różnice, jakie mogą wyniknąć z tytułu przyjęcia takich ograniczeń, dla parametru $\beta = 0.3$ oraz dla funkcji $f(x) = 0.3x(x - 0.3)$ wygenerowano odpowiednio szereg liniowy i nieliniowy.



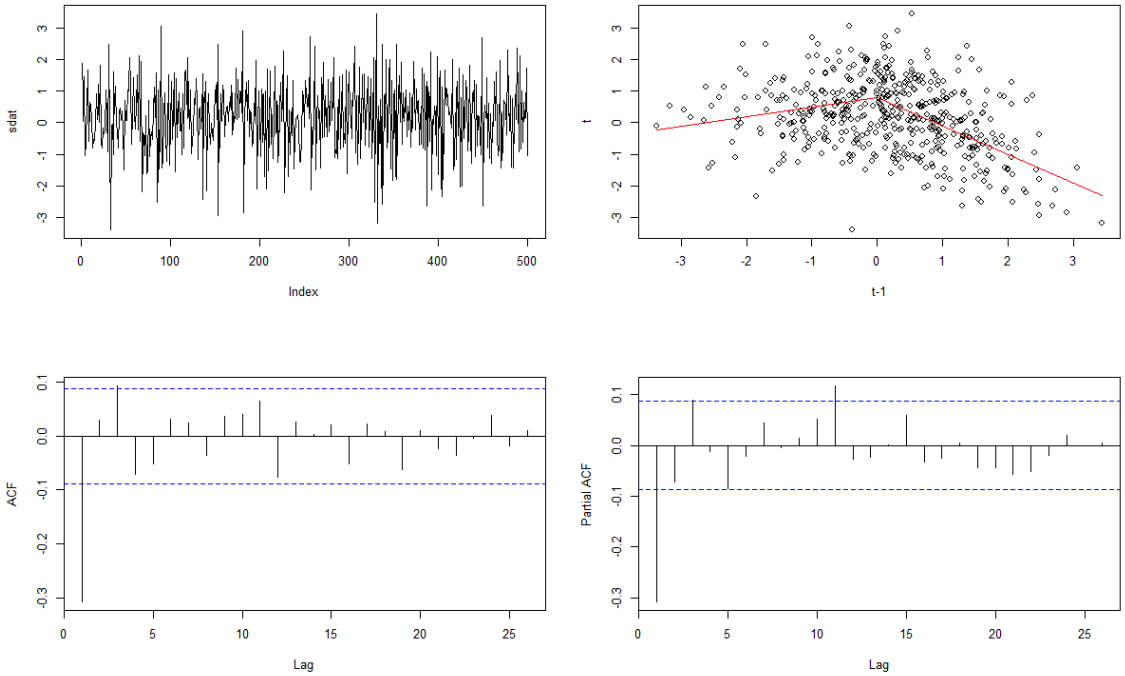
Rysunek 1.3: Wykresy wygenerowanych szeregów czasowych oraz ich wykresy rozrzutu z wyświetlonymi na czerwono funkcjami generującymi te szeregi.

Jak widać na wykresie, wizualne rozróżnienie szeregu liniowego i nieliniowego nie jest możliwe, natomiast jest widoczna zależność dla pierwszego opóźnienia w przypadku obu z nich na wykresach rozrzutu. W przypadku modelu nieliniowego standardowe testy ACF i PACF nie wykazują autokorelacji pierwszego rzędu:



Rysunek 1.4: Wyniki testów ACF i PACF dla szeregów liniowych i nieliniowych.

Na tym etapie można odnieść wrażenie, że istnieje twarda granica pomiędzy światem zależności liniowych i nieliniowych. W rzeczywistości jednak różnica pomiędzy takimi szeregami bywa nieuchwytna i przy stosowaniu niepełnych procedur testowania bardzo łatwo jest o pomyłkę - proces nieliniowy błędnie zakwalifikowany może zostać jako liniowy, a to może mieć poważne konsekwencje w prognozowaniu lub wnioskowaniu statystycznym. Aby to zademonstrować, wygenerowany został jeszcze jeden zależny od jednego opóźnienia szereg nieliniowy:



Rysunek 1.5: Na górze wykres wygenerowanego procesu nieliniowego w czasie oraz jego wykres rozrzutu z wyświetlonymi na czerwono funkcjami generującymi ten szereg. Na dole wyniki testów ACF i PACF.

W tym przypadku pomimo nieliniowości wygenerowanego procesu testy ACF i PACF wykazały istotną statystycznie autokorelację liniową. Prawidłowa identyfikacja procesów stochastycznych stojących za nieliniowymi szeregami czasowymi, których dotyczy analiza potrafi być zatem zadaniem bardzo złożonym. W niesprzyjających okolicznościach standardowa analiza zależności wyłącznie liniowych może prowadzić do mylnych wniosków. Uproszczenia tego procesu i ulepszenia dokładności stawianych prognoz szeregów nieliniowych można zatem szukać wśród modeli nieparametrycznych - nie zakładają one rodzaju zależności a priori, ale pozwalają na odpowiednie dostosowanie się do istniejących już danych.

2 Wykorzystane modele i testy statystyczne

2.1 Modele uczenia maszynowego

KNN

K-najbliższych sąsiadów (K-nearest neighbors jest nieparametrycznym, nadzorowanym algorytmem uczenia maszynowego opierającym się na dystansach. Dla klasyfikacji wykorzystuje on w drodze głosowania k najbliższych sąsiadów, natomiast w przypadku regresji jest z nich liczona średnia. Do jego działania wymagane jest podanie kilku parametrów, czyli:

1. Parametr 'K': jest to liczba najbliższych sąsiadów biorących udział w głosowaniu. Wyznaczyć go można za pomocą algorytmów walidacyjnych.

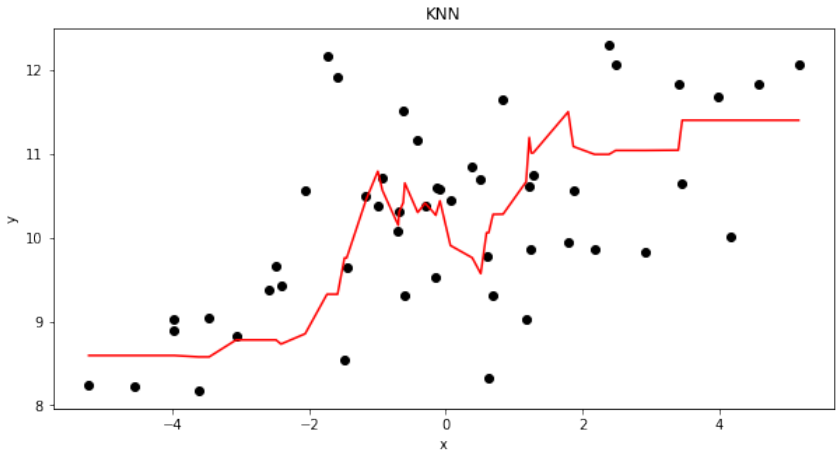
2. Rodzaj dystansu: jest to sposób liczenia odległości pomiędzy sąsiadami. Zazwyczaj jest on predefiniowany w zależności od rodzaju problemu, najczęściej wykorzystywane są dwie miary wyrażone wzorem:
$$(\sum_{i=1}^{n} |X_i - Y_i|^p)^{\frac{1}{p}} \tag{12}$$
Jeśli parametr $p = 1$ to jest to metryka Manhattan, w przypadku $p = 2$ jest to dystans Euklidesowy (NIST, 2022).

3. Wagi – jest to sposób ważenia sąsiadów. Mogą one być jednolite lub zmniejszające się proporcjonalnie do dystansu pomiędzy sąsiadami.

Jego implementację możemy podzielić na dwie fazy:

1. Uczenie – w tej fazie zapisywane są dane do postaci uczenia nadzorowanego, czyli zapisywana jest macierz danych X oraz macierz prawidłowych odpowiedzi Y.

2. Predykcja – w tej fazie wyliczana jest macierz odległości pomiędzy wszystkimi danymi należącymi do zbioru X oraz danej, dla której liczona jest predykcja. KNN jest zaliczany do algorytmów *"lazy learner"* (Rhys, 2020), ponieważ wszystkie obliczenia następują dopiero w trakcie liczenia predykcji.

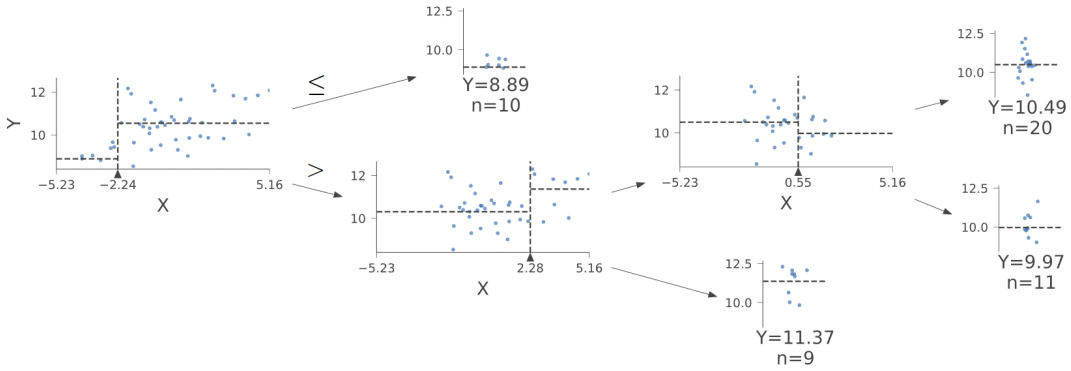


Rysunek 2.1: Wykres rozrzutu danych wygenerowanych z nałożoną linią regresji KNN dla $k = 4$

Drzewa decyzyjne i las losowy

Drzewo decyzyjne (decision tree) jest jedną z najstarszych metod nadzorowanego uczenia maszynowego stosowanych zarówno w klasyfikacji jak i regresji. Algorytm ten polega na znalezieniu na podstawie zbioru danych X takiej kombinacji decyzji binarnych, dla której przyjęta funkcja błędu jest najmniejsza. Drzewa decyzyjne składają się z korzenia, węzłów i liści. Każde drzewo decyzyjne zaczyna się od korzenia, który jest pierwszym węzłem rozpoczynającym proces decyzyjny. Następnym elementem jest węzeł, na którym definiowane są kolejne warunki i rozdzielający powstałe podzbiory zbioru X na optymalne, a ostatnią częścią drzewa decyzyjnego są liście, które reprezentują odpowiedź na zadany problem.

Ze względu na łatwość wizualizacji, drzewa decyzyjne są szczególnie cenione w problemach wymagających interpretacji takich jak na przykład automatyzacja procesów biznesowych. Do wytrenowania drzewa decyzyjnego wykorzystuje się algorytm „recursive binary splitting”. W klasyfikacji jako kryterium zazwyczaj wykorzystuje się np. entropię lub indeks Gini, natomiast w przypadku regresji najczęściej stosowanym algorytmem jest redukcja wariancji. Polega ona na iteracyjnym znalezieniu takiego punktu A, dla którego wartość sumy kwadratów reszt od średnich liczonych oddzielnie po obu stronach punktu A jest najmniejsza spośród wszystkich kombinacji. Ten punkt A jest progiem korzenia, który dzieli zbiór na pierwsze dwie części. Następnie analogicznie wyznaczane są progi w węzłach do momentu, w którym spełniony zostanie dowolny warunek pre-pruning, lub dalszy podział nie będzie możliwy.



Rysunek 2.2: Wizualizacja algorytmu recursive binary splitting na przykładzie empirycznym

Drzewa decyzyjne ze względu na swoją budowę są bardzo podatne na przeuczenie. Uniknąć go można stosując przycinanie (pruning). Przycinanie podzielić można na dwa rodzaje:

1. Pre-pruning – polega na zatrzymaniu dodatkowego rozgałęziania drzewa przed wzięciem pod uwagę całego zestawu treningowego.

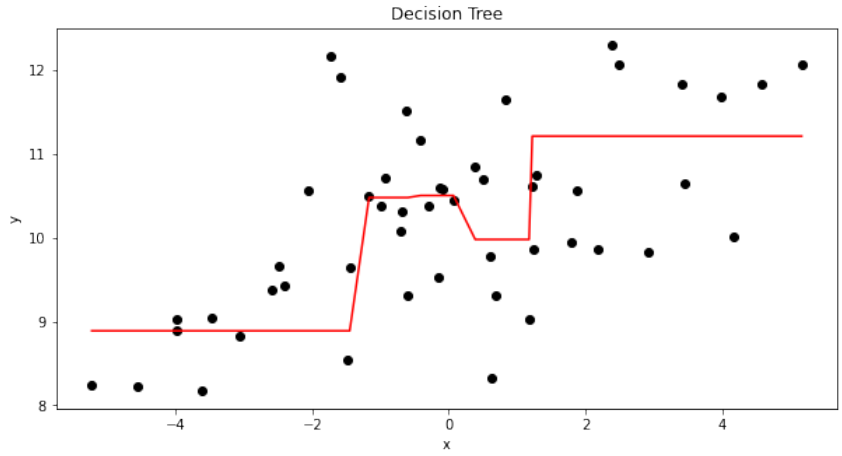
2. Post-pruning – najpierw pozwala drzewu rozwinąć się maksymalnie, a następnie pozbywa się najmniej potrzebnych rozgałęzień, stosownie do wybranego algorytmu.

Do wykorzystania metody pre-pruning, a zatem powstrzymania drzewa od nadmiernego podziału, potrzebne jest zdefiniowanie parametrów modelu korzystając z algorytmu walidacji krzyżowej (cross-validation). W badaniu poruszonym w dalszej części pracy wykorzystane zostaną parametry:

1. Minimalna liczba obserwacji do możliwości dalszego podziału drzewa – odpowiednia wartość tego parametru powoduje, że drzewo nie będzie się dzielić za sprawą przypadków, które wystąpiły bardzo małą liczbę razy. Im wyższa jego wartość tym lepsza generalizacja.

2. Minimalna liczba obserwacji wymagana dla każdego liścia – dla odpowiedniej wartości tego parametru nie będą tworzone osobne liście dla przypadków, które wystąpiły zbyt małą liczbę razy. Wraz ze zwiększeniem jego wartości wzrasta generalizacja, a więc i podatność na niedouczenie.

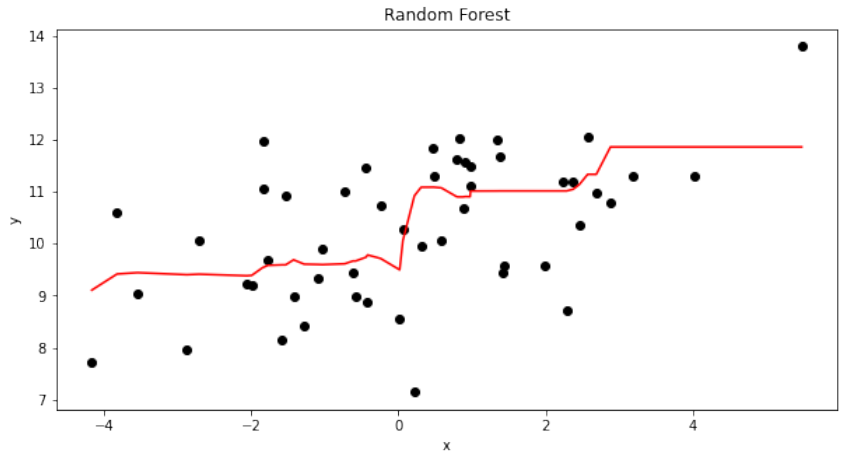
3. Maksymalna wysokość drzewa – zapobiega tworzeniu się zbyt wielu podziałów za pomocą ograniczenia liczby razy, jaką drzewo może się podzielić.



Rysunek 2.3: Wykres rozrzutu danych wygenerowanych z nałożoną linią regresji oszacowanej przez drzewo decyzyjne

Rozwinięciem algorytmu CART jest las losowy (random forest). Tak samo jak ten poprzedni, może on być stosowany zarówno do klasyfikacji jak i regresji. Las losowy składa się z wielu różnych od siebie wzajemnie drzew decyzyjnych. W jego implementacji wszystkie te drzewa wykonują swoje predykcje, a finalny wynik jest średnią arytmetyczną wszystkich wyników drzew. Aby drzewa różniły się od siebie nawzajem stosowany jest bootstrapping danych treningowych.

Las losowy dziedziczy po CART wszystkie parametry, natomiast oprócz nich niezbędne jest także zdefiniowanie ilości estymatorów – jest to ilość drzew decyzyjnych w całym lesie. Wraz ze wzrostem tego parametru zwiększa się dokładność modelu. Szacowany jest on za pomocą algorytmów walidacyjnych.



Rysunek 2.4: Wykres rozrzutu danych wygenerowanych z nałożoną linią regresji oszacowanej przez las losowy

2.2 Modele ekonometryczne i testy statystyczne

Modele liniowe

Jeżeli ε_t jest stacjonarnym procesem czysto losowym ze średnią μ równą zero oraz wariancją σ^2, proces X_t wyrażony wzorem:

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_r X_{t-p} + \varepsilon_t \tag{13}$$

jest procesem autoregresyjnym AR(p), którego wzór można zapisać także jako:

$$y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t \tag{14}$$

Liniowe modele autoregresyjne szacowane mogą być za pomocą różnych metod takich jak na przykład: równania Yule-Walkera, metoda największej wiarygodności lub metoda najmniejszych kwadratów. Kolejnym rodzajem liniowego procesu jest model średniej ruchomej MA(q). W odróżnieniu jednak od modeli AR(p), zamiast poprzednich wartości szeregu Xi pod uwagę brane są poprzednie błędy ε według wzoru:

$$X_t = \mu + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t \tag{15}$$

W rozumieniu ścisłym model średniej ruchomej również jest procesem regresyjnym, jednak ze względu na swoją naturę nie jest możliwe oszacowanie jej parametrów analogicznie do modelu AR(p) za pomocą metody najmniejszych kwadratów. Najczęściej stosowaną metodą jest metoda największej wiarygodności. Naturalnym połączeniem powyższych dwóch modeli jest model ARMA(p, q). Wyrażony jest on wzorem:

$$X_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t \tag{16}$$

Podobnie jak w przypadku modelu MA(q) nie jest możliwym uzyskanie estymatorów regresji bezpośrednio za pomocą metody najmniejszych kwadratów, zamiast tego najczęściej wykorzystywanym algorytmem jest metoda największej wiarygodności stosowana dla wszystkich parametrów α oraz β jednocześnie.

W przypadku wyżej wymienionych modeli największym ograniczeniem wydaje się być wymóg stacjonarności – większość szeregów czasowych analizowanych przez ekonometrię takich jak na przykład dane makroekonomiczne, ceny instrumentów finansowych lub w ostatnich latach dane epidemiologiczne są szeregami niestacjonarnymi. Ten problem jest jednak z praktycznego punktu widzenia dość prosty do ominięcia – co prawda sam szereg może być niestacjonarny, natomiast jego zmiany mogą być realizacją stacjonarnego procesu (Hyndman, 2018) Modele wykorzystujące tę własność nazywa się zintegrowanymi. Do wykorzystania mechani-

zmu różnicowania korzysta się z operatora opóźnienia B:

$$By_t = y_{t-1} \tag{17}$$

Alternatywną notacją jest L (B: 'backshift' - przesunięcie w tył, L: 'Lag' - opóźnienie). Za pomocą tak zdefiniowanego operatora opóźnienia łatwo można wyznaczyć ogólną formę różnicowania stopnia d:

$$(1-B)^d y_t \tag{18}$$

Modelami wykorzystującymi operator różnicowania są to modele zintegrowane, takie jak ARI(p, d), IMA(d, q), ARIMA(p, d, q). Niesie to za sobą bardzo ważne implikacje - jeżeli szereg czasowy jest niestacjonarny, ale jego różnice są stacjonarne i spełniają założenia modelu niezintegrowanego, to możliwe jest zaprognozowanie go za pomocą modeli liniowych. Szereg taki nazywany jest zintegrowanym w stopniu d:

$$I(d) \tag{19}$$

Model przełącznikowy

Jednym z najpopularniejszych modeli nieliniowych w ekonometrii jest rodzina modeli przełącznikowych SETAR (Self-Exciting Threshold AutoRegressive model). Stanowi on rozszerzenie tradycyjnych modeli autoregresyjnych o przełączenie pomiędzy reżimami, które zdefiniowane są za pomocą progów. Model SETAR o dwóch reżimach można zapisać jako:

$$\begin{cases} y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sigma\varepsilon_t, & y_{t-d} > r \\ y_t = \theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i} + \sigma\varepsilon_t, & y_{t-d} \leq r \end{cases} \tag{20}$$

gdzie ϕ i θ są współczynnikami regresji, p rzędem procesu AR, i jest progiem, d jest dodatnim opóźnieniem dla progu, ε_t jest procesem losowym idd z zerową średnią i stałą wariancją, θ i ϕ są liczbami rzeczywistymi takimi, że $\theta \neq \phi$. Dla uproszczenia został wykorzystany taki sam rząd autoregresji dla obu reżimów, ale wykorzystane mogą być także różne od siebie liczby (R. S. Tsay i Chen, 2019).

Model ARCH

Model ARCH (AutoRegressive Conditional Heteroskedasticity) opisuje zależności drugiego stopnia w szeregu y_t (niewykazującym autokorelacji) za pomocą funkcji kwadratowej od opóźnionych wartości y_{t-i}. Równania modelu ARCH(q) określone są jako:

$$
\begin{aligned}
y_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2
\end{aligned}
\tag{21}
$$

gdzie ε_t jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie ze średnią zero i wariancją jeden oraz $\omega > 0$ i $\alpha \geq 0$ dla $i = 1, \ldots, q$. Równanie to implikuje zależność skutkującą tym, że duże wartości y_{t-1}^2 pociągają za sobą dużą zmienność σ_t^2. Oznacza to, że w szeregu generowanym przez proces ARCH prawdopodobieństwo wystąpienia po dużej zmianie kolejnych dużych zmian jest większe niż prawdopodobieństwo wystąpienia mniejszych zmian. Ta cecha czyni je szczególnie przydatnymi w modelowaniu szeregów czasowych charakteryzujących się występowaniem skupisk zmienności. (Doman i Doman, 2009)

Test ADF

Test ADF (Augmented Dickey Fuller) jest testem pierwiastka jednostkowego. W zależności od rodzaju, podstawowy test DF (Dickey Fuller) polega na oszacowaniu regresji:

$$
\begin{aligned}
Y_t &= \alpha y_{t-1} + \phi \Delta Y_{t-1} + \varepsilon_t \\
Y_t &= \mu + \alpha y_{t-1} + \phi \Delta Y_{t-1} + \varepsilon_t \\
Y_t &= \mu + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + \varepsilon_t
\end{aligned}
\tag{22}
$$

gdzie ΔY_{t-1} jest pierwszą różnicą dla opóźnienia $t-1$. Hipotezy tego testu zdefiniowane są jako:

H_0: Szereg jest niestacjonarny

H_1: Szereg jest stacjonarny

Test ADF jest rozszerzeniem testu DF do:

$$
Y_t = \mu + \beta t + \alpha y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \varepsilon_t \tag{23}
$$

Dla testu ADF hipotezy są takie same co dla testu DF. Odrzucenie hipotezy zerowej oznacza brak pierwiastka jednostkostkowego, a zatem stacjonarność szeregu (Prabhakaran, 2022).

Test KPSS

Test KPSS (Kwiatkowski–Phillips–Schmidt–Shin) jest również testem pierwiastka jednostkowego. Ideą testu jest:

$$
\begin{aligned}
Y_t &= Y_t = X_t + Z_t \\
X_t &= X_{t-1} + V_t, \qquad V_t \sim WN(0, \sigma^2) \\
Z_t &= \mu + \varepsilon_t
\end{aligned}
\tag{24}
$$

Przy prawdziwości H_0 wariancja σ_v^2 jest zerowa, zaś wartości zmiennej X_t są stałe w czasie (**SGH.**). Hipotezy tego testu zdefiniowane są jako:

H_0: Szereg jest stacjonarny

H_1: Szereg jest niestacjonarny

Test ACF i PACF

Test ACF (AutoCorrelation Function) jest prostym testem sprawdzającym autokorelację szeregu czasowego. Określony jest wzorem:

$$
r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{25}
$$

Test PACF (Partial AutoCorrelation Function) również testuje autokorelację szeregu, natomiast pozbywa się on informacji pomiędzy momentem t a $t-k$. Estymuje się go zazwyczaj za pomocą równań Yule-Walkera lub KMNK (istnieje więcej metod) (Lewinson, 2022).

Test Ljunga-Boxa

Test Ljunga-Boxa sprawdza istotność statystyczną autokorelacji rzędów od t do $t-m$ (e. S. Tsay, 2005). Statystyka testowa określona jest wzorem:

$$
Q(m) = T(T+2) \sum_{\iota=1}^{m} \frac{\hat{\rho}_\iota^2}{T-\iota} \tag{26}
$$

Hipotezy tego testu zdefiniowane są jako:

H_0: $\rho_1 = \cdots = \rho_m = 0$

H_1: $\rho_i \neq 0$

Test Jarque-Bera

Test Jarque-Bera jest testem na normalność rozkładu. Jego statystyka testowa oparta jest na 3 i 4 momencie rozkładu i opisana jest wzorem:

$$JB = \frac{n}{6}(S^2 + \frac{1}{4}(K-3)^2) \tag{27}$$

gdzie S jest współczynnikiem asymetrii próbki a K jej kurtozą. Hipotezy zdefiniowane są jako:

H_0: Rozkład jest normalny

H_1: Rozkład nie jest normalny

Test Shapiro-Wilka

Test Shapiro-Wilka jest alternatywnym testem, który nie bierze pod uwagę 3 i 4 momentu. Jego statystyka wyrażona jest wzorem:

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{28}$$

gdzie parametr a brany jest z zewnętrznych tablic Shapiro-Wilka uwarunkowanych parametrem n (Javier Fernandez, 2022). Hipotezy tego testu zdefiniowane są jako:

H_0: Rozkład jest normalny

H_1: Rozkład nie jest normalny

Test Kołmogorowa-Smirnowa

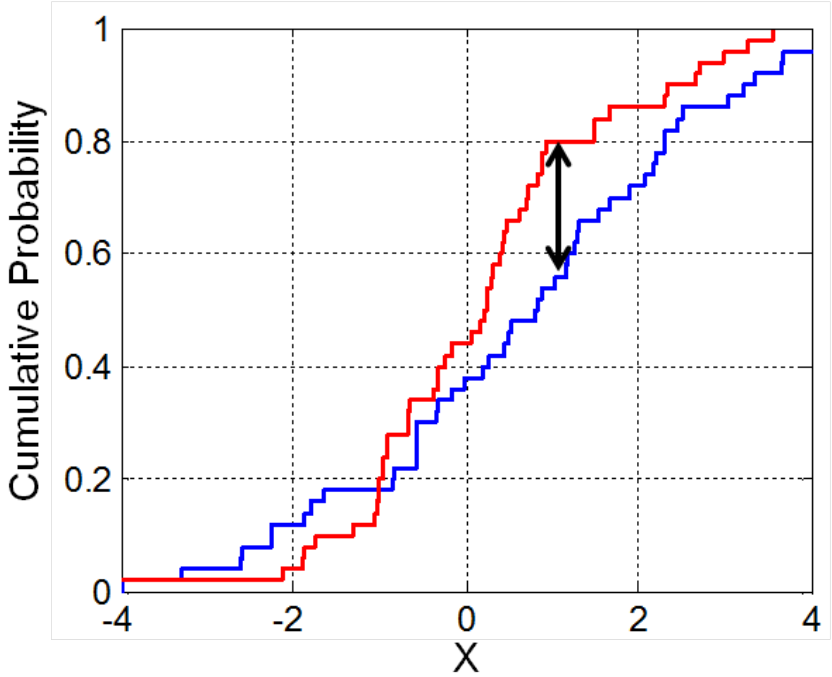
Test Kołmogorowa-Smirnowa jest nieparametrycznym testem dobroci dopasowania. W zależności od podstawionej dystrybucji, może pełnić również rolę testowania normalności rozkładu (Javier Fernandez, 2022). Statystyką testową jest największy dystans pomiędzy rozkładem teoretycznym a empirycznym:

$$D_{n,m} = sup|F_{1,n}(x) - F_{2,m}(x)| \tag{29}$$

Hipotezy tego testu rozumianego jako test na normalność zdefiniowane są jako:

H_0: Rozkład jest normalny

H_1: Rozkład nie jest normalny



Rysunek 2.5: Wizualizacja testu Kołmogorowa-Smirnowa. Źródło: Wikipedia

2.3 Metryki błędu

Do prawidłowego postawienia prognozy należy wybrać pewną metrykę błędu. Zależy ona głównie od oczekiwań prognosty – czasami wymagania postawione w danym procesie decyzyjnym są różne. Celem może być zminimalizowanie prognoz bardzo odstających od wartości rzeczywistej, ale czasem nie mają one większego znaczenia i liczy się długookresowy efekt. Wybrane metryki błędów to:

ME

ME (Mean Error) jest to średnia arytmetyczna błędów. Pokazuje ona, w którą stronę i o ile średnio prognoza się myli.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) \tag{30}$$

MSE

MSE (Mean Squared Error) - jest to metryka odpowiadająca na pytanie jaki średni kwadrat błędu jest uzyskiwany dla danej prognozy.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} \tag{31}$$

RMSE

RMSE (Root Mean Squared Error) - jest to metryka będąca pierwiastkiem MSE. Odpowiada na pytanie jakie jest średnie odchylenie prognozy od danej w tych samych jednostkach. Penalizuje ona wartości bardzo odstające.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{32}$$

MAE

MAE (Mean Absolute Error) - jest to średni absolutny błąd.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{33}$$

MAPE

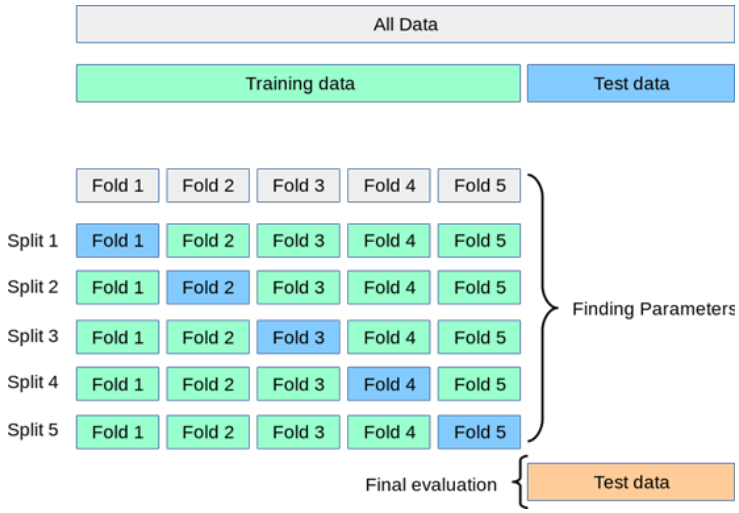
MAPE (Mean Absolute Percentage Error) - jest to średni absolutny błąd procentowy. Odpowiada on na pytanie o ile procent przeciętnie prognozy odchylają się od danych empirycznych.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{y_i} \tag{34}$$

2.4 Rodzaje walidacji

Algorytmy uczenia maszynowego są bardzo podatne na przeuczenie. Jest to sytuacja, w której predykcje modelu są zbyt dokładne, a wytrenowany model staje się tak zawiły, że przestaje generalizować wiedzę, a zaczyna uczyć się także składnika losowego zapamiętując dane, na których został opracowany. Charakterystyczne do tej sytuacji są bardzo niskie funkcje błędu w predykcjach w próbce (in-sample), które jednak stają się bardzo wysokie dla zbioru testowego (out-of-sample) (IBM, 2022). Aby zapobiec takiej sytuacji wykorzystuje się techniki walidacji.

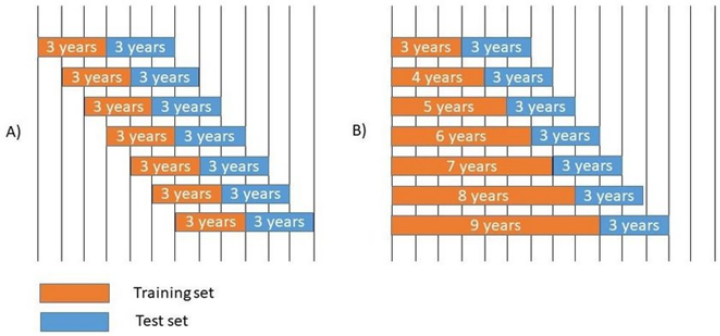
Najczęściej wykorzystywaną dzisiaj techniką jest K-Fold Cross-Validation – metoda ta polega na podzieleniu zbioru treningowego na K podzbiorów, spośród których każdy musi być dokładnie 1 raz zbiorem walidacyjnym, natomiast reszta z nich musi zostać wykorzystana do wytrenowania modelu. Następnie dla każdej kombinacji parametrów modelu obliczana jest średnia arytmetyczna i wybierany jest ten model, w którym wartość funkcji błędu jest najmniejsza.



Rysunek 2.6: Wizualizacja algorytmu K-Fold Cross-Validation
Źródło: (Sklearn, 2022)

Używanie standardowych algorytmów walidacji do szeregów czasowych nie jest jednak z reguły preferowaną metodą. Jest to spowodowane tym, że szeregi czasowe nie zawsze spełniają założenia idd, istnieje w nich często seryjna zależność. Ponadto istnieje ryzyko wykorzystania informacji z przyszłości, aby znaleźć lepsze dopasowanie parametrów do próbek z przeszłości (Packt Editorial Staff, 2022). Aby temu przeciwdziałać w niniejszej pracy zaimplementowany został inny rodzaj walidacji polegający na umieszczeniu ruchomego okna na danych w taki sposób, aby dany model nie miał dostępu do danych z przyszłości w trakcie estymacji. Istnieją dwa ogólne rodzaje algorytmu Rolling Window Cross-Validation:

1. z stałą długością zestawu treningowego, ale ruchomego w czasie,
2. z zwiększającą się wraz z każdą kolejną iteracją algorytmu długością zbioru treningowego.



Rysunek 2.7: Wizualna reprezentacja algorytmu Rolling Window. A) Walidacja o stałym oknie treningowym. B) Walidacja o rozszerzającym się oknie treningowym.

Źródło: (Shojaei i Flood, 2018)

2.5 Metodyka badania

Niniejsza praca ma na celu odpowiedź na pytanie: Czy wybrane modele uczenia maszynowego są w stanie dostarczyć podobnej jakości prognozy w stosunku do modeli ekonometrycznych? Aby na nie odpowiedzieć, zostaną przeprowadzone trzy badania:

1. W pierwszym badaniu zostanie wygenerowany proces SETAR(2, 1, 1).
2. W drugim badaniu wykorzystane zostaną dane populacji rysiów w Kanadzie.
3. W trzecim badaniu wykorzystane zostaną dane giełdowe wolumenu indeksu SP500.

W każdym przypadku porównane zostaną modele: SETAR, KNN, drzewo decyzyjne i las losowy. W dwóch pierwszych przypadkach porównany zostanie model liniowy AR, natomiast do danych giełdowych zostanie zaimplementowany model ARMA - ze względu skomplikowane prawa rządzące tego typu szeregami.

Zaimplementowany proces walidacji modeli opartych na uczeniu maszynowym podzielić można na etapy:

1. Podzielenie wszystkich danych na zestaw treningowy i testowy.
2. Wyodrębnienie z zestawu treningowego ruchomego okna o stałej długości od pierwszej obserwacji.
3. Wytrenowanie modelu ograniczonego parametrami.
4. Postawienie prognozy na 1 obserwację do przodu i zapisanie jej.
5. Przesunięcie ruchomego okna o 1 obserwację do przodu.
6. Powtórzenie pkt. 3,4 i 5 aż do końca zestawu treningowego
7. Zmiana hiperparametrów.
8. Powtórzenie pkt. 6 aż cała procedura wykonana zostanie dla wszystkich parametrów.
9. Selekcja parametrów modelu, który posiada najlepszą wartość błędu RMSE.

Ze względu na różną długość wektora wartości błędów modeli uczenia maszynowego i modeli ekonometrycznych spowodowaną specyficzną formą walidacji, nie zostaną porównane predykcje (in-sample) tych dwóch klas modeli, końcowe wnioski wyciągnięte zostaną na podstawie zbioru testowego. Aby uniknąć sytuacji, w której modele uczenia maszynowego dostałyby więcej informacji niż modele ekonometryczne, zostało przyjęte założenie, że opóźnienie dla modeli KNN, drzewa decyzyjnego i lasu losowego jest równe opóźnieniu maksymalnemu któregokolwiek z modeli ekonometrycznych.

Dla wszystkich reszt z modeli przeprowadzona zostanie weryfikacja statystyczna. Zarówno predykcje jak i prognozy zostaną porównane wybranymi metrykami błędu, natomiast jako decydujące kryterium zostało wybrane RMSE.

3 Badanie danych symulowanych

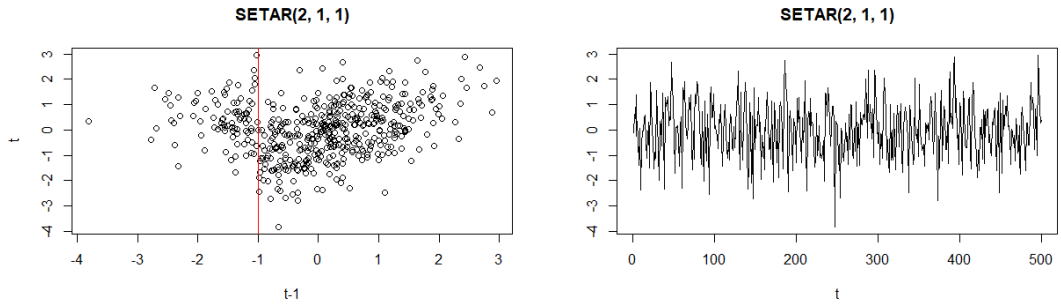
3.1 Prognozowanie symulowanego procesu SETAR

Charakterystyka szeregu

Wygenerowany został proces SETAR(2, 1, 1) o parametrach:

$$\begin{cases} y_t = -0,3 + 0,6y_{t-1}, & y_{t-1} > -1 \\ y_t = 0,8 + 0,2y_{t-1}, & y_{t-1} \leq -1 \end{cases} \tag{35}$$

Dane te zostały podzielone na dwa podzbiory: treningowy oraz testowy w proporcjach 8:2. Pierwszy z nich zostanie wykorzystany do przeprowadzenia analizy statystycznej oraz oszacowania wszystkich modeli, natomiast drugi będzie przeznaczony do policzenia prognozy i jej oceny.



Rysunek 3.1: Wykres rozrzutu dla 1 opóźnienia wygenerowanego szeregu oraz wykres samego szeregu.

Na początek zbadana zostanie stacjonarność szeregu za pomocą testów ADF i KPSS:

	ADF	KPSS
Statystyka	-6.5604	0.066535
P-value	0.01	0.1

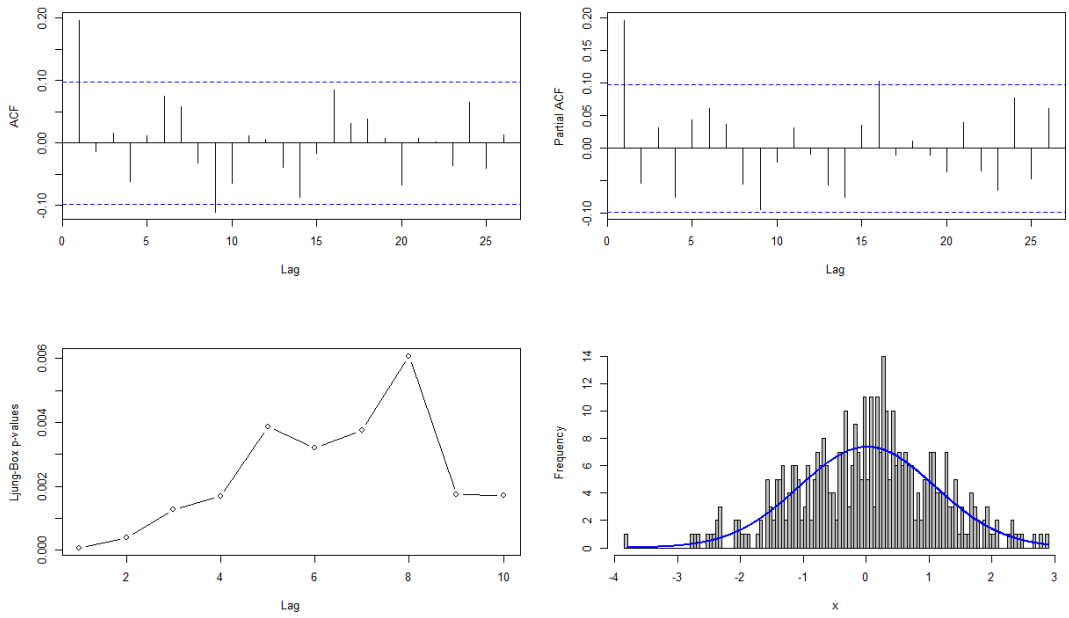
Tablica 3.1: Statystyki i p-values testów ADF i KPSS

Oba testy potwierdziły stacjonarność wygenerowanego szeregu. Następnie zbadana zostanie jego dystrybucja za pomocą testu Jarque-Bera, Shapiro-Wilka i Kołmogorowa Smirnowa:

	Statystyka	P-value
Jarque-Bera	1.5586	0.4587
Shapiro-Wilk	0.99673	0.5983
Kołmogorow-Smirnow	0.039627	0.5563

Tablica 3.2: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

P-values wszystkich trzech testów jest większe od $\alpha = 0.05$, a zatem nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Zbadany zostanie histogram, autokorelacja oraz cząstkowa autokorelacja szeregu:



Rysunek 3.2: Wyniki testów ACF, PACF (górny rząd) oraz p-values Ljunga-Boxa i histogram szeregu.

Zarówno test ACF jak i PACF wykazał istotną autokorelację dla opóźnienia równego 1. Potwierdzone to zostaje za pomocą testu Ljunga-Boxa - wartości p-values są dla wszystkich badanych rzędów mniejsze od wartości $\alpha = 0.05$, zatem można odrzucić hipotezę zerową o braku autokorelacji szeregu. Histogram szeregu wskazuje na rozkład normalny.

3.2 Szacowanie modeli ekonometrycznych

Szacowanie modelu AR

Ponieważ test PACF wykazał jeden rząd opóźnienia z istotną cząstkową autokorelacją, estymowany model jest modelem AR(1):

Parametr	Błąd	standardowy	Statystyka	P-value
ar1	0.198237	0.048908	4.0533	5.051e-05***

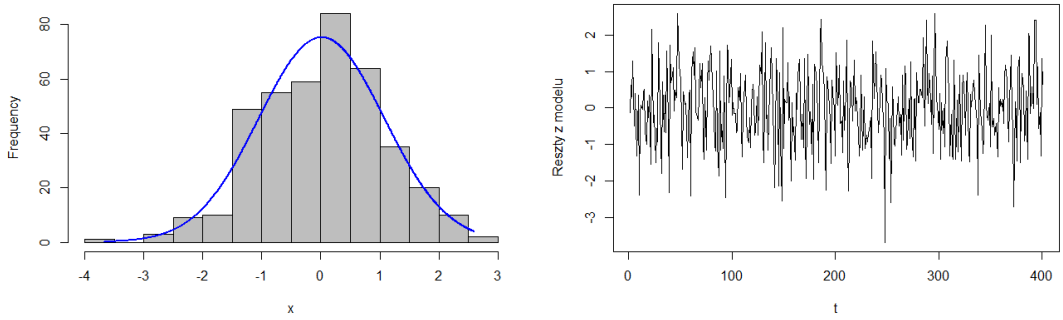
Tablica 3.3: Oszacowane parametry modelu AR(1), jego błąd standardowy, statystyka Z oraz p-value

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-6.552	0.06671
P-value	0.01	0.1

Tablica 3.4: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana dystrybucja reszt tego modelu:

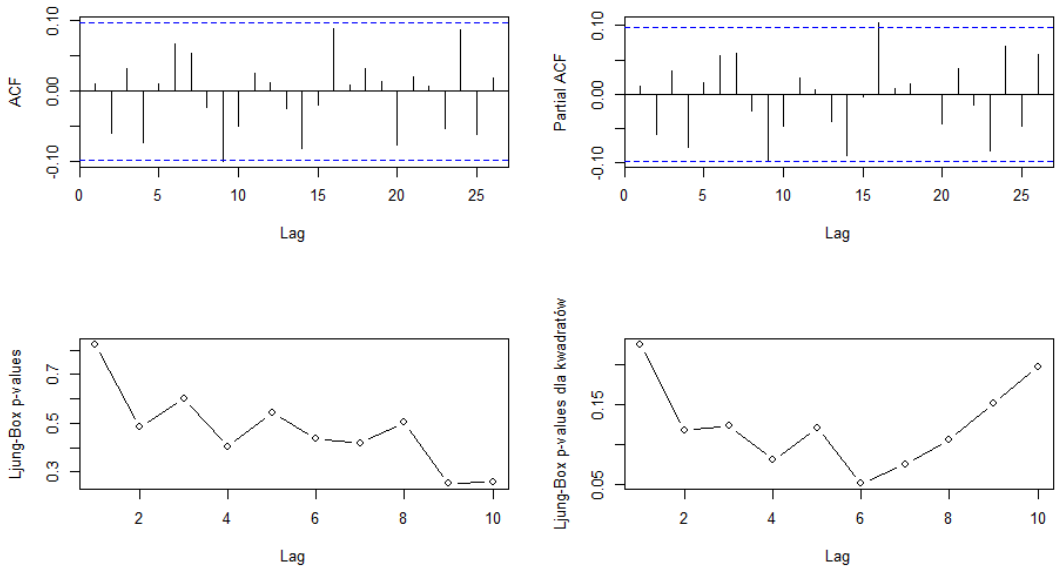


Rysunek 3.3: Histogram reszt modelu AR i ich wykres w zależności od czasu.

	Statystyka	P-value
Jarque-Bera	1.2115	0.5457
Shapiro-Wilk	0.9959	0.3851
Kołmogorow-Smirnow	0.047579	0.3241

Tablica 3.5: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

Histogram reszt z modelu wygląda jak rozkład normalny, a p-values wszystkich trzech testów jest większe od $\alpha = 0.05$, a zatem nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Wykres reszt nie wskazuje na zachowanie jakiejkolwiek struktury, wskazują na rozkład niezależny od czasu. Zbadana zostanie następnie autokorelacja oraz cząstkowa autokorelacja szeregu:



Rysunek 3.4: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Żaden test nie wykazał istotnej statystycznie autokorelacji reszt modelu. Wszystkie zależności liniowe szeregu zostały wyjaśnione za pomocą modelu AR(1). Ponadto reszty nie wykazują efektu ARCH.

Szacowanie modelu SETAR

Specyfikacja modelu SETAR jest bardziej złożona od procedur związanych z podstawowym modelem AR, a ponieważ dane na których wykonywana jest analiza zostały wygenerowane, dla uproszczenia maksymalne opóźnienie zostanie ograniczone do $d = 1$. Oszacowany w ten sposób został model SETAR(2, 1, 1):

	Parametr	Błąd standardowy	Statystyka t	P-value
$\phi_{0,0}$	0.263171	0.347529	0.7573	0.4493
$\phi_{0,1}$	-0.070983	0.206828	-0.3432	0.7319
$\phi_{1,0}$	-0.314460	0.058000	-5.4217	1.028e-07 ***
$\phi_{1,1}$	0.669685	0.064682	10.3535	« 2.2e-16 ***
r	-1.004	-	-	-
d	1	-	-	-

Tablica 3.6: Oszacowane parametry modelu SETAR, ich błędy standardowe, statystyki t i p-values

Gdzie zbiór parametrów ϕ odwzorowany jest jako $\phi_{I(y_{t-1}),p_q}$ dla:

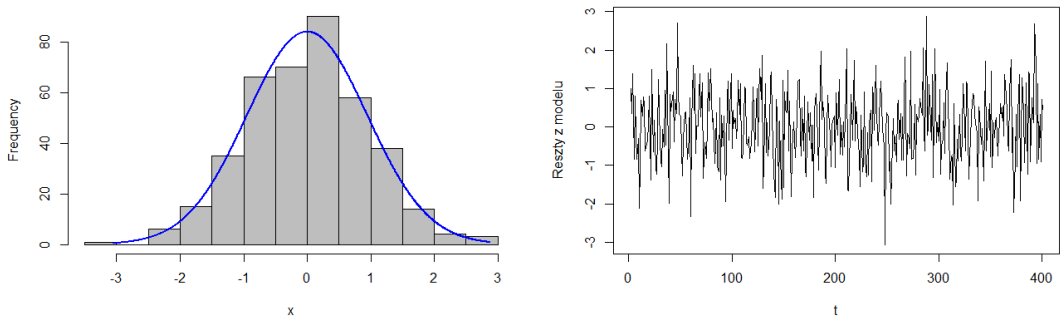
$$\begin{cases} I(N) = 0 \text{ dla } y_{t-d} \leq r \\ I(N) = 1 \text{ dla } y_{t-d} > r \end{cases} \tag{36}$$

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-6.6884	0.062285
P-value	0.01	0.1

Tablica 3.7: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana dystrybucja reszt tego modelu:

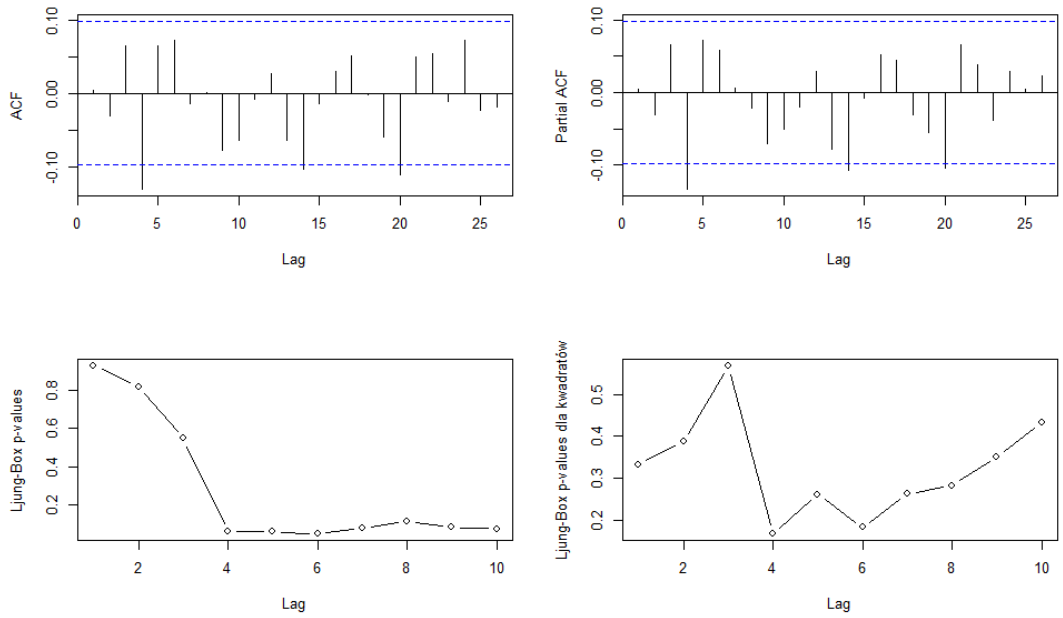


Rysunek 3.5: Histogram reszt modelu SETAR i ich wykres w zależności od czasu.

	Statystyka	P-value
Jarque-Bera	0.088362	0.9568
Shapiro-Wilk	0.99798	0.9189
Kołmogorow-Smirnow	0.028467	0.9029

Tablica 3.8: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

Histogram szeregu wygląda wizualnie jak rozkład normalny, a p-values wszystkich trzech testów jest większe od $\alpha = 0.05$, a zatem nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Wykres reszt nie wskazuje na występowanie żadnych zależności ani heteroskedastyczności. Zbadana zostanie następnie autokorelacja oraz cząstkowa autokorelacja szeregu:

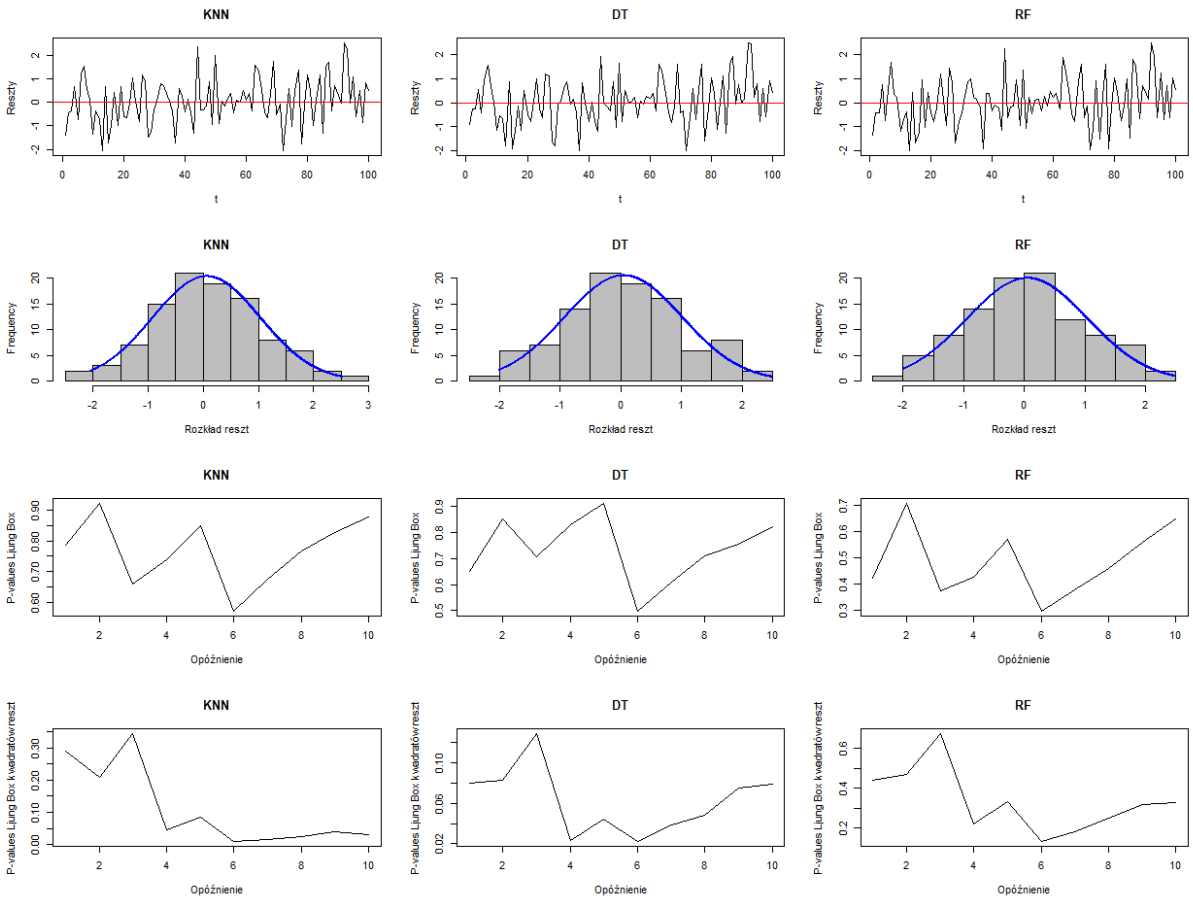


Rysunek 3.6: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Testy ACF i PACF wykazały istotne autokorelacje składnika resztowego 4 rzędu, natomiast nie zostało to potwierdzone za pomocą testu Ljunga-Boxa (p-value jest równe 0.0587, zatem jest większe od poziomu istotności $\alpha = 0.05$). Pierwsze p-value testu Ljunga-Boxa mniejsze od $\alpha = 0.05$ jest na szóstym opóźnieniu, co nie ma odzwierciedlenia w testach ACF i PACF, zatem uznane zostaje, że szereg nie wykazuje autokorelacji. Ponadto reszty nie wykazują efektu ARCH.

3.3 Szacowanie modeli uczenia maszynowego

Przeanalizowane zostały reszty ze wszystkich trzech modeli:



Rysunek 3.7: W kolejnych rzędach: wykresy reszt w zależności od czasu, histogramy, p-values testu Ljunga-Boxa reszt oraz kwadratów reszt

W przypadku wszystkich trzech modeli wykresy reszt nie wskazują na zachowanie się zależności, a ich rozkłady przypominają wizualnie rozkład normalny - zostanie to zweryfikowane za pomocą dalszych testów. Wszystkie p-values testu Ljunga-Boxa są większe niż $\alpha = 0.05$, zatem nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji reszt oraz kwadratów reszt. Szeregi nie są autoskorelowane liniowo i nie posiadają efektu ARCH. Przeprowadzono szereg testów na normalność składnika resztowego:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	0.51354	0.7735	0.99223	0.8376	0.053765	0.9347
DT	0.4625	0.7935	0.98907	0.5905	0.059443	0.8716
RF	0.96406	0.6175	0.99051	0.7059	0.058158	0.8877

Tablica 3.9: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

W przypadku wszystkich modeli p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-

Smirnowa są większe od $\alpha = 0.05$, zatem nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Rozkład szeregu zostaje uznany za rozkład normalny. W następnym kroku zbadana zostanie jego stacjonarność:

	Statystyka ADF	P-value ADF	Statystyka KPSS	P-value KPSS
DT	-4.7492	0.01	0.57971	0.02448
RF	-4.9318	0.01	0.4818	0.04576
KNN	-4.8956	0.01	0.53148	0.03458

Tablica 3.10: Statystyki i p-values testów ADF i KPSS

P-values wszystkich modeli dla testów ADF oraz KPSS jest mniejsze od $\alpha = 0.05$, zatem błędy zostają uznane za niestacjonarne. Aby porównać dopasowanie modeli policzone zostały funkcje błędu:

	KNN	DT	RF
ME	0.0596	0.0455	**0.04011**
MSE	0.97755	**0.96846**	**1.00405**
RMSE	0.95561	**0.93791**	1.00811
MAE	0.77498	**0.75631**	0.81585
MAPE	5.83257	**4.46139**	10.18543

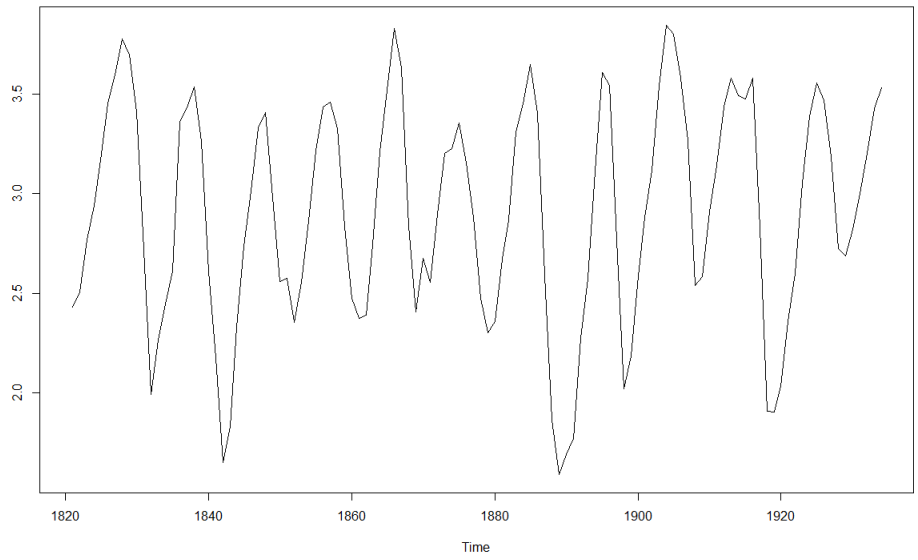
Tablica 3.11: Wartości funkcji błędów dla reszt z predykcji modeli uczenia maszynowego

Najmniejszy błąd RMSE posiada model DT, zatem zostaje on uznany za najlepiej dopasowany spośród modeli uczenia maszynowego.

4 Prognozowanie danych rzeczywistych

4.1 Charakterystyka szeregu

Do zamodelowania danych rzeczywistych, charakteryzujących się własnościami nieliniowymi wybrane zostały dane ”Annual Canadian lynx trappings”. Zostały one podzielone w proporcjach 7:3 na podzestaw treningowy oraz testowy, oraz wyciągnięty został z nich log_{10}:



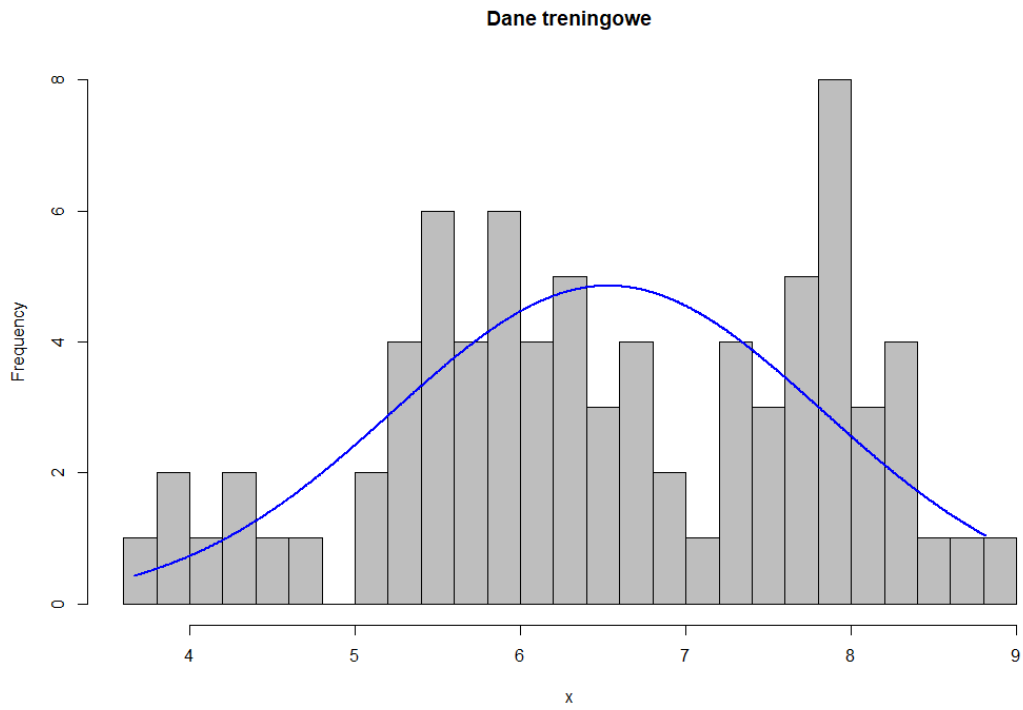
Rysunek 4.1: Wykres szeregu.

Na początek zbadana zostanie stacjonarność szeregu za pomocą testów ADF i KPSS:

	ADF	KPSS
Statystyka	-4.5378	0.089086
P-value	0.01	0.1

Tablica 4.1: Statystyki i p-values testów ADF i KPSS

Oba testy potwierdziły stacjonarność wygenerowanego szeregu. Następnie zbadana zostanie jego dystrybucja za pomocą histogramu, testu Jarque-Bera, Shapiro-Wilka i Kołmogorowa Smirnowa:

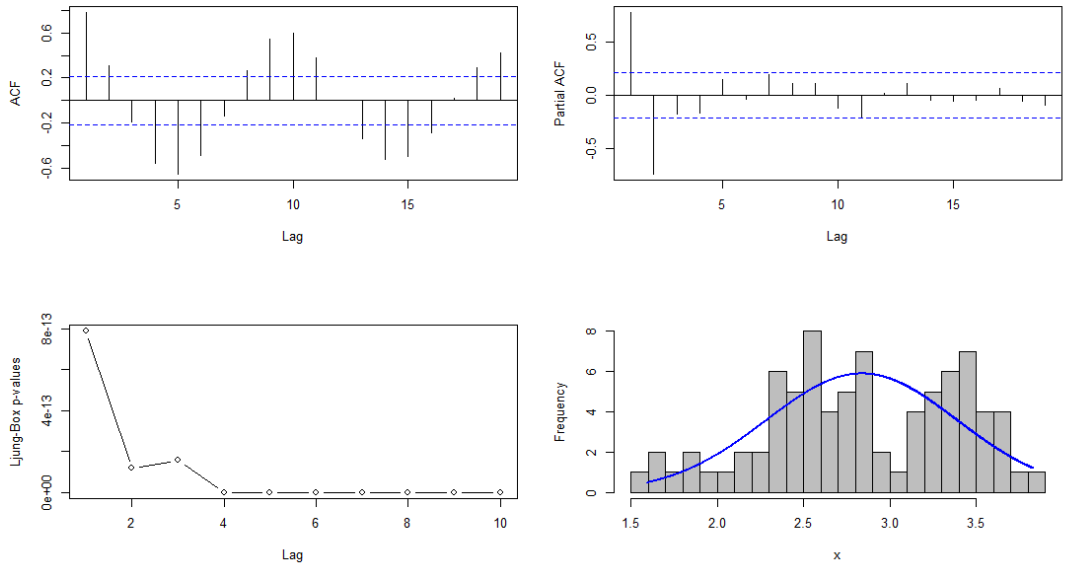


Rysunek 4.2: Histogram szeregu.

	Statystyka	P-value
Jarque-Bera	2.6562	0.265
Shapiro-Wilk	0.97023	0.05342
Kołmogorow-Smirnow	0.9442	< 2.2e-16

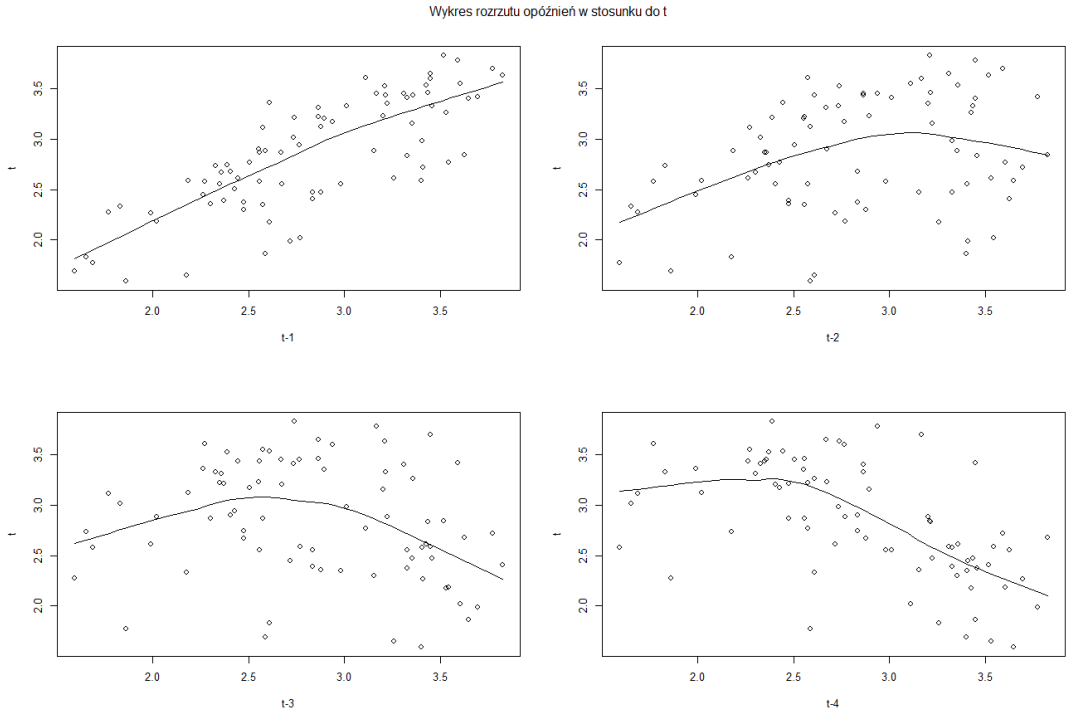
Tablica 4.2: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

Histogram szeregu wizualnie nie przypomina rozkładu normalnego. W przypadku testu Jarque-Bera oraz Shapiro-Wilka p-value jest większe od $\alpha = 0,05$, jednak w przypadku testu Kołmogorowa p-value jest mniejsze od $\alpha = 0,05$, a więc odrzucona jest hipoteza zerowa o normalności rozkładu.



Rysunek 4.3: Wyniki testów ACF, PACF (górny rząd) oraz p-values Ljunga-Boxa i histogram szeregu.

Test PACF wykazał istotną cząstkową autokorelacja dla opóźnień 1 i 2, a test ACF również dla opóźnień 1 i 2, występuje bardzo wyraźna cykliczność istotnych autokorelacji wyższych rzędów. Dodatkowo potwierdzone to zostało za pomocą testu Ljunga-Boxa. Ponieważ p-values do 10 rzędu są mniejsze od $\alpha = 0,05$, można odrzucić hipotezę zerową o braku autokorelacji. Szereg jest autoskorelowany. Sprawdzono również autokorelację kwadratów tego szeregu - p-values do 10 rzędu są mniejsze od $\alpha = 0,05$, zatem odrzucono hipotezę zerową o braku autokorelacji kwadratów tego szeregu, występuje efekt ARCH.



Rysunek 4.4: Wykresy rozrzutu szeregu czasowego w momencie od t do $t - i$ z nałożoną linią regresji lokalnej.

Z wykresów rozrzutu wynika, że od drugiego opóźnienia występują bardzo wyraźne nieliniowości.

4.2 Szacowanie modeli ekonometrycznych

Szacowanie modelu AR

Ponieważ test PACF wykazał dwa rzędy opóźnienia z istotną cząstkową autokorelacją, estymowany model jest modelem AR(2):

	Parametr	Błąd standardowy	Statystyka z	P-value
ar1	1.364877	0.070261	19.426	<2.2e-16 ***
ar2	-0.748450	0.069374	-10.789	<2.2e-16 ***
intercept	2.840894	0.064689	43.916	<2.2e-16 ***

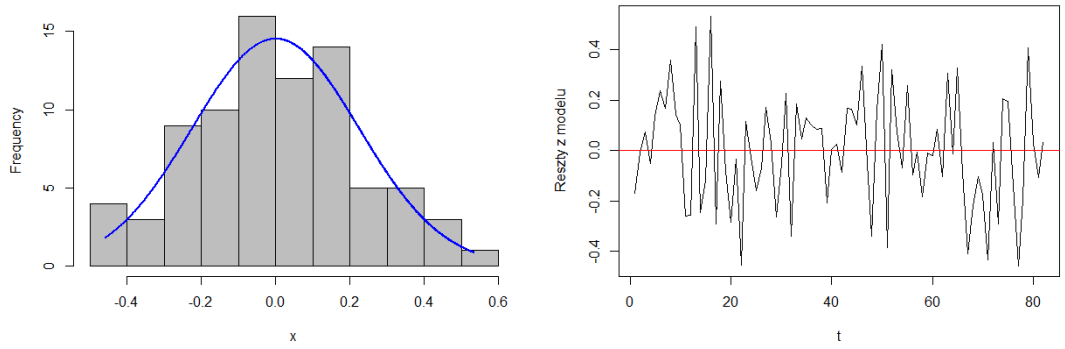
Tablica 4.3: Oszacowane parametry modelu AR(2), jego błędy standardowy, statystyki Z oraz p-values

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-3.8966	0.16972
P-value	0.01854	0.1

Tablica 4.4: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana dystrybucja reszt tego modelu:



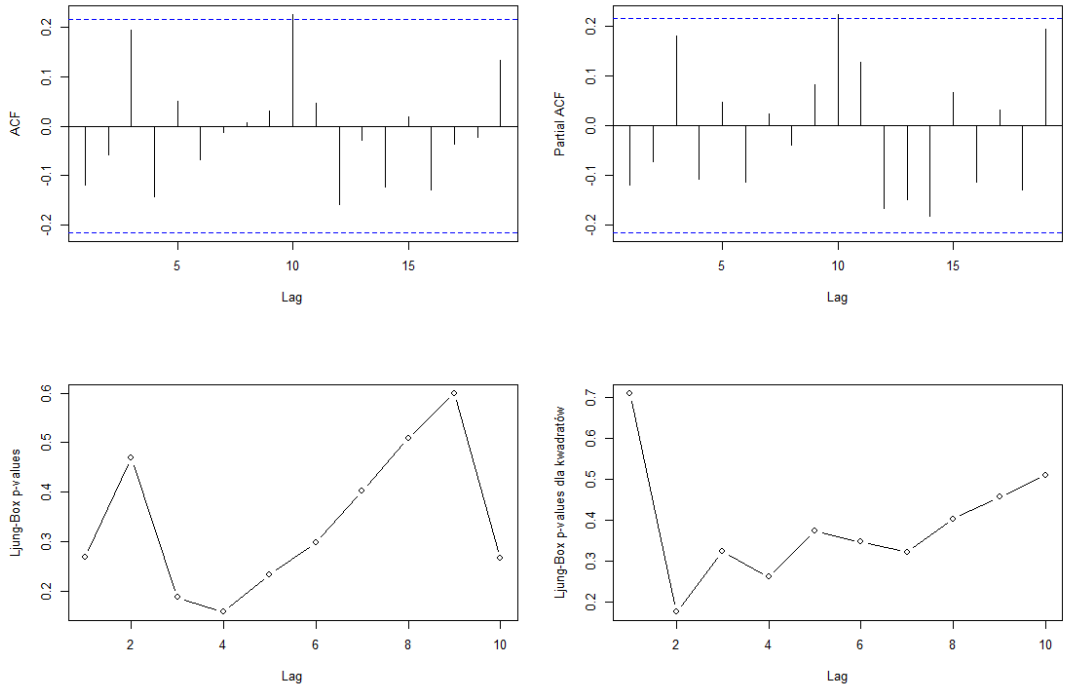
Rysunek 4.5: Histogram reszt modelu AR i ich wykres w zależności od czasu.

Rozkład reszt przypomina rozkład normalny, jednak widać w nim dwa szczyty. Zostanie to rozstrzygnięte za pomocą testów. Wykres reszt wskazuje na losowy rozkład reszt względem czasu.

	Statystyka	P-value
Jarque-Bera	0.54187	0.7627
Shapiro-Wilk	0.99184	0.8901
Kołmogorow-Smirnow	0.32314	4.037e-08

Tablica 4.5: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

W przypadku testów Jarque-Bera oraz Shapiro-Wilka nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu, jednak p-value testu Kołmogorowa-Smirnowa jest mniejsze od $\alpha = 0.05$, zatem w jego przypadku odrzucona zostaje hipoteza zerowa o normalności rozkładu. Rozkład reszt modelu AR(2) nie jest rozkładem normalnym. Zostanie następnie zbadana autokorelacja reszt oraz kwadratów reszt tego modelu:



Rysunek 4.6: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Żaden test nie wykazał istotnej statystycznie autokorelacji. Szereg nie wykazuje także efektu ARCH.

Szacowanie modelu SETAR

Oszacowany został model SETAR(2,3,3) o parametrach:

	Parametr	Błąd standardowy	Statystyka z	P-value
$\phi_{0,const}$	0.906115	0.207111	4.3750	3.917e-05 ***
$\phi_{0,1}$	0.973434	0.135917	7.1620	4.898e-10 ****
$\phi_{0,2}$	0.043399	0.187493	0.2315	0.81759
$\phi_{0,3}$	-0.289062	0.110648	-2.6124	0.01088 *
$\phi_{1,const}$	0.471476	0.712126	0.6621	0.50998
$\phi_{1,1}$	1.688850	0.167842	10.0622	1.688e-15 ***
$\phi_{1,2}$	-1.528838	0.275226	-5.5549	4.152e-07 ***
$\phi_{1,3}$	0.637023	0.267472	2.3816	0.01981 *
r	2.981	-	-	-
d	2	-	-	-

Tablica 4.6: Oszacowane parametry modelu SETAR, ich błędy standardowe, statystyki t i p-values

Gdzie zbiór parametrów ϕ odwzorowany jest jako $\phi_{I(y_{t-1}),p_q}$ dla:

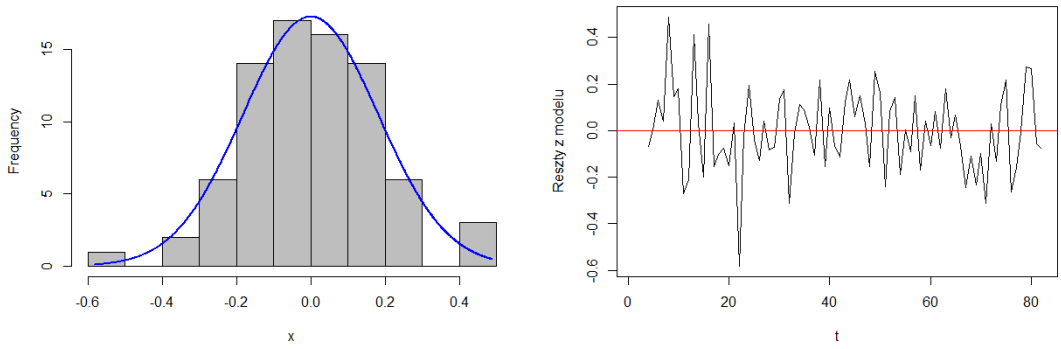
$$\begin{cases} I(N) = 0 \text{ dla } y_{t-d} \leq r \\ I(N) = 1 \text{ dla } y_{t-d} > r \end{cases} \tag{37}$$

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-3.5248	0.13232
P-value	0.04543	0.1

Tablica 4.7: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana dystrybucja reszt tego modelu:



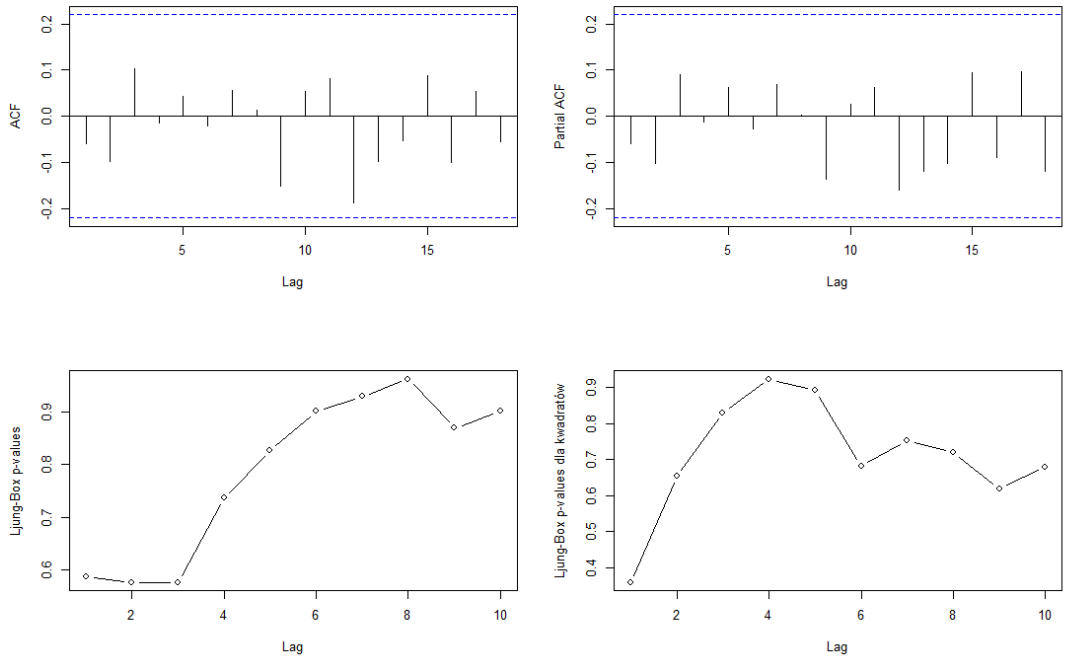
Rysunek 4.7: Histogram reszt modelu SETAR i ich wykres rozrzutu.

Histogram reszt przypomina wizualnie rozkład normalny, a ich wykres wskazuje na ich niezależność w czasie. Są jednak widoczne pojedyncze wartości odstające:

	Statystyka	P-value
Jarque-Bera	2.0184	0.3645
Shapiro-Wilk	0.98396	0.4257
Kołmogorow-Smirnow	0.36518	5.964e-10

Tablica 4.8: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

W przypadku testów Jarque-Bera oraz Shapiro-Wilka nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Jednak p-value testu Kołmogorowa-Smirnowa jest mniejsze od $\alpha = 0.05$, zatem w jego przypadku odrzucona zostaje hipoteza zerowa o normalności rozkładu. Rozkład reszt modelu SETAR(2,3,3) nie jest rozkładem normalnym.

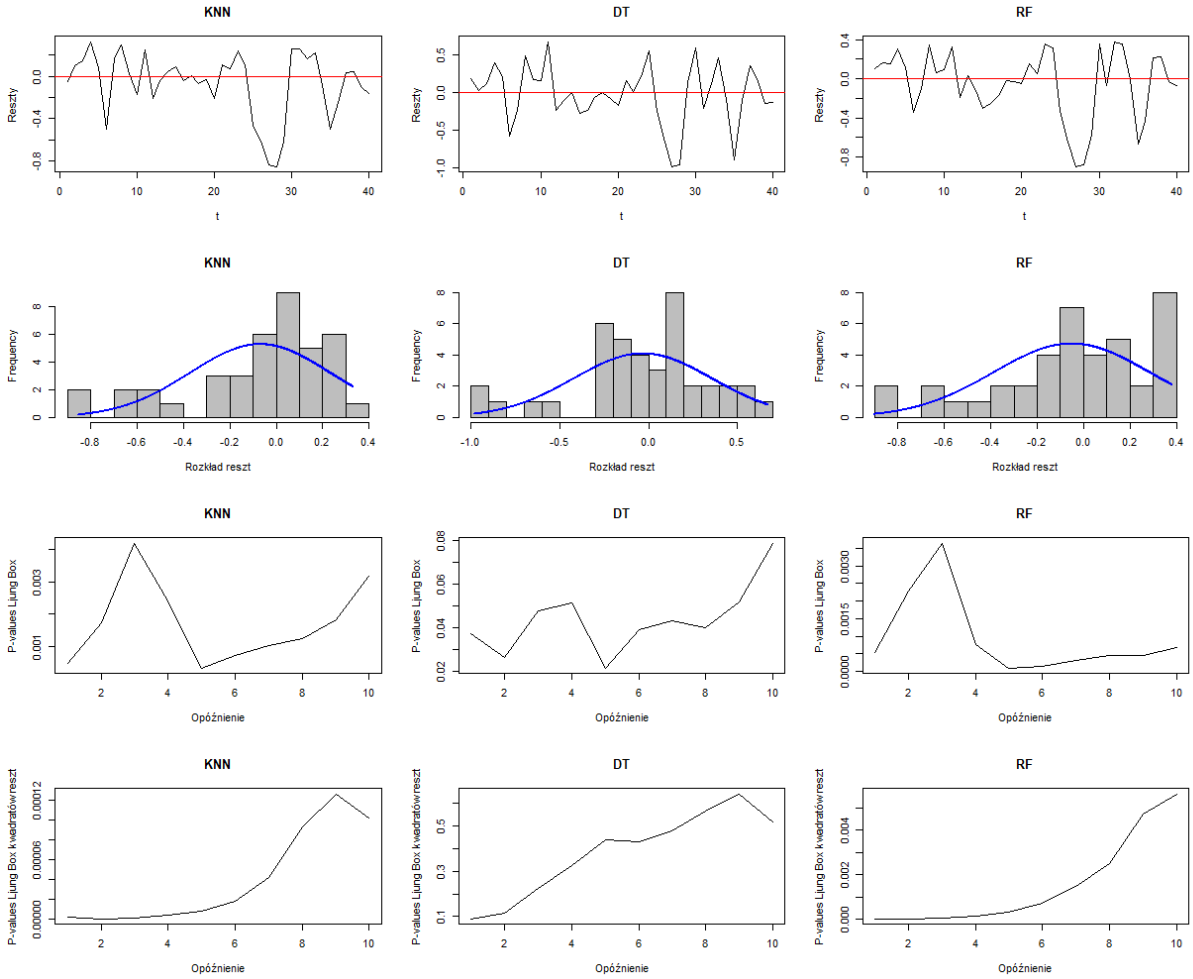


Rysunek 4.8: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Żaden test nie wykazał istotnej statystycznie autokorelacji reszt modelu dla żadnego rzędu, dodatkowo potwierdzone to zostaje za pomocą testu Ljunga-Boxa. P-values testu Ljunga-Boxa kwadratów reszt są większe niż $\alpha = 0.05$ nie wykazują efektu ARCH.

4.3 Szacowanie modeli uczenia maszynowego

Przeanalizowane zostały reszty ze wszystkich trzech modeli:



Rysunek 4.9: W kolejnych rzędach: wykresy reszt w zależności od czasu, histogramy, p-values testu Ljunga-Boxa reszt oraz kwadratów reszt

Reszty wszystkich trzech modeli wykazują wartości odstające, ale nie wskazują one wizualnie na zachowane zależności. Wszystkie rozkłady reszt przypominają wizualnie rozkład normalny, zostanie to zweryfikowane za pomocą dalszych testów. Wszystkie p-values testu Ljunga-Boxa dla reszt i kwadratów reszt dla rzędu równego przynajmniej 1 są mniejsze niż $\alpha = 0.05$, zatem odrzucona zostaje hipoteza zerowa o braku autokorelacji reszt oraz kwadratów reszt. Reszty są autoskorelowane liniowo posiadają efekt ARCH. Przeprowadzono szereg testów na normalność składnika resztowego:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	6.6655	0.03569	0.93657	0.0146	0.30565	0.0002599
DT	2.1211	0.3463	0.95917	0.1064	0.25416	0.004165
RF	3.7312	0.1548	0.95705	0.08798	0.2983	0.0003993

Tablica 4.9: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

W przypadku wszystkich modeli p-values testów Jarque-Bera i Shapiro-Wilka są większe od $\alpha = 0.05$, ale p-value testu Kołmogorowa-Smirnowa jest mniejsze od $\alpha = 0.05$ zatem odrzucona zostaje hipoteza zerowa o nomalności rozkładu. W następnym kroku zbadana zostanie jego stacjonarność:

	Statystyka ADF	P-value ADF	Statystyka KPSS	P-value KPSS
DT	-3.8986	0.01999	0.042716	0.1
RF	-3.6798	0.03711	0.16602	0.1
KNN	-3.422	0.06494	0.19004	0.1

Tablica 4.10: Statystyki i p-values testów ADF i KPSS

P-values wszystkich modeli dla testu ADF są mniejsze od $\alpha = 0.05$, oraz dla testu KPSS są większe od $\alpha = 0.05$, zatem błędy zostają uznane za stacjonarne. Aby porównać dopasowanie modeli policzone zostały funkcje błędu:

	KNN	DT	RF
ME	-0.0164	0.0255	**-0.01565**
MSE	**0.31499**	0.43299	0.34795
RMSE	**0.09922**	0.18748	0.12107
MAE	**0.23705**	0.33226	0.26709
MAPE	**0.09657**	0.13106	0.10795

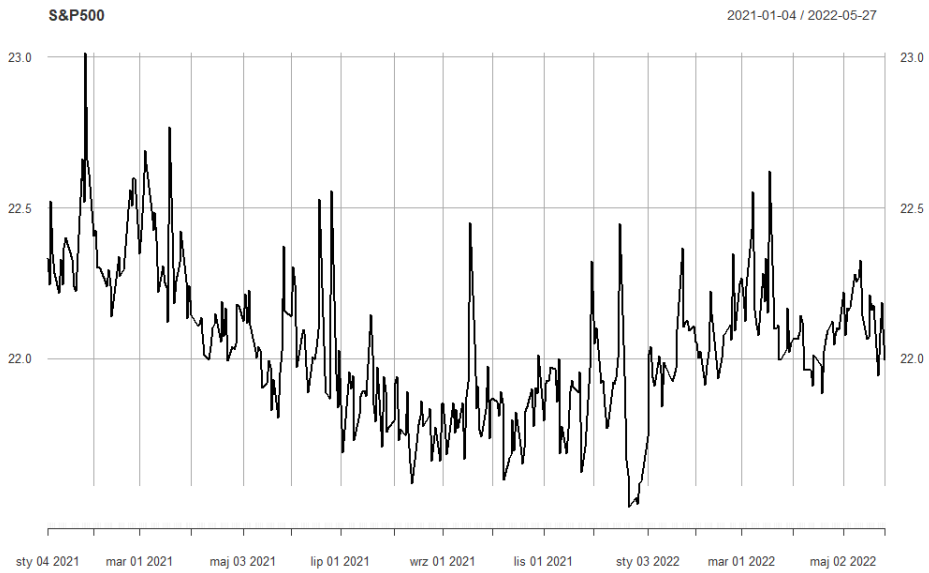
Tablica 4.11: Wartości funkcji błędów dla reszt z predykcji modeli uczenia maszynowego

Najmniejszy błąd RMSE posiada model KNN, zatem zostaje on uznany za najlepiej dopasowany spośród modeli uczenia maszynowego.

5 Prognozowanie danych giełdowych

5.1 Charakterystyka szeregu

Prognozowanie szeregów finansowych jest bardzo złożonym procesem ze względu na ich bardzo wysoką zmienność, często niskie autokorelacje liniowe oraz trudne do wykrycia zależności nieliniowe. Ponieważ wszyscy gracze na giełdzie na bieżąco próbują uzyskać przewagę nad resztą wykorzystując wszystkie dostępne dla nich metody ilościowe i jakościowe, uzyskanie wiarygodnej prognozy cen jest bardzo utrudnione. Do niniejszego badania wykorzystane zostaną zlogarytmowane dzienne dane wolumenu indeksu SP500 w okresie od 4 stycznia 2021 roku do 27 maja 2022 roku:



Rysunek 5.1: Wykres zlogarytmowanego wolumenu SP500.

Dane te zostały podzielone w proporcjach 7:3 na szeregi treningowy i testowy. Na początku zbadana została stacjonarność tego szeregu za pomocą testów ADF i KPSS:

	ADF	KPSS
Statystyka	-3.3868	3.1331
P-value	0.05729	0.01

Tablica 5.1: Statystyki i p-values testów ADF i KPSS

Na poziomie istotności $\alpha = 0,05$ test ADF wskazuje na niestacjonarność szeregu, dlatego w dalszym kroku sprawdzona zostanie stacjonarność przyrostów:

	ADF	KPSS
Statystyka	-8.869	0.016156
P-value	0.01	0.1

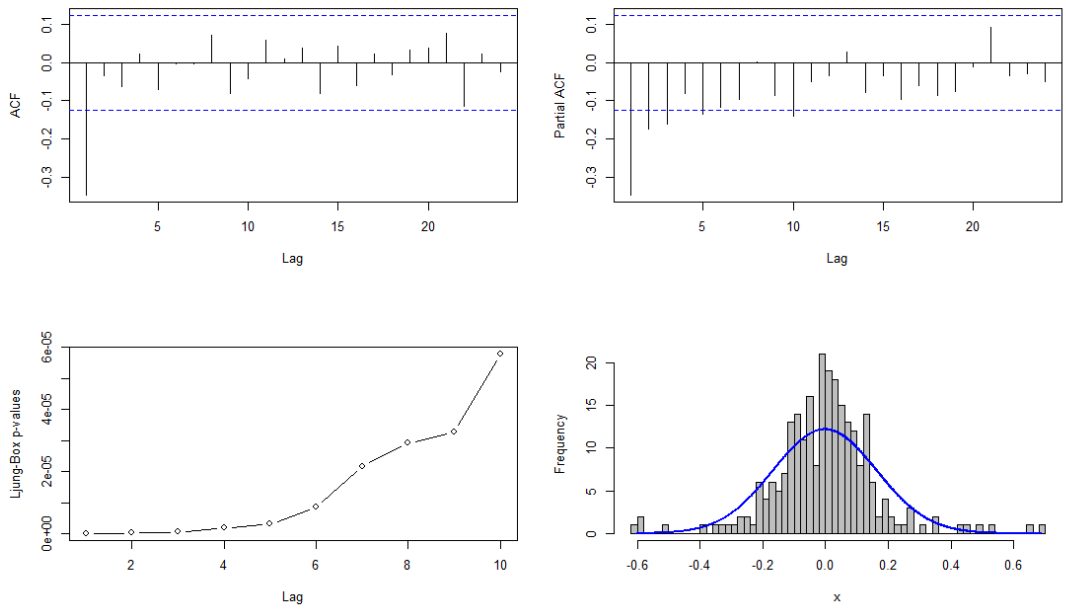
Tablica 5.2: Statystyki i p-values testów ADF i KPSS

Oba testy jednoznacznie wykazały, że przyrosty analizowanego szeregu są stacjonarne. W takim przypadku zbadana zostanie ich rozkład za pomocą histogramu, testu Jarque-Bera, Shapiro-Wilka i Kołmogorowa Smirnowa:

	Statystyka	P-value
Jarque-Bera	146.73	< 2.2e-16
Shapiro-Wilk	0.93247	4.204e-09
Kołmogorow-Smirnow	0.36303	< 2.2e-16

Tablica 5.3: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

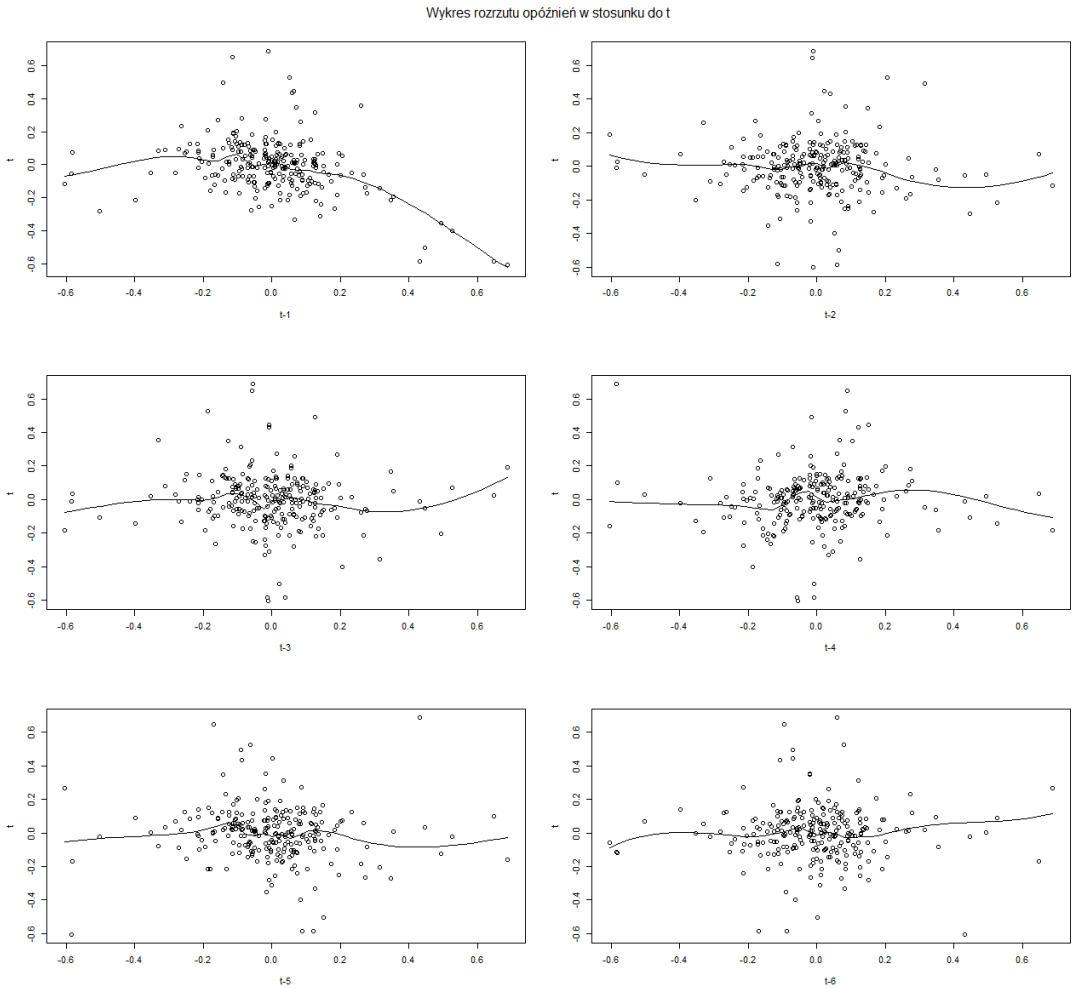
Wszystkie trzy testy jednoznacznie odrzuciły normalność rozkładu, natomiast histogram wskazuje leptokurtyczny rozkład normalny. Rozkład szeregu nie zostaje uznana za rozkład normalny. W dalszej części analizy zbadana zostanie autokorelacja szeregu:



Rysunek 5.2: Wyniki testów ACF, PACF (górny rząd) oraz p-values Ljunga-Boxa i histogram szeregu.

Test ACF wykazał istotną autokorelację dla 1 opóźnienia, oraz test PACF istotną autokorelację cząstkową dla do 3 opóźnienia włącznie. Potwierdzone to zostało za pomocą testu Ljunga-Boxa.

Zbadano także kwadraty zwrotów na występowanie autokorelacji - p-values dla wszystkich rzędów jest mniejszy od $\alpha = 0,05$, zatem w szeregu występuje efekt ARCH.



Rysunek 5.3: Wykresy rozrzutu szeregu czasowego w momencie od t do $t - i$ z nałożoną linią regresji lokalnej.

Wykres rozrzutu wskazuje na nieliniowość w pierwszym opóźnieniu.

5.2 Szacowanie modeli ekonometrycznych

Szacowanie modelu ARMA

Estymowany został model ARMA(1, 1):

	Parametr	Błąd standardowy	Statystyka z	P-value
ar1	0.4346663	0.0775981	5.6015	2.125e-08 ***
ma1	-0.9205735	0.0404338	-22.7674	< 2.2e-16 ***
const	-0.0025838	0.0013415	-1.9261	0.0541 .

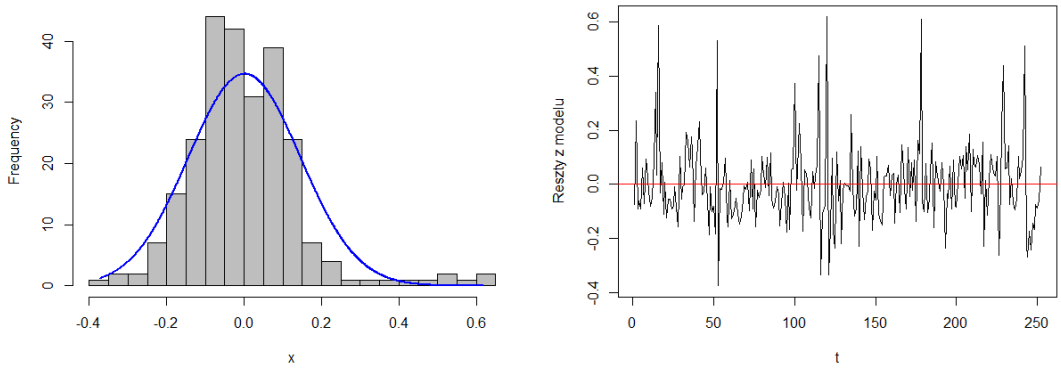
Tablica 5.4: Oszacowane parametry modelu ARMA(1, 1), jego błąd standardowy, statystyka Z oraz p-value

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-5.6907	0.094017
P-value	0.01	0.1

Tablica 5.5: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana dystrybucja reszt tego modelu:

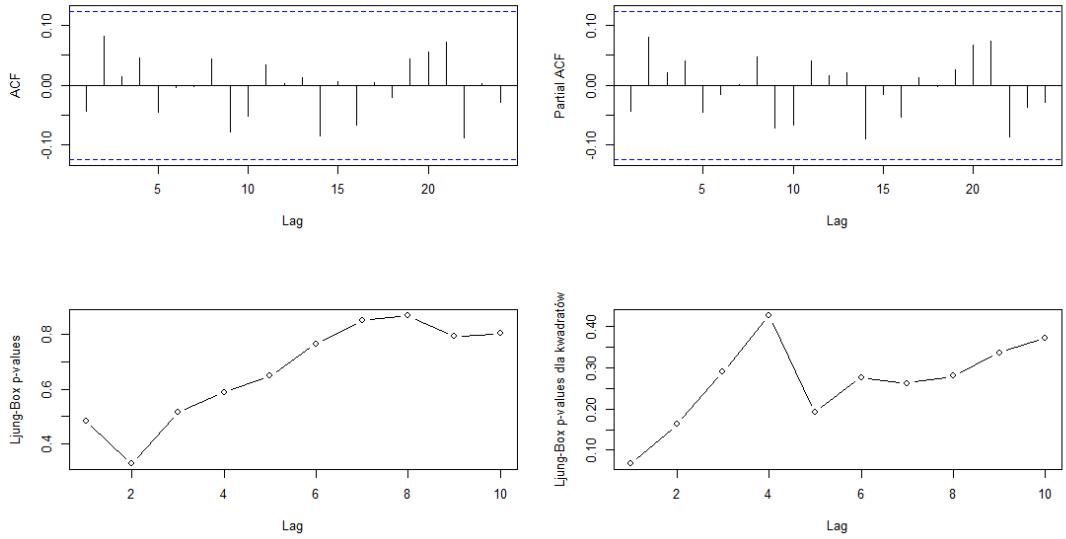


Rysunek 5.4: Histogram reszt modelu ARMA i ich wykres w zależności od czasu.

Histogram reszt wskazuje na rozkład normalny, jednak widoczne są dwa szczyty. P-values wszystkich trzech testów jest mniejsze od $\alpha = 0.05$, zatem odrzucona jest hipoteza zerowa o normalności rozkładu. Wykres reszt wskazuje na występowanie wartości odstających, ale rozkładają się one niezależnie od czasu. Zbadana zostanie następnie autokorelacja oraz cząstkowa autokorelacja szeregu:

	Statystyka	P-value
Jarque-Bera	244.74	< 2.2e-16
Shapiro-Wilk	0.90957	3.352e-11
Kołmogorow-Smirnow	0.38502	< 2.2e-16

Tablica 5.6: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa



Rysunek 5.5: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Żaden test nie wykazał istotnej statystycznie autokorelacji reszt modelu. Wszystkie zależności liniowe szeregu zostały wyjaśnione za pomocą modelu ARIMA(1, 0, 1). Ponadto reszty nie wykazują efektu ARCH.

Szacowanie modelu SETAR

Oszacowany został model SETAR(2,1,5) o parametrach:

	Parametr	Błąd standardowy	Statystyka z	P-value
$\phi_{0,const}$	0.013843	0.010932	1.2662	0.206649
$\phi_{0,1}$	-0.558252	0.077597	-7.1942	7.689e-12 ***
$\phi_{1,const}$	-0.017232	0.032207	-0.5351	0.593101
$\phi_{1,1}$	-0.337387	0.108588	-3.1070	0.002113 **
$\phi_{1,2}$	-0.341590	0.169159	-2.0193	0.044545 *
$\phi_{1,3}$	-0.637389	0.151884	-4.1966	3.797e-05 ***
$\phi_{1,4}$	-0.405347	0.152269	-2.6620	0.008283 **
$\phi_{1,5}$	-0.642352	0.145434	-4.4168	1.507e-05 ***
r	0.0595	-	-	-
d	1	-	-	-

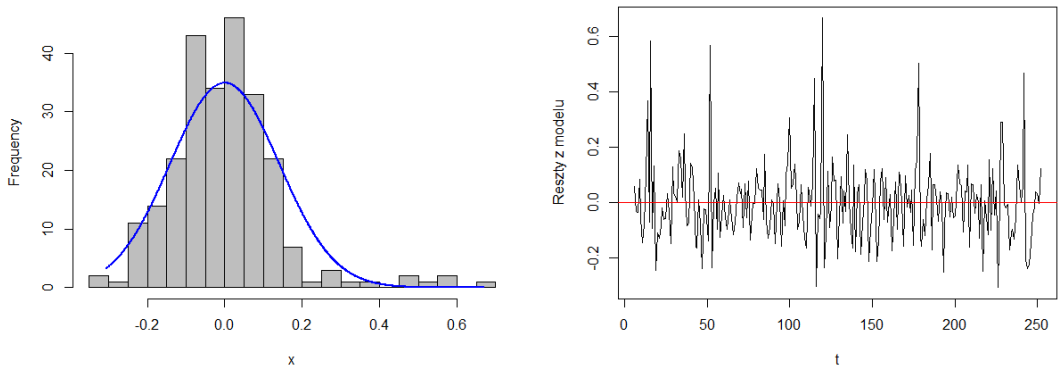
Tablica 5.7: Oszacowane parametry modelu SETAR, ich błędy standardowe, statystyki t i p-values

Zostaną sprawdzone reszty z tego modelu. Na początek zbadana zostanie ich stacjonarność:

	ADF	KPSS
Statystyka	-7.307	0.045719
P-value	0.01	0.1

Tablica 5.8: Statystyki i p-values testów ADF i KPSS

Oba testy wykazały stacjonarność reszt z modelu. W następnym kroku zbadana zostanie zbadana ich dystrybucja::



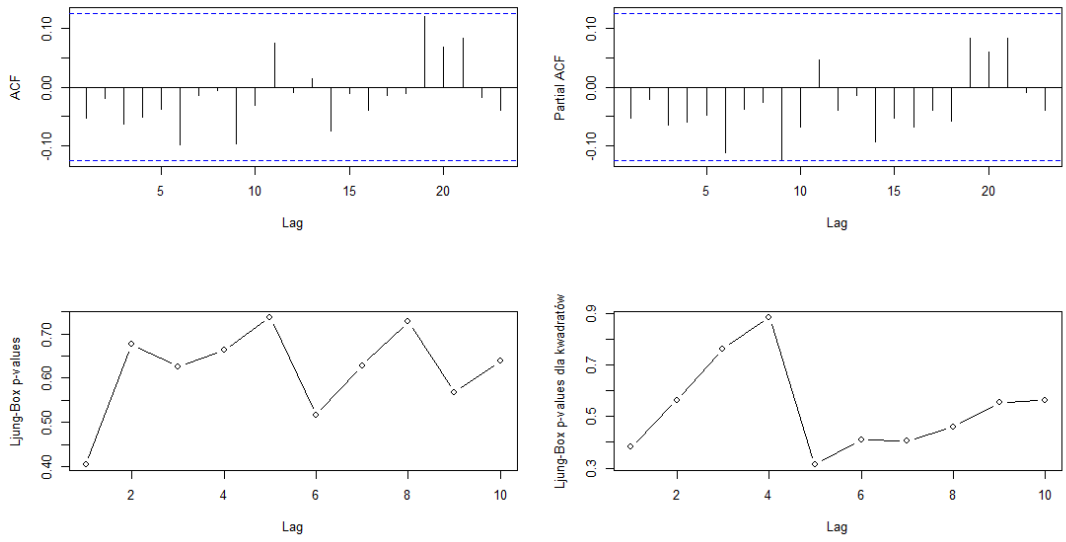
Rysunek 5.6: Histogram reszt modelu SETAR i ich wykres w zależności od czasu.

Histogram reszt wizualnie przypomina rozkład normalny, a ich wykres wskazuje na rozkład niezależny od czasu. Widoczne są także wartości odstające.

	Statystyka	P-value
Jarque-Bera	264.27	< 2.2e-16
Shapiro-Wilk	0.91166	6.574e-11
Kołmogorow-Smirnow	0.39249	< 2.2e-16

Tablica 5.9: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

Wszystkie p-values są mniejsze od $\alpha = 0.05$, zatem odrzucona zostaje hipoteza zerowa o normalności rozkładu. Rozkład reszt modelu SETAR(2,1,5) nie jest rozkładem normalnym. Zostanie następnie zbadana autokorelacja reszt oraz ich kwadratów:

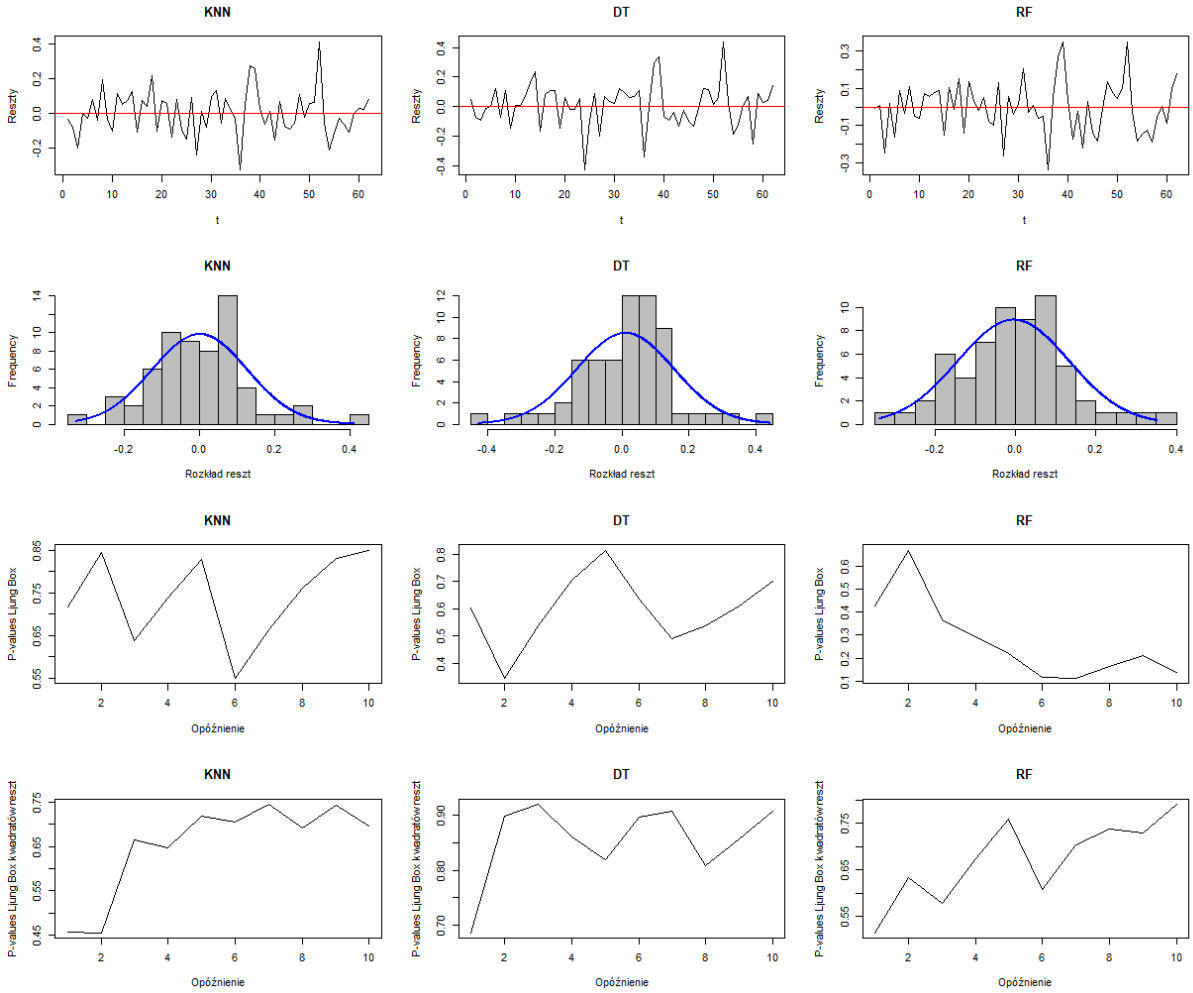


Rysunek 5.7: Wyniki testów ACF, PACF (górny rząd) oraz p-values testu Ljunga-Boxa (dolny rząd) dla reszt oraz kwadratów reszt modelu.

Testy ACF i PACF nie wykazały autokorelacji, potwierdzone to zostało za pomocą testu Ljunga-Boxa którego p-values jest powyżej $\alpha = 0.05$, więc nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji. Szereg nie wykazuje także efektu ARCH.

5.3 Szacowanie modeli uczenia maszynowego

Przeanalizowane zostały reszty ze wszystkich trzech modeli:



Rysunek 5.8: W kolejnych rzędach: wykresy reszt w zależności od czasu, histogramy, p-values testu Ljunga-Boxa reszt oraz kwadratów reszt

Reszty wszystkich trzech modeli wykazują wartości odstające, ale nie wskazują one na zależność od czasu. Wszystkie rozkłady reszt przypominają wizualnie rozkład normalny, zostanie to zweryfikowane za pomocą dalszych testów. Wszystkie p-values testu Ljunga-Boxa są większe niż $\alpha = 0.05$, zatem nie ma podstaw do odrzucenia hipotezy zerowej o braku autokorelacji reszt oraz kwadratów reszt. Reszty nie są autoskorelowane liniowo i nie posiadają efektu ARCH. Przeprowadzono szereg testów na normalność składnika resztowego:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	5.2837	0.07123	0.97464	0.2267	0.38876	5.998e-09
DT	6.6294	0.03634	0.96026	0.04271	0.37143	3.485e-08
RF	1.4754	0.4782	0.97807	0.3313	0.38281	1.109e-08

Tablica 5.10: Statystyki i p-values testów Jarque-Bera, Shapiro-Wilka i Kołmogorowa-Smirnowa

W przypadku modeli KNN oraz lasu losowego, p-values testów Jarque-Bera oraz Shapiro-Wilka są większe od $\alpha = 0.05$, zatem w przypadku tych testów brak jest podstaw do odrzucenia hipotezy zerowej o normalności rozkładu, zaś p-value testu Kołmogorowa-Smirnowa jest mniejsze od $\alpha = 0.05$, zatem hipotezę zerową o normalności rozkładu można odrzucić. Rozkład reszt z modeli KNN oraz lasu losowego nie zostaje uznany za rozkład normalny. W przypadku modelu drzewa decyzyjnego wszystkie trzy p-values są mniejsze od $\alpha = 0.05$, zatem hipoteza zerowa o normalności rozkładu również zostaje odrzucona, rozkład ten nie zostaje uznany za rozkład normalny.

	Statystyka ADF	P-value ADF	Statystyka KPSS	P-value KPSS
DT	-3.8986	0.01999	0.042716	0.1
RF	-4.4458	0.01	0.035039	0.1
KNN	-4.505	0.01	0.038436	0.1

Tablica 5.11: Statystyki i p-values testów ADF i KPSS

P-values wszystkich modeli dla testu ADF są mniejsze od $\alpha = 0.05$, oraz dla testu KPSS są większe od $\alpha = 0.05$, zatem błędy zostają uznane za stacjonarne. Aby porównać dopasowanie modeli policzone zostały funkcje błędu:

	KNN	DT	RF
ME	0.00095	0.01073	**-0.00026**
MSE	**0.12504**	0.14453	0.14317
RMSE	**0.01563**	0.02089	0.0205
MAE	**0.09628**	0.10696	0.11089
MAPE	3.15921	2.90387	**2.07169**

Tablica 5.12: Wartości funkcji błędów dla reszt z predykcji modeli uczenia maszynowego

Najmniejszy błąd RMSE posiada model KNN, zatem zostaje on uznany za najlepiej dopasowany spośród modeli uczenia maszynowego.

6 Wyniki badań

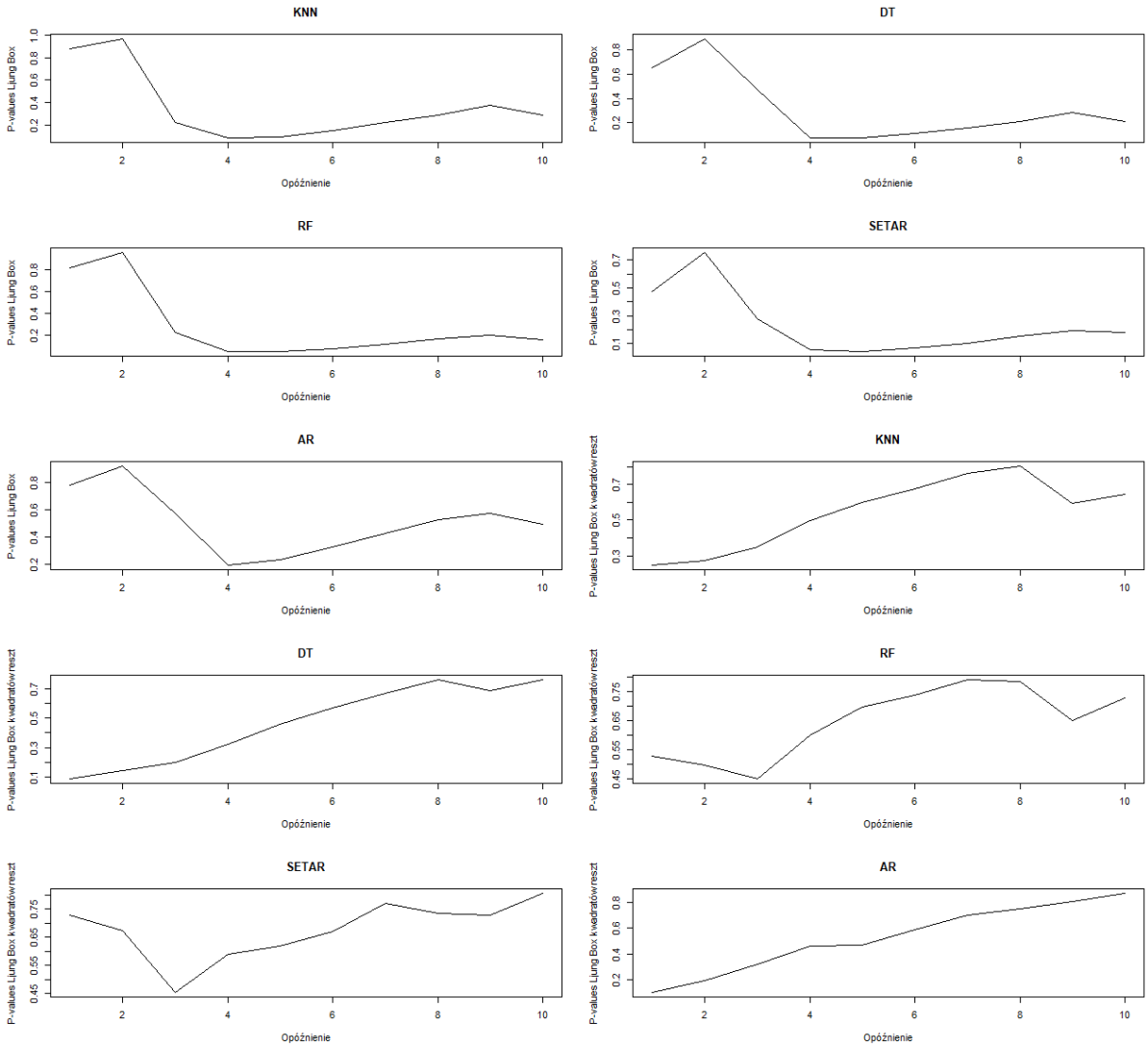
6.1 Wyniki dla danych symulowanych

Dla zestawu treningowego została obliczona prognoza. Do porównania wszystkich użytych modeli policzone zostały funkcje błędu:

	KNN	DT	RF	AR	SETAR
ME	0.09412	0.04245	0.05546	**0.02976**	0.04466
MSE	0.89608	0.91343	0.87611	1.00255	**0.84069**
RMSE	0.80296	0.83436	0.76757	1.0051	**0.70676**
MAE	0.70055	0.71933	0.6909	0.81103	**0.6479**
MAPE	1.43125	1.56942	1.53252	**1.22234**	1.40835

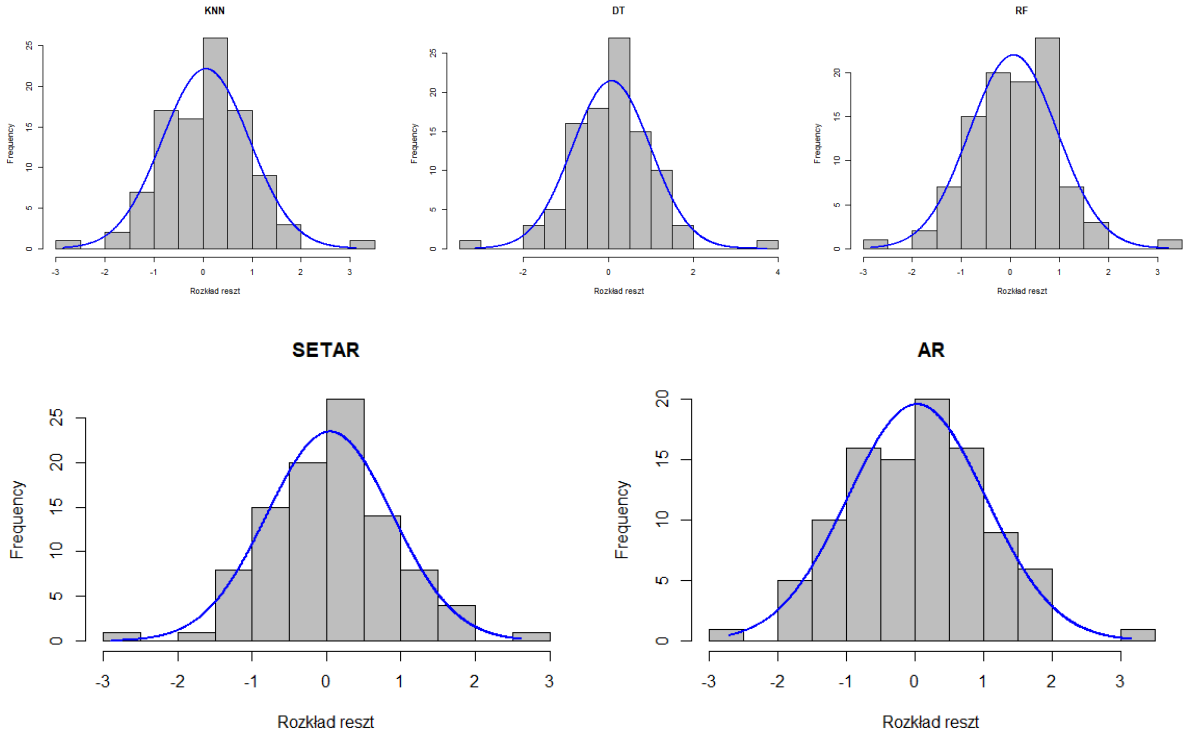
Tablica 6.1: Wartości funkcji błędów dla reszt z prognoz wszystkich modeli zestawu testowego

Najlepszym algorytmem pod względem RMSE okazał się być SETAR. Najlepszym algorytmem uczenia maszynowego jest w tym przypadku las losowy, natomiast najgorszy wynik otrzymał model liniowy AR. Należy jednak zauważyć, że dane brane pod uwagę to dane wygenerowane właśnie ze wzoru na model SETAR. Zbadane zostały autokorelacje reszt oraz kwadratów reszt tych prognoz:



Rysunek 6.1: P-values testu Ljunga-Boxa reszt oraz kwadratów reszt wszystkich modeli dla zestawu testowego

W przypadku każdego modelu p-values testu Ljunga-Boxa okazało się większe od $\alpha = 0.05$ dla przynajmniej pierwszego rzędu, nie ma zatem podstaw do odrzucenia hipotezy zerowej o braku autokorelacji reszt i kwadratów reszt z prognoz pochodzących ze wszystkich analizowanych modeli.



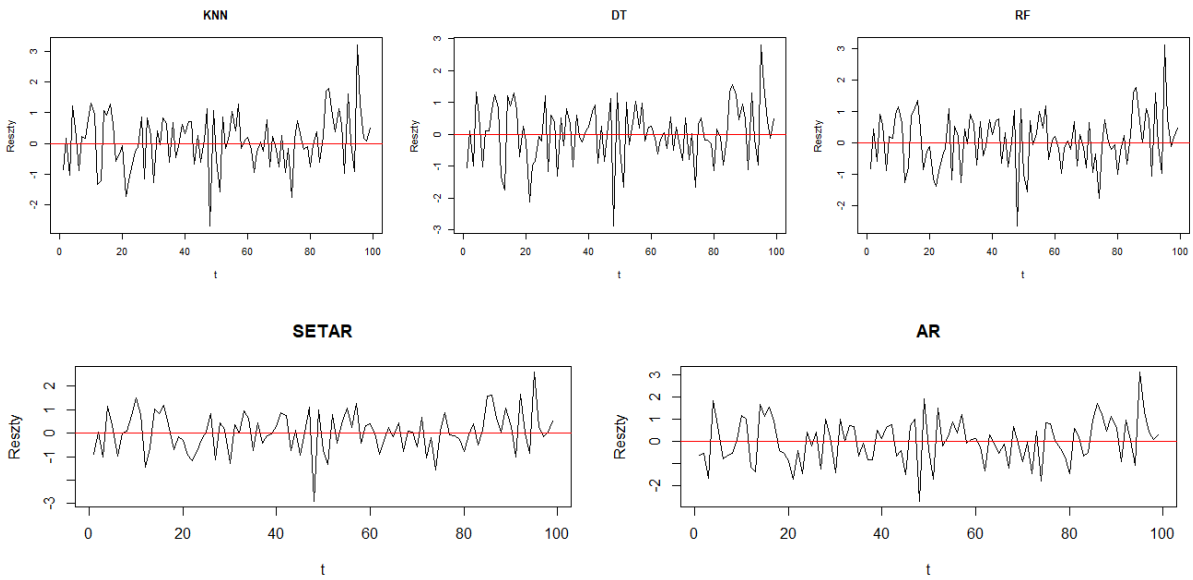
Rysunek 6.2: Histogramy reszt ze wszystkich modeli dla zestawu testowego

Histogramy reszt wszystkich modeli wskazują na normalność ich rozkładu. Zostanie to potwierdzone testami na normalność:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	4.9295	0.08503	0.98216	0.2006	0.11036	0.1663
DT	22.83	1.103e-05	0.96721	0.01429	0.10488	0.2109
RF	5.9154	0.05194	0.98266	0.2184	0.082758	0.4812
SETAR	3.5959	0.1656	0.98585	0.3722	0.091301	0.3594
AR	0.04273	0.9789	0.9921	0.8332	0.051912	0.9397

Tablica 6.2: Statystyki oraz p-values testów na normalność reszt ze wszystkich modeli dla zestawu testowego

Wszystkie 3 testy wykazały p-value mniejsze od $\alpha = 0.05$ w przypadku reszt modelu drzewa decyzyjnego, zatem odrzucona zostaje hipoteza zerowa o normalności ich rozkładu. P-values wszystkich testów pozostałych modeli są większe od $\alpha = 0.05$, zatem brak jest podstaw do odrzucenia hipotezy zerowej, rozkład reszt tych modeli zostaje uznany za normalny. Sprawdzone zostaną wykresy reszt wszystkich modeli:



Rysunek 6.3: Wykresy reszt względem czasu ze wszystkich modeli dla zestawu testowego

W przypadku reszt wszystkich modeli widoczne są wartości odstające, jednak wizualnie ich rozkład nie wskazuje na zależność od czasu.

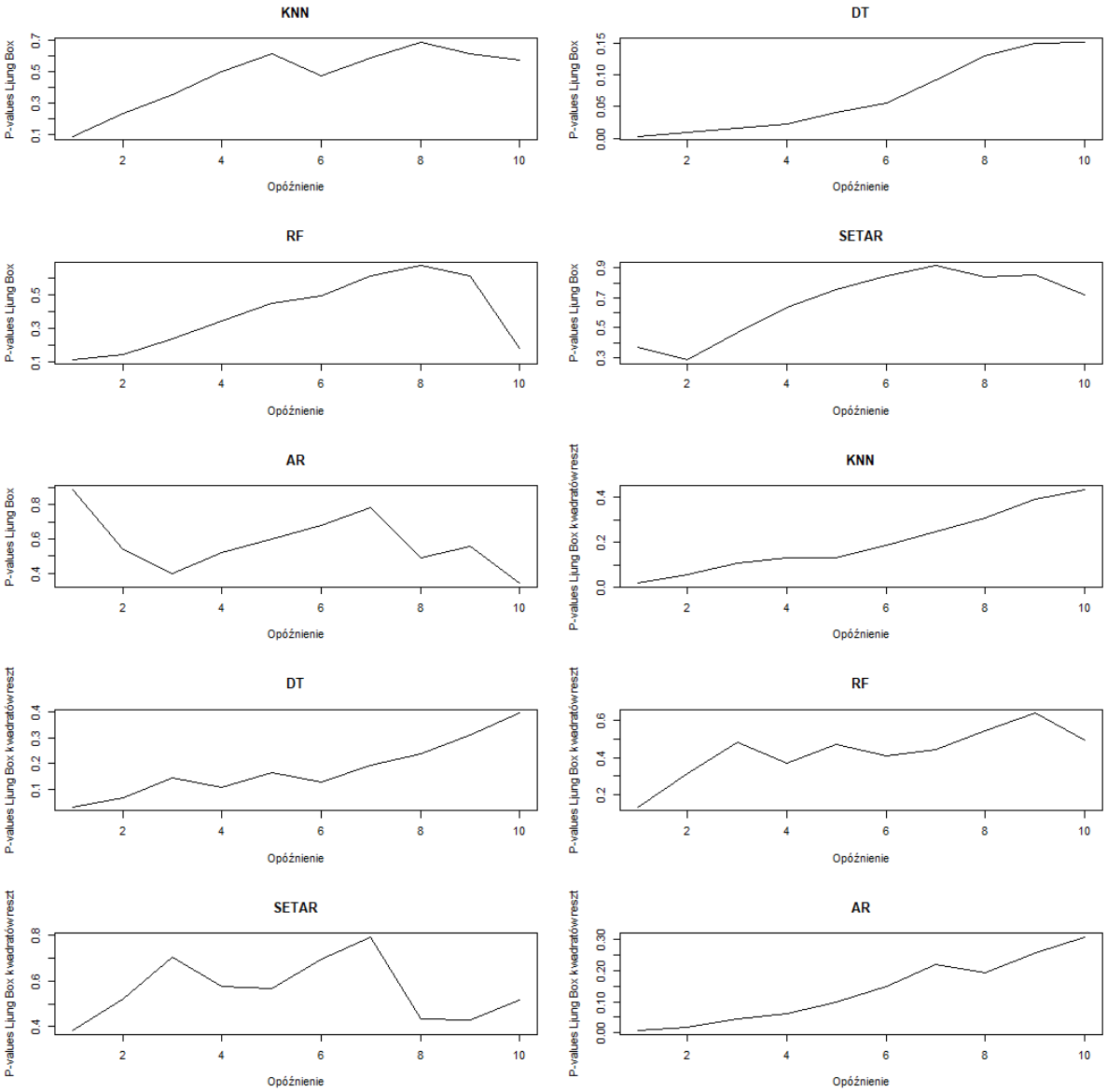
6.2 Wyniki dla danych rzeczywistych

Dla zestawu treningowego została obliczona prognoza. Do porównania wszystkich użytych modeli policzone zostały funkcje błędu:

	KNN	DT	RF	AR	SETAR
ME	**0.04621**	0.04722	0.0489	0.08486	0.0811
MSE	0.2788	0.36923	0.34778	0.23609	**0.23207**
RMSE	0.07773	0.13633	0.12095	0.05574	**0.05386**
MAE	0.21888	0.2887	0.25231	0.2005	**0.17923**
MAPE	0.07373	0.09744	0.0882	0.07035	**0.06008**

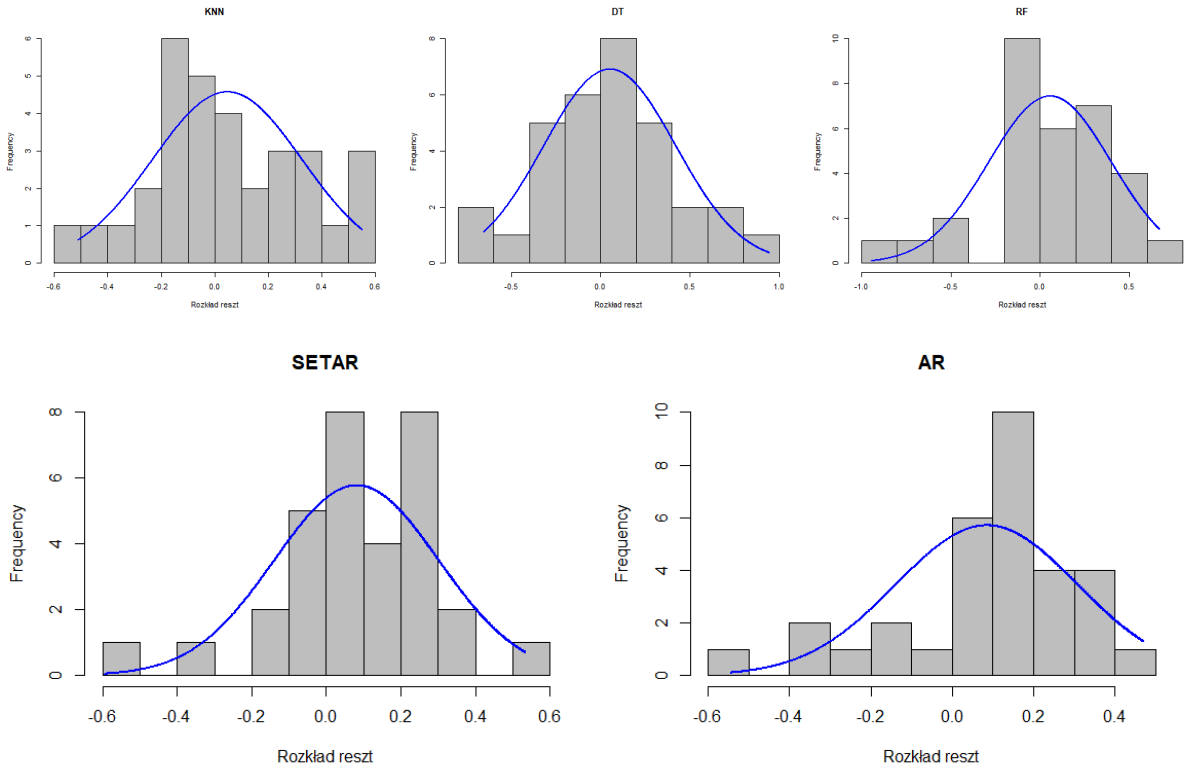
Tablica 6.3: Wartości funkcji błędów dla reszt z prognoz wszystkich modeli zestawu testowego

Najlepszym algorytmem pod względem RMSE okazał się być SETAR. Najlepszym algorytmem uczenia maszynowego jest w tym przypadku KNN, natomiast najgorszy wynik otrzymał model drzewa decyzyjnego. Zbadane zostały autokorelacje reszt oraz kwadratów reszt tych prognoz:



Rysunek 6.4: P-values testu Ljunga-Boxa reszt oraz kwadratów reszt wszystkich modeli dla zestawu testowego

W przypadku modelu drzewa decyzyjnego p-value testu Ljunga-Boxa okazało się mniejsze od $\alpha = 0.05$, zatem odrzucona zostaje hipoteza zerowa o braku autokorelacji składnika resztowego. W przypadku reszty modeli nie ma podstaw do odrzucenia hipotezy zerowej. P-values testu Ljunga-Boxa dla kwadratów reszt modelu AR jest mniejsze od $\alpha = 0.05$, zatem występuje w nich efekt ARCH.



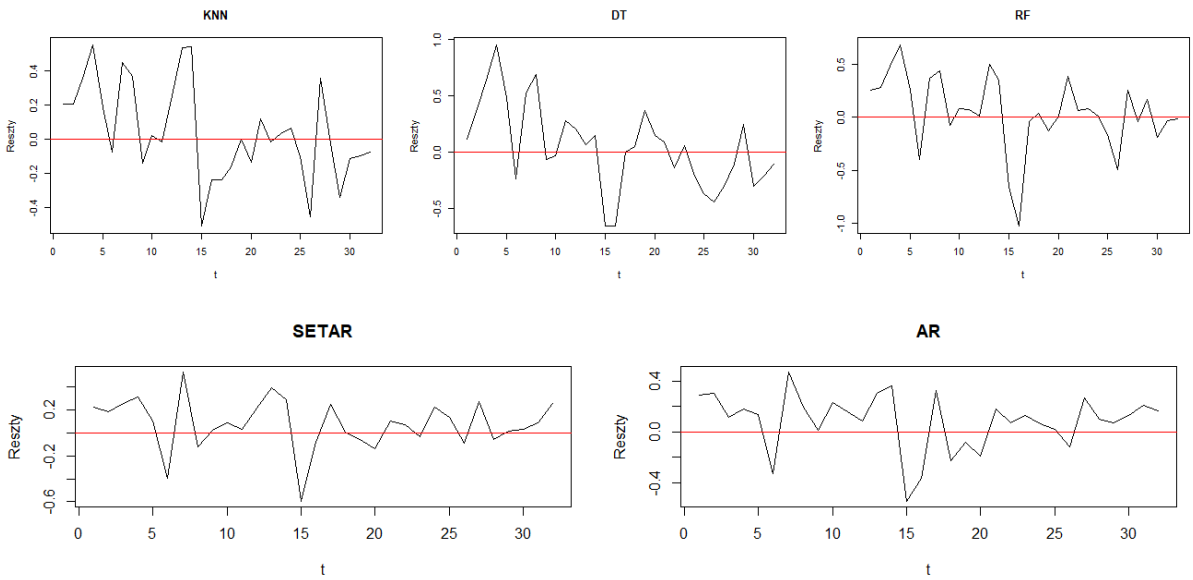
Rysunek 6.5: Histogramy reszt ze wszystkich modeli dla zestawu testowego

Histogram reszt modelu drzewa decyzyjnego przypomina rozkład normalny, pozostałe modele nie wykazują podobieństwa do rozkładu normalnego. Zostanie przeprowadzony szereg testów na normalność rozkładu:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	0.59183	0.7439	0.96643	0.4071	0.31098	0.003015
DT	0.31571	0.854	0.98673	0.9547	0.266	0.0173
RF	5.1888	0.07469	0.93417	0.05124	0.30724	0.003526
SETAR	7.3685	0.02512	0.94359	0.09472	0.38166	0.000106
AR	5.9569	0.05087	0.92539	0.02919	0.32883	0.00139

Tablica 6.4: Statystyki oraz p-values testów na normalność reszt ze wszystkich modeli dla zestawu testowego

Na poziomie istotności $\alpha = 0.05$ test Jarque-Bera odrzucił hipotezę zerową o normalności rozkładu dla modelu SETAR, test Shapiro-Wilka dla modelu AR, natomiast test Kołmogorowa-Smirnowa odrzucił hipotezę zerową dla wszystkich rozkładów. Żaden rozkład spośród wyżej wymienionych nie jest rozkładem normalnym.



Rysunek 6.6: Wykresy reszt względem czasu ze wszystkich modeli dla zestawu testowego

Reszty modeli KNN i drzewa decyzyjnego wskazują na występowanie trendu, a reszty modeli lasu losowego, SETAR i AR rozkładają się niezależnie od czasu.

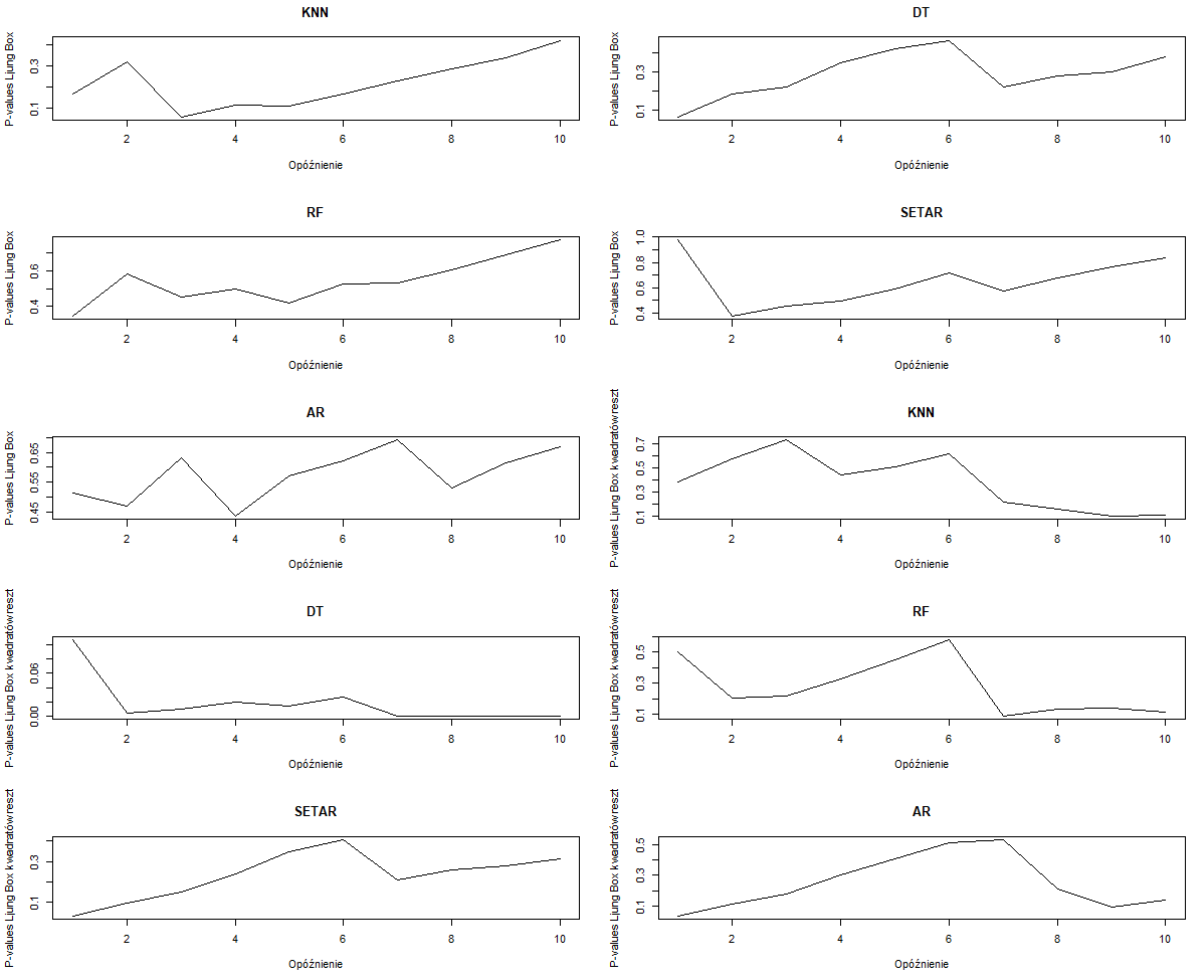
6.3 Wyniki dla danych giełdowych

Dla zestawu treningowego została obliczona prognoza. Do porównania wszystkich użytych modeli policzone zostały funkcje błędu:

	KNN	DT	RF	ARMA	SETAR
ME	**-0.00242**	0.0051	0.00266	0.03123	0.00926
MSE	0.13528	0.13545	**0.11968**	0.11987	0.12508
RMSE	0.0183	0.01835	**0.01432**	0.01437	0.01565
MAE	0.10168	0.10152	0.09186	**0.08654**	0.09264
MAPE	13.42426	**8.33185**	12.3185	10.5075	12.64461

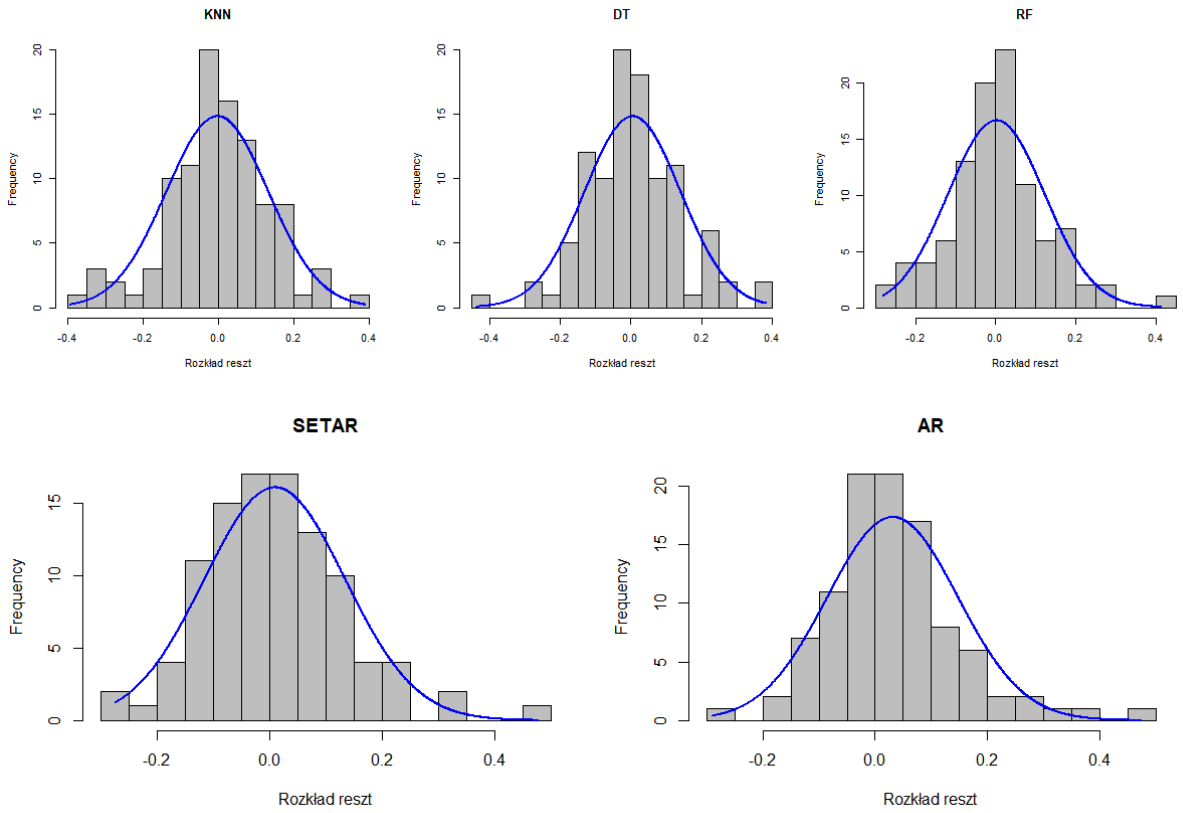
Tablica 6.5: Wartości funkcji błędów dla reszt z prognoz wszystkich modeli zestawu testowego

Najlepszym algorytmem pod względem RMSE okazał się być las losowy. Najlepszym modelem ekonometrycznym jest model ARMA, natomiast najgorszym modelem ze wszystkich jest model drzewa decyzyjnego. Należy jednak zauważyć, że w tym przypadku nie został wykorzystany model AR, tylko model ARMA biorący pod uwagę reszty swojego modelu autoregresyjnego – żaden inny model nie miał dostępu do takiej informacji.



Rysunek 6.7: P-values testu Ljunga-Boxa reszt oraz kwadratów reszt wszystkich modeli dla zestawu testowego

W przypadku każdego modelu p-values testu Ljunga-Boxa okazało się większe od $\alpha = 0.05$, nie ma zatem podstaw do odrzucenia hipotezy zerowej o braku autokorelacji reszt. P-values testu Ljunga-Boxa kwadratów reszt z prognoz modeli SETAR, ARMA i drzewa decyzyjnego są mniejsze od $\alpha = 0.05$, występuje w nich zatem efekt ARCH.



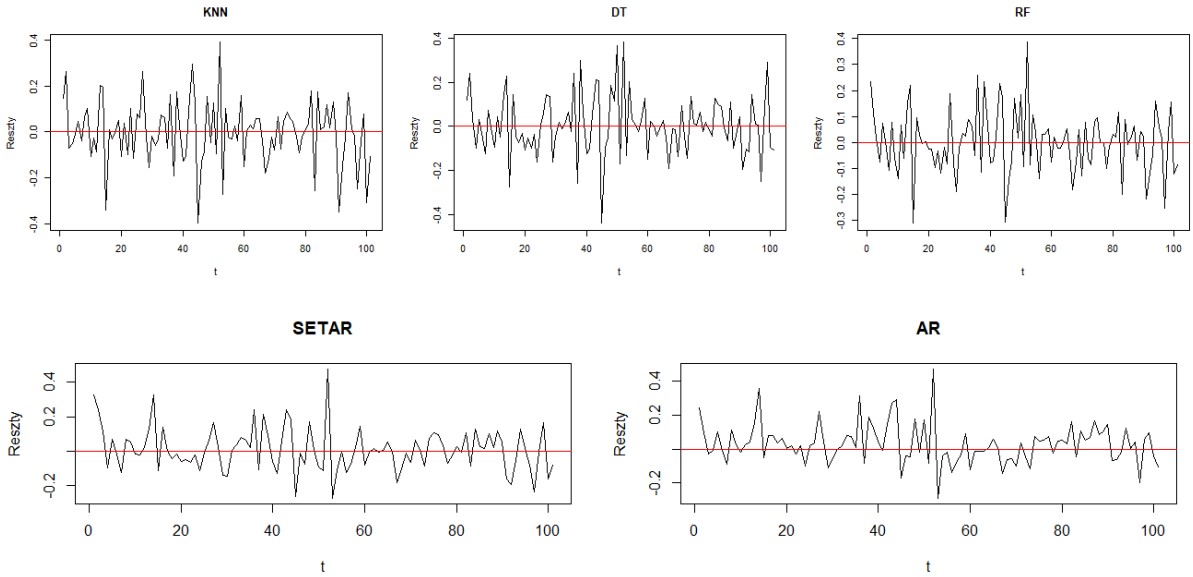
Rysunek 6.8: Histogramy reszt ze wszystkich modeli dla zestawu testowego

Histogramy reszt wszystkich modeli wskazują na normalność ich rozkładu. Zostanie to potwierdzone testami na normalność:

	Statystyka J-B	P-value J-B	Statystyka S-W	P-value S-W	Statystyka K-S	P-value K-S
KNN	4.0047	0.135	0.98088	0.151	0.38092	3.73e-13
DT	3.65	0.1612	0.9819	0.1815	0.38311	2.66e-13
RF	3.414	0.1814	0.97989	0.1262	0.38909	1.047e-13
SETAR	15.92	0.0003492	0.97022	0.02194	0.39306	5.584e-14
AR	26.607	1.669e-06	0.95864	0.003029	0.4136	1.998e-15

Tablica 6.6: Statystyki oraz p-values testów na normalność reszt ze wszystkich modeli dla zestawu testowego

Na poziomie istotności $\alpha = 0.05$ test Jarque-Bera odrzucił hipotezę zerową o normalności rozkładu dla modelu SETAR i AR, test Shapiro-Wilka również dla modelu AR i SETAR, natomiast test Kołmogorowa-Smirnowa odrzucił hipotezę zerową dla wszystkich rozkładów. Żaden rozkład spośród wyżej wymienionych nie jest rozkładem normalnym. Zbadane zostaną wykresy rozrzutu reszt:



Rysunek 6.9: Wykresy reszt względem czasu ze wszystkich modeli dla zestawu testowego

W przypadku reszt wszystkich modeli widoczne są wartości odstające, jednak wizualnie ich rozkład nie wskazuje na zależność od czasu.

6.4 Podsumowanie wyników

Celem badania było ustalenie, czy wybrane nieparametryczne modele uczenia maszynowego niewymagające założeń dotyczących rodzaju zależności są w stanie zapewnić prognozy porównywalne lub lepsze niż modele ekonometryczne. Aby odpowiedzieć na to pytanie, sprawdzone zostanie które modele zapewniają najmniejszą wartość funkcji błędu (RMSE):

	SETAR(2,1,1)	Lynx	SP500
Optymalny model	SETAR	SETAR	RF
Optymalny model ekonometryczny	SETAR	SETAR	ARMA
Optymalny model uczenia maszynowego	RF	KNN	RF

Tablica 6.7: Tabela przedstawia najlepiej dopasowane modele do danych z badań według RMSE.

Tak jak widać na powyższej tabeli, w dwóch na trzy badania najlepszym modelem okazał się model ekonometryczny SETAR. W przypadku danych wolumenu indeksu SP500 optymalnym modelem jest las losowy. Warto zauważyć, że ani razu nie pojawił się model drzewa decyzyjnego. Samo kryterium najmniejszej wartości RMSE może jednak być zwodnicze, dlatego dla pewności sprawdzone zostaną różnice procentowe RMSE we wszystkich modelach:

	KNN	DT	RF	AR/ARMA	SETAR
SETAR(2,1,1)	13,61%	18,05%	8,60%	42,21%	0,00%
Lynx	44,32%	153,12%	124,56%	3,49%	0,00%
SP500	27,79%	28,14%	0,00%	0,35%	9,29%

Tablica 6.8: Tabela przedstawia o ile procent prognozy w zbiorze testowym różnią się od modelu optymalnego pod względem RMSE. 0% oznacza model optymalny.

Ponieważ w przypadku danych symulowanych to model SETAR był ich generatorem, nie jest dziwnym, że okazał się być najlepiej dopasowany. Niewiele jednak różni się od niego model lasu losowego, ponieważ tylko o 8,6%. Zdecydowanie najgorzej wypadł model liniowy AR - nie powinno to jednak być niespodzianką, ponieważ nie zostało zachowane założenie o liniowości relacji (pomimo statystyk potwierdzających je). W przypadku danych populacji rysi pod względem prognozy modele SETAR i AR okazały się być bardzo podobne - różnica procentowa błędu wynosi jedynie 3,49%. Spośród modeli uczenia maszynowego najlepszy wynik osiągnęło KNN, zaś najgorsze były modele drzewa decyzyjnego i lasu losowego - o ponad 100%. Dla danych wolumenu indeksu SP500 najlepszym modelem okazał się być model lasu losowego, natomiast model ARMA wykazał praktycznie identyczną funkcję błędu - różnica wynosi zaledwie 0,35%. Model SETAR różnił się o 9,29%, zaś modele KNN i drzewa decyzyjnego w przybliżeniu o 28%. Czy zostały jednak wyjaśnione wszystkie zależności w tych szeregach?

	KNN	DT	RF	AR/ARMA	SETAR
SETAR	Tak	Tak	Tak	Tak	Tak
Lynx	Tak	Autokorelacja	Tak	Efekt ARCH	Tak
SP500	Tak	Tak	Tak	Efekt ARCH	Efekt ARCH

Tablica 6.9: Tabela przedstawia wyniki testów Ljunga-Boxa dla błędów i kwadratów błędów prognoz.

Z tabeli wynika, że model Drzewa Decyzyjnego nie był w stanie wyjaśnić nawet autokorelacji liniowych w danych populacji rysi. Model liniowy AR wyjaśnił co prawda autokorelacje liniowe, ale pozostał efekt ARCH. W danych giełdowych natomiast reszty modeli ekonometrycznych wykazały warunkową heteroskedastyczność, natomiast modele uczenia maszynowego nie.

Wnioski

W szeregach SP500 oraz populacji rysi widać było wyraźnie, że pomimo obecności nieliniowości w danych, prostsze modele czasami potrafią zapewnić bardzo podobnej jakości prognozy niż modele nieliniowe. Nieuprawnionym wnioskiem byłoby jednak stwierdzenie, że dzieje się tak zawsze - badanie z użyciem danych wygenerowanych wykazało ponad wszelką wątpliwość, że model liniowy AR nie był w stanie zapewnić prognozy o zbliżonej funkcji błędu do modeli nieliniowych, gdy nieliniowość ta jest silna. W przypadku danych populacji rysi pomimo nieliniowości modele uczenia maszynowego dostarczyły dużo gorszej jakości prognozy w stosunku do modeli ekonometrycznych - przyczyny takiego stanu rzeczy można szukać w stosunkowo małej długości danych, a zatem ograniczonej możliwości interpolacji danych - wszystkie te modele nie opierają się na równaniach regresji, tylko na średnich arytmetycznych liczonych według odpowiedniej reguły.

Czy zatem modele KNN, las losowy oraz drzewa decyzyjne są w stanie zapewnić podobnej jakości prognozy, jak modele ekonometryczne? Tak. Czy są one w stanie zapewnić lepszej jakości prognozy niż modele ekonometryczne? Tak, ale to zależy od cech danego szeregu. Należy zwrócić uwagę, że w niniejszym badaniu została przyjęta zasada, że modele uczenia maszynowego otrzymują maksymalną ilość opóźnień zastosowaną w modelach ekonometrycznych - oznacza to, że bez takiego ograniczenia możliwa jest jeszcze lepsza optymalizacja tych modeli, a zatem nadal istnieje pole do poprawy jakości prognozy.

Kontynuacja badań

Pomimo satysfakcjonujących wyników badań należy zauważyć, że wykorzystane w nich modele uczenia maszynowego są algorytmami prostymi i niewykorzystującymi nowych metod usprawniających ich siłę predykcyjną. Dlatego kontynuację badań należy moim zdaniem rozpocząć od wykorzystania innych, skuteczniejszych algorytmów od tych, które zostały sprawdzone w tym badaniu. Użycie modeli zespołowych można rozszerzyć na wykorzystanie metod wzmacniających - przykładem mogą być: XGBoost, CatBoost, AdaBoost, LightGBM.

W przypadku KNN najbardziej intuicyjnym rozszerzeniem może być wprowadzenie regularyzacji. Należy jednak zauważyć, że jest to algorytm wyciągający średnią spośród najbliższych obserwacji - nie jest zatem estymowana funkcja gładka. Aby rozszerzyć dalsze badania można wykorzystać alternatywne algorytmy, takie jak na przykład estymator Nadaraya-Watsona.

Najbardziej interesującą alternatywą dla modelu drzewa decyzyjnego jest moim zdaniem model MARS - jest to bardzo podobny algorytm wykorzystujący klasyczną metodę najmniejszych kwadratów zamiast średniej arytmetycznej dla estymacji predykcji. Jest to zatem algorytm przypominający w działaniu SETARa, natomiast nie jest oparty na dokładnym testowaniu cech statystycznych danych, a raczej na empirycznych metrykach błędu. Oznacza to, że nie wymaga on wiedzy na temat prognozowanego szeregu, a zatem jest mniej podatny na błędną identyfikację przez prognostę.

Oprócz zmiany modeli uczenia maszynowego zaimplementowane mogą zostać także metody inżynierii cech - w niniejszym badaniu jedynymi informacjami branymi pod uwagę były opóźnienia. W rzeczywistości jednak rzadko są to jedyne informacje dostępne dla prognosty. Wykorzystać można na przykład dane dotyczące czasu - takie jak dzień, miesiąc lub kwartał (w zależności od interwału), lub inne szeregi czasowe. W takiej sytuacji porównane powinny one być do modeli wektorowych typu TVAR lub TVECM.

Zakończenie

Celem niniejszego badania była odpowiedź na pytanie: Czy wybrane modele uczenia maszynowego są w stanie dostarczyć podobnej jakości prognozy w stosunku do modeli ekonometrycznych?

Na pytanie to została udzielona odpowiedź twierdząca. W pierwszym badaniu, w którym pod uwagę brany był szereg wygenerowany za pomocą procesu SETAR(2,1,1), modele uczenia maszynowego co prawda nie zdołały osiągnąć lepszej wartości funkcji błędu jak model SETAR, natomiast osiągnęły one lepsze wyniki od modelu AR. W drugim badaniu wszystkie modele uczenia maszynowego osiągnęły wyniki znacznie gorsze niż modele ekonometryczne zarówno liniowe jak i nieliniowe - sugeruje to, że problem leży gdzie indziej, np. w krótkim zestawie danych. W trzecim badaniu model lasu losowego osiągnął najlepszy wynik spośród wszystkich modeli, a w przeciwieństwie do modeli ARMA i SETAR jego błędy nie wykazywały efektu ARCH.

Modele uczenia maszynowego są w stanie zapewnić prognozy porównywalne a także lepsze od modeli uczenia maszynowego. Metody regresji nieparametrycznej mogą w znacznym stopniu uprościć stawianie dobrej jakości prognoz bez konieczności dogłębnej weryfikacji statystycznej szeregów czasowych branych pod uwagę.

Bibliografia

Dodge, Y. (2010). *The concise encyclopedia of statistics*. Springer.

Doman, M. i Doman, R. (2009). *Modelowanie zmienności i ryzyka: Metody ekonometrii finansowej*. Oficyna a Wolters Kluwer business.

Elliott, G., Granger, C. W. J. i Timmermann, A. (2006-). *Handbook of economic forecasting* (1st ed., T. 24). Elsevier North-Holland.

Hanke, J. E. i Wichern, D. W. (2014a). *Business forecasting* (Pearson new international edition). Pearson Educated Limited.

Hanke, J. E. i Wichern, D. W. (2014b). *Business forecasting* (Pearson new international edition). Pearson Educated Limited.

Henrik Madsen and Jan Holst. (2009). Modeling Non-Linear and Non-Stationary Time Series.

Hyndman, A., R.J. (2018). Forecasting: Principles and Practice. https://otexts.com/fpp2/backshift.html

IBM. (2022). What is overfitting? **retrieved**2022**from**https://www.ibm.com/cloud/learn/overfitting

Javier Fernandez. (2022). Choose the appropriate normality test. https://towardsdatascience.com/choose-the-appropriate-normality-test-d53146ca1f1c

Lewinson, E. (2022). A Step-by-Step Guide to Calculating Autocorrelation and Partial Autocorrelation. **retrieved**2022**from**https://towardsdatascience.com/a-step-by-step-guide-to-calculating-autocorrelation-and-partial-autocorrelation-8c4342b784e8

Maddala, G. S. (1994). *Introduction to econometrics* (2. ed., 5. [Dr.]). Macmillan.

Mahmoud, H. F. F. (2019). Parametric versus Semi and Nonparametric Regression Models. https://doi.org/10.48550/ARXIV.1906.10221

NASA. (2022). Weather Forecasting Through the Ages. https://aqua.nasa.gov/sites/default/files/references/Wx_Forecasting.pdf

NIST. (2022). *Minkowski distance*. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/minkdist.htm

Packt Editorial Staff. (2022). Cross-Validation strategies for Time Series forecasting. https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/

PennState. (2022). Online Master of Applied Statistics program. https://online.stat.psu.edu/stat500/lesson/9/9.2/9.2.3

Prabhakaran, S. (2022). Augmented Dickey Fuller Test (ADF Test) – Must Read Guide. **retrieved**2022**from**http
//www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/
PWN. (2022). prognozowanie. **retrieved**2022**from**https://sjp.pwn.pl/slowniki/prognozowanie.
html
Raphals, L. (2013). *Divination and Prediction in Early China and Ancient Greece*. Cambridge University Press.
Rhys, H. (2020). *Machine Learning with R, the tidyverse, and mlr*. Manning.
SGH. (2022). MODELOWANIE POLSKIEJ GOSPODARKI z R. https://web.sgh.waw.pl/~mrubas/
EFzR/pdf/R1Prezentacja.pdf
Shojaei, A. i Flood, I. (2018). Univariate Modeling of the Timings and Costs of Unknown Future Project Streams: A Case Study. *International Journal on Advances in Systems and Measurements*, *11*, 38.
Sklearn. (2022). K-Fold Cross-Validation. https://scikit-learn.org/stable/_images/grid_search_
cross_validation.png
Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
Tsay, e. S. (2005). *Analysis of financial time series*. Wiley.
Tsay, R. S. i Chen, R. (2019). *Nonlinear time series analysis*. Wiley.
Wikipedia. (2022a). *Cross-sectional data*. https://en.wikipedia.org/wiki/Cross-sectional_data
Wikipedia. (2022b). *Time series*. **retrieved**2022**from**https://en.wikipedia.org/wiki/Time_series

Spis rysunków

Spis tablic