Analiza metod detekcji anomalii na podstawie przebiegu kursów instrumentów finansowych

Aleksandra Dzieniszewska Eryk Warchulski

Abstract

Celem niniejszego dokumentu jest przedstawienie wstępnych założeń dotyczących realizacji projektu analizy algorytmów detekcji anomalii w szeregach czasowych, tj. TSAD (z ang. **time series anomaly detection**). W jego ramach zostanie omówiona domena problemu, tj. przebiegi wartości instrumentów finansowych na różnych typach rynków, model zjawiska detekcji anomalii oraz plan eksperymentów. Opisane zostaną również metryki jakości uzyskiwanych rozwiązań jak i źródło pozyskiwanych danych.

1 Wprowadzenie

Przebieg kursów aktywów finansowych jak na przykład kursów akcji spółek lub indeksów giełdowych jest fundamentem działania inwestorów na rynkach. Jego zachowanie determinuje strategie oraz ryzyko inwestycyjne. W związku z powyższym kursy akcji są szczególnie interesujące nie tylko dla bezpośrednich aktorów rynkowych, ale również ekonomistów lub statystyków, którzy opracowują modele zachowań tych kursów lub rynków w ogólności. Współcześnie bardzo dużo uwagi poświęca się zagadnieniom predykcji przyszłych wartości kursów, co stanowi złożony i trudny problem, biorąc pod uwagę fakt, że jedna z dominujących teorii rynkowych, tj. EFM (z ang. **efficient-market hypothesis**) uważa za niemożliwe dokonywanie przydatnych predykcji [1] przyszłych kursów, a ponadto uznaje się, że rynki finansowe nie są obojętne na predykcje jak np. pogoda [2]. Niemniej jednak problem predykcji przyszłych wartości kursów nie jest jedynym zagadnieniem, które ma praktyczne znaczenie. Możliwość odróżnienia niestandardowych fluktuacji kursów i tym samym punktów odstających może stanowić istotną informację dla inwestorów i zaangażować dedykowane takim zdarzeniom procesy biznesowe. Szczególnie istotne wydaje się to z punktu widzenia zautomatyzowanych rynków giełdowych, w których akcje podejmowane przez aktorów (systemy decyzyjne) mierzone są w milisekundach, a dane giełdowe mogą być traktowane jako strumieniowe [3]. Wówczas wykrycie anomalii i jej odpowiednie obsłużenie wydaje się być krytyczne dla gracza rynkowego.
Dalszy rozkład dokumentu jest następujący: w rodziale drugim (2) krótko scharakteryzowane są wielkości, których przebieg zostanie poddany baddaniu w ramach projektu. Rodział (3) definiuje podstawowe pojęcia związane z zadaniem detekcji anomalii. W rozdziale (4) zarysowane są idee, które stoją za stosowanymi przez autorów algorytmami. Rozdział (5) i (6) zawierają kolejno – opis danych, które bedą stosowane, oraz plan eksperymentów.

2 Rynki finansowe

W ramach realizacji projektu użyte zostaną łącznie dwa zbiory danych, na które składają się przebiegi kursów instrumentów finansowych. Jeden z nich będzie pochodzić z rzeczywistego

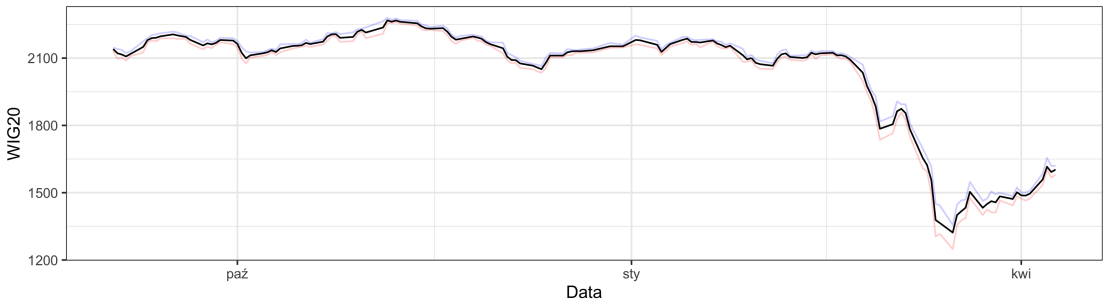

Figure 1: Przebieg indeksu giełdowego WIG20. Rysunek własny.

rynku finansowego, a drugi z rynku wirtualnego. Motywacją takiego podejścia jest chęć zaobserwowania zachowania się rynku w świecie, w którym na wartość kursu wpływa bardzo duża liczba czynników jawnych lub niejawnych, oraz w świecie, w którym zbiór możliwych akcji podejmowanych przez aktorów jest znacząco ograniczony oraz sam świat ma raczej charakter statyczny i iteracyjny. Ponadto dodatkową motywacją za skorzystaniem z danych pochodzących z rynku wirtualnego jest łatwość określenia zdarzeń, które powodują nagłe zmiany przebiegu kursu – co zostanie bardziej szczegółowo opisane w podsekcji poświęconej temu rynkowi.

2.1 Rynki rzeczywiste

Indeks giełdowy WIG20 jest statystyką (średnia ważona kapitalizacją spółek) obrazującą zmianę cen akcji dwudziestu największych spółek akcyjnych notowanych na Warszawskiej Giełdzie Papierów Wartościowych. Wartość indeksu pozwala ocenić inwestorom ogólny kierunek zmian cen i stan rynku.

Przebieg indeksu WIG20 w wybranym przedziale czasowym pokazany jest na rysunku poniżej.

2.2 Rynki wirtualne

WoW Token jest przedmiotem w grze MMO-RPG (*Massively multiplayer online role-playing game*) *World of Warcraft*, który przez gracza może zostać wykorzystany w następujące sposoby:

- gracz może kupić token za rzeczywistą walutę (USD, GBP, EUR, TWD, KRW), a następnie sprzedać go w umieszczonym w grze domu aukcyjnym (*Auction House*)
- gracz może kupić token, płacąc fikcyjną walutą obowiązującą w świecie *World of Warcraft* (dalej G), a następnie wymienić go na przedłużenie abonamentu gry lub wymienić go na bon w internetowym sklepie wydawcy gry.

Kurs wymiany G/USD jest stały i wynosi 20 [1] natomiast kurs tokena w świecie gry jest zmienny i zależy głównie od podaży i popytu [4].

Przebieg ceny tokena w grze na serwerach europejskich w wybranym czasie jest pokazany na poniższym rysunku.

[1] To samo tyczy się jakiejkolwiek innej waluty rzeczywistej.

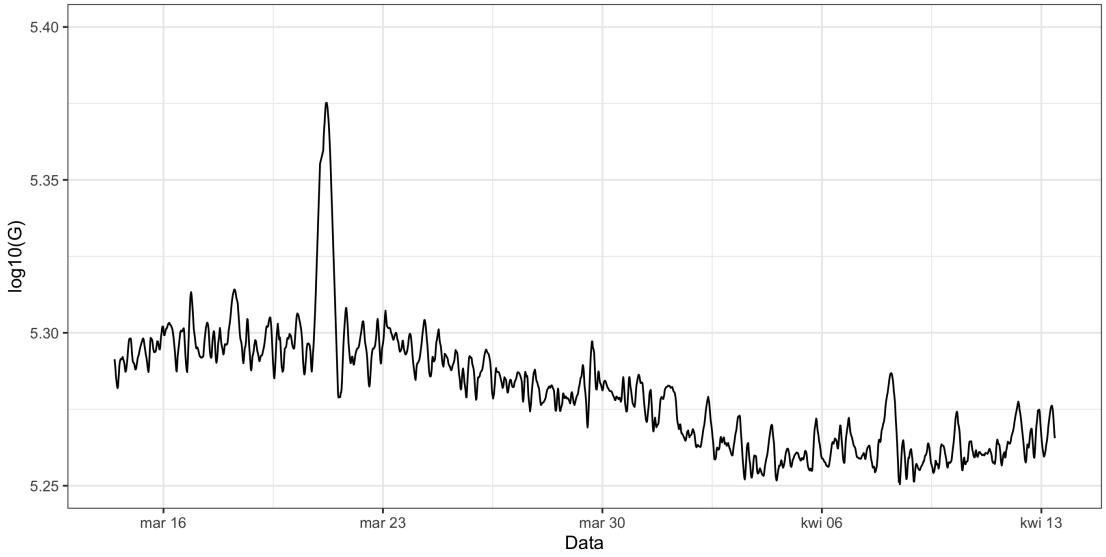

Figure 2: Przebieg wartości WoW tokna. Rysunek własny.

3 Zadanie detekcji anomalii

W rozdziale tym zostanie zdefiniowany w ogólny sposób problem detekcji anomalii w szeregach czasowych. Konkretyzacja ogólnych pojęć zdefiniowanych poniżej zostanie przedstawiona w rozdziale dotyczącym algorytmów detekcji.

Przez zadanie detekcji anomalii w trakcie realizacji projektu będziemy rozumieć następujący problem:

Szereg czasowy niech $(X_i)_{i\in T}$ będzie procesem stochastycznym określonym na pewnej przestrzeni probabilistycznej, a zbiór indeksów T będzie interpretowany jako zbiór chwil czasowych jednakowo odległych od siebie. Realizację tego procesu, tj. uporządkowany zbiór $\{x_t\}_{t=1,\dots,N}$, będziemy nazywać szeregiem czasowym. Nie zakładamy ponadto niczego względem stacjonarności tego szeregu lub rozkładu prawdopodobieństwa, z którego jest on generowany poza faktem, że jego nośnikiem jest zbiór liczb rzeczywistych $\mathbb{R}$.

Anomalia Anomalią w szeregu czasowym będziemy nazywali punkt w tym szeregu x_k, który według przyjętego kryterium *odstaje* od pozostałych punktów w najbliższym sąsiedztwie $x_{k-s},\dots,x_{k+s}$ (anomalia lokalna) lub względem wszystkich punktów w szeregu (anomalia globalna). Kryterium *odstawania* jest silnie zależne od kontekstu i procedury detekcji anomalii. Często przyjmowane kryterium wygląda następująco [5]:

$$|x_t - s(\{X_t\})| > \tau \tag{1}$$

przy czym $s(\star)$ jest pewną funkcją.

Detekcja anomalii Detekcją anomalii będziemy nazywali procedurę, pozwalającą wykryć w szeregu czasowym zbiór indeksów T_A punktów uznawanych za anomalię:

$$AD\colon \{X_t\} \to T_A \subset T.$$

4 Algorytmy detekcji anomalii

W rozdziale tym zostaną opisane algorytmy służące do detekcji anomalii, które w ramach projektu zamierzamy zbadać. Opis ten nie będzie zawierał dokładnego pseudokodu, a jedynie pewien formalizm matematyczny, który pozwala zobrazować idee stojące za omawianymi metodami. Ponadto wskazane i omówione zostaną parametry tych metod.

Omawiane metody

4.1 MAD

Metoda *MAD*, tj. *median absolute deviation* jest stosunkowo prostym sposobem detekcji anomalii opartym na oknie kroczącym (z ang. *moving window*) o długości k. Dla każdego punktu metoda estymuje dwie wielkości:

1. medianę w oknie
2. bezwzględne odchylenie mediany.

Model można sformalizować w następujący sposób: dla każdego punktu szeregu czasowego x_i liczona jest wielkość

$$MAD_i = median_i[x_i - median_j(x_j)].$$

przy czym $median_j(\star)$ jest medianą j-tego okna.

Następnie każdy punkt szeregu czasowego porównywany jest z odpowiadającą mu wartość $MAD_\star$ – jeśli wartość bezwzględna różnicy między tym punktem, a wartością MAD jest większa od ustalonego progu, to punkt jest klasyfikowany jako anomalia. W nawiązaniu do (1) – kryterium konkretyzuje się do postaci:

$$|x_t - MAD_t| > \tau.$$

4.2 STL-ESD

Metoda STL (z ang. *Seasonal-Trend decomposition using Loess*) opiera się na dekompozycji addytywnej szeregu czasowego na trzy składowe w następującej formie

$$x_t = \tau_t + s_t + r_t,\ t = 1, \ldots, N$$

gdzie x_t jest obserwowaną wartością szeregu czasowego, τ_t jest składową trendu, s_t składową sezonowości, a r_t składową rezyduów. Dekompozycja taka jest dedykowana szeregom czasowym z zauważalnymi wolnozmiennymi fluktuacjami sezonowości oraz szybkozmienną składową trendu [6]. Zakłada się ponadto, że składowa rezyduów zawiera całą pozostałą informację o szeregu czasowym – m.in. szum. Można to sformalizować w następujący sposób:

$$r_t = a_t + \epsilon_t$$

przy czym składowa ϵ_t jest składową szumu, a a_t modeluje poszukiwane anomalie w postaci nagłych przyrostów wartości (t.zw. *peak*'ów). Dekompozycja STL jest procedurą iteracyjną i szczegółowo zostanie opisana w dokumentacji końcowej projektu. Na potrzeby tego dokumentu należy zaznaczyć, że wynikiem dekompozycji jest składowa a_t, a składowa:

- szumu ϵ_t jest usuwana wskutek operacji filtracji (ang. *denoising*) przy pomocy średniej/-mediany kroczącej lub probabilistycznemu wygładzaniu eksponencjalnego (PEWMA) [7]

- trendu τ_t jest usuwana wskutek operacji detrendyzacji, która w najprostszym wariancie sprowadza się do zastosowania operatora różnicowego

$$\nabla x_t = x_t - x_{t-1}$$

  lub do filtrów kroczących

- sezonowości s_t, która jest usuwana w standardowej wersji metody STL przy pomocy LOESS (ang. *locally estimated scatterplot smoothing*), tj. metody łączącej działanie średniej kroczącej z regresją wielomianową [8].

STL zawiera trzy parametry sterujące metodą:

1. n_p liczbę obserwacji, która jest rozważana w każdym cyklu wyliczania składowej sezonowości
2. n_i liczbę iteracji wewnętrznej pętli algorytmu
3. n_o liczbę iteracji pętli zewnętrznej

oraz trzy lub więcej parametrów sterujących detrendyzacją, usunięciem szumu oraz ekstrakcją sezonowości.

Po otrzymaniu składowej a_t stosowany jest test statystyczny ESD (z ang. _Extreme Studentized Deviate), który służy do wykrycia anomalii w próbie o rozkładzie *asymptotycznie* normalnym. Jest on uogólnieniem testu Grubbs'a, w którym rozkład normalny próby jest warunkiem koniecznym. Test ESD rozważa dwie hipotezy

H_0: **1** *w próbie* $\{a_1, \ldots, a_n\}$ *nie ma punktów odstających (anomalii)*

H_1: **1** *w próbie* $\{a_1, \ldots, a_n\}$ *jest co najwyżej* K *punktów odstających.*

i-ta statystyka testowa ESD wyliczana jest w następujący sposób:

$$T_i^{ESD} = \frac{\max_i |a_i - \bar{a}|}{\sigma}$$

gdzie $\bar{a}$ oznacza średnią z próby, a σ odchylenie standardowe z próby. Po wyliczeniu T_i^{ESD} usuwana jest z próby obserwacja maksymalizująca licznik i liczona jest kolejna statystyka testowa na podstawie zmodyfikowanej próby.

Po wyliczeniu $\{T_i^{ESD}\}_{i=1,\ldots,K}$ wyliczane są wartości krytyczne testu λ_i będące funkcją rozkładu t-studenta z $N - i$ stopniami swobody.

Liczbę anomalii w próbie znajduje się przez podanie największego i takiego, że $T_i^{ESD} > \lambda_i$.

W kontekście definicji procedury detekcji anomalii podanej (3) należy wybrać z próby i największych elementów i ich znaczniki umieścić w zbiorze T_A.

Dekompozycja szeregu czasowego przez algorytm STL znajduje się na poniższym obrazku.

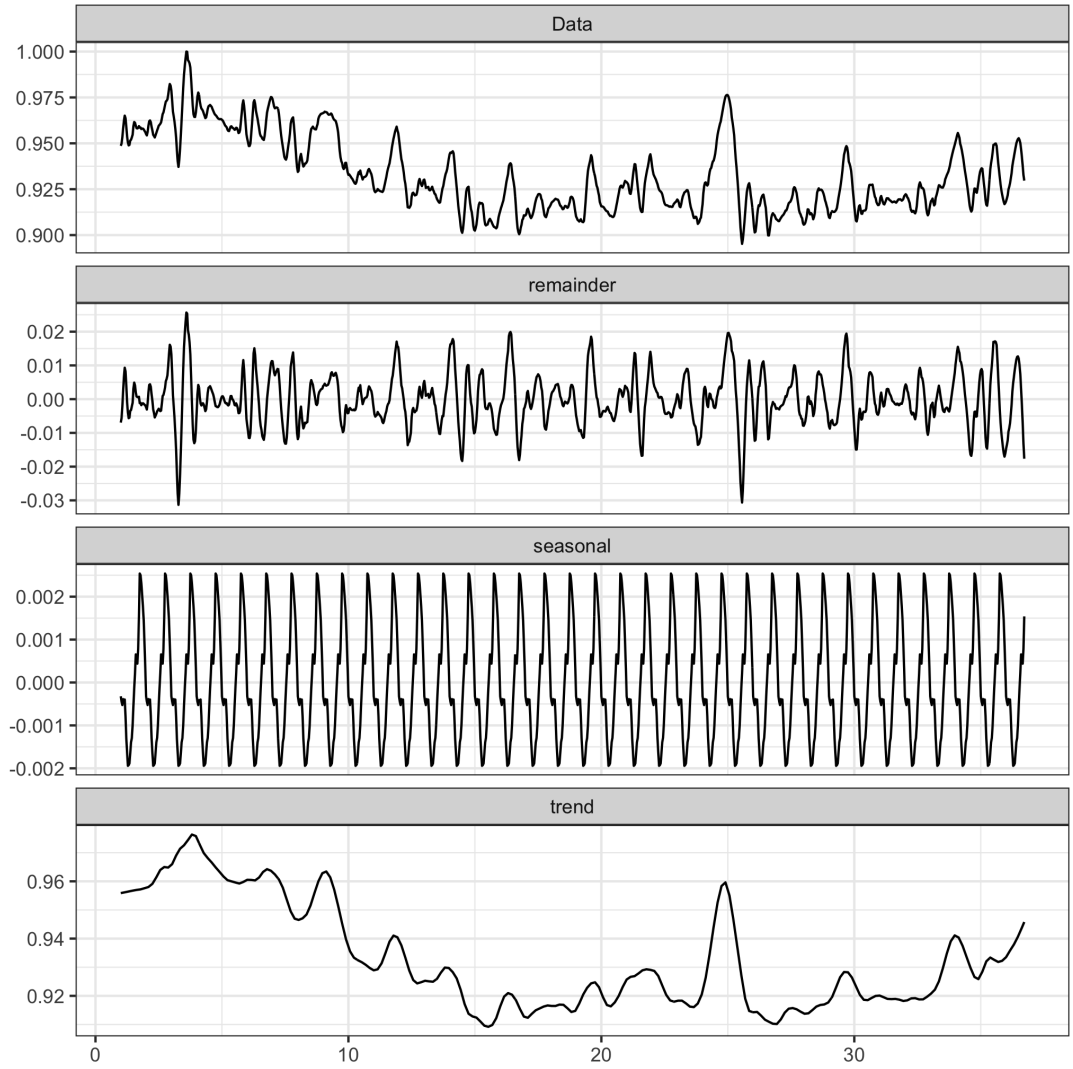

Figure 3: Dekompozycja szeregu czasowego metodą STL. Rysunek własny.

4.3 LDCD

Metoda LDCD (z ang. *Lazy Drifting Conformal Detector*) [9] jest procedurą detekcji anomalii opartą na algorytmie k-NN, która nie czyni żadnych założeń odnośnie rozkładu szeregu czasowego i jego właściwości.

Algorytm bazuje na kroczącym oknie historii o długości l, które zanurza (ang. *embed*) szereg czasowy w przestrzeni l-wymiarowej. Operacja taka umożliwia zastosowanie metod wielowymiarowych do przyporządkowania punktom szeregu czasowego punktów niedopasowania (z ang. *non-conformity score*) α_t. Punkty te są przyznawane w oparciu o metodę k-NN. Metoda ta ponadto utrzymuje w trakcie swojego działa dwa zbiory:

- zbiór referencyjny T_t o wielkości n

- zbiór korekcyjny A_t o wielkości m

przy czym $t \geq m + n$.
W trakcie działania algorytmu każdemu przetwarzanemu punktowi x_t przyporządkowywany jest punkt α_t, który służy do wyliczenia p-wartości:

$$p(x_t, A_t) = \frac{1}{m+1}|\{i = 1, \ldots, m | \alpha_{t-i} \geq \alpha_t\}|.$$

Na jej podstawie przyznawany jest z kolei t.zw. *conformal abnormality score*, tj.

$$p(t)_{ab} = 1 - p(x_t, A_t)$$

Punkt przyznany przetwarzanemu przykładowi umieszczany jest na czele kolejki A_t, a jej ogon jest zastępowany przez punkt ówcześnie poprzedzający go.

Pełniejsze znaczenie obu struktur i dokładne działanie algorytmu zostanie omówione w dokumetancji końcowej.

5 Charakterystyka zbioró wdanych

Zbiory danych użyte do analizy działania algorytmów zostały częściowo omówione w rozdziale 2. W niniejszym rozdziale zostaną wskazane źródła, z których pochodzą dane, oraz podana zostanie krótka charakterystyka tych danych.

Kompletna analiza statystyczna danych, na którą będzie składało się zbadanie rozkładu danych, autokorelacji szeregu czasowego lub jego innych właściwości jak stacjonarność zostanie umieszczona w dokumentacji końcowej.

5.1 WIG20

Historyczne dane `WIG20` dostępne są do pobrania ze strony stooq.pl z interwałem:

- dziennym
- tygodniowym
- miesięcznym
- kwartalnym
- rocznym.

Dane pobierane są w formacie `CSV` i nie zawierają wartości brakujących. Struktura zbioru danych przedstawiona jest na poniższym obrazku.

```
Rows: 2,143
Columns: 2
$ price <int> 195600, 195037, 194398, 193379, 192695, 192146, 191602, 191378,…
$ time  <dttm> 2020-03-14 10:30:21, 2020-03-14 10:50:21, 2020-03-14 11:10:21,…
```

Figure 4: Wynik wypisania na konsolę fragmentu zbioru danych.

Należy jednak zaznaczyć, że rozważaną w projekcie wartością szeregu czasowego jest uśredniona wartość indeksu z wartości otwarcia, zamknięcia oraz wartości najmniejszej i największej danej sesji.

Wartości indeksu WIG20 dostępne są od 16 kwietnia 1991 do dnia dzisiejszego (02.04.2020) z tą uwagą, że nie we wszystkich przypadkach zachowany jest wymagany interwał, co jest zaznaczone na rysunku poniżej.

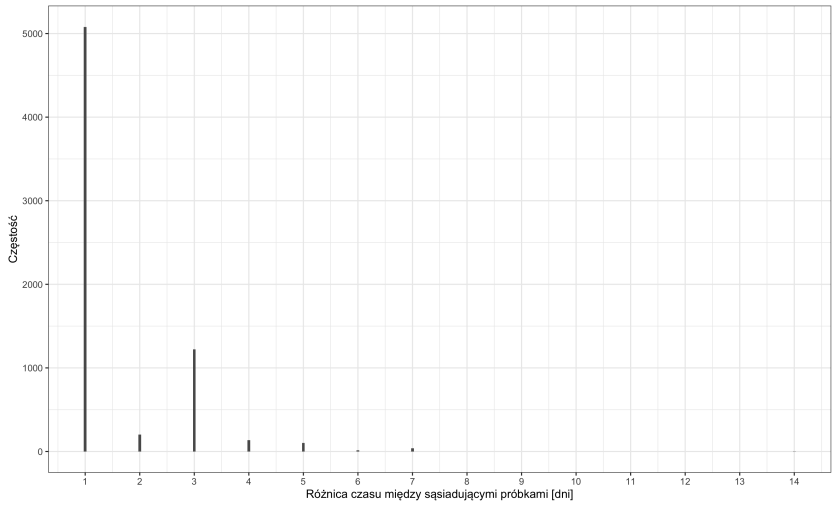

Figure 5: Histogram odstępów między sąsiadującymi próbkami w zbiorze danych WIG20. Rysunek własny.

W związku z tym do badań zostanie użyty najdłuższy podciąg, w którym odstęp między danymi wynosi jedną dobę.

5.2 WoW Token

Historyczne kursy tokena dostępne są na stronie wowtokenprices.com. Serwis udostępnia API, które umożliwia pobrania danych z kwantem 20-minutowym, które nie zawierają wartości brakujących, od początku istnienia tokena, tj. od roku 2015. Kurs tokena różni się między regionami, w których znajdują się serwery. Dostępnych jest 5 regionów, tj.

1. europejski
2. amerykański
3. chiński
4. tajwański
5. koreański.

Struktura zbioru danych z ostatniego miesiąca widoczna jest na na poniższym obrazku.

```
Rows: 2,143
Columns: 2
$ price <int> 195600, 195037, 194398, 193379, 192695, 192146, 191602, 191378,…
$ time  <dttm> 2020-03-14 10:30:21, 2020-03-14 10:50:21, 2020-03-14 11:10:21,…
```

Figure 6: Wynik wypisania na konsolę fragmentu zbioru danych.

6 Plan badań

Celem projektu jest zbadanie i porównanie metod detekcji anomalii w szeregach czasowych przy pomocy danych scharakteryzowanych we wcześniejszych rozdziałach.

Przyjęta metodologia zakłada, że zadanie detekcji anomalii zostanie potraktowane jak zadanie klasyfikacji binarnej. Oznacza to, że przykłady w zbiorze danych zostaną oznaczone flagą, która informuje o tym czy dany przykład jest anomalią czy też nie.

Przydział flagi będzie odbywał się na dwa sposoby:

1. autorzy manualnie wybiorą podzbiór zdarzeń, których zajście spowodowało znaczny przyrost lub spadek wartości indeksu/tokena i oznaczą je jako anomalie. Przykłady takich zdarzeń są następujące:

- kryzys gospodarczy z 2008 roku (ogłoszenie upadłości przez bank *Lehman Brothers*) (WIG20)
- rozpoczęcie stanu epidemicznego w Polsce w związku z pandemią koronawirusa SARS-CoV-2 (WIG20)
- atak Iranu na bazy wojskowe USA na początku 2020 roku (WIG20)
- zwiększenie współczynnika przyrostu zdobywanego doświadczenia o 100% (WoW Token)
- wydanie nowej części gry *World of Warcraft* (WoW Token).

2. do zbiorów danych zostaną wprowadzone anomalie w sposób losowy.

Dzięki takiemu podejściu możliwe będzie porównanie metod detekcji w dość ogólnym kontekście (dane równie dobrze mogłyby być syntetycznie generowane) oraz sprawdzenie jak metody radzą sobie z wykrywaniem anomalii, które odpowiadają przełomowym zdarzeniom i mają odzwierciedlenie na rynkach. Przyjęcie takiej metodologii podyktowane jest faktem, że w dostepnych zbiorach danych autorom nie udało się znaleźć zbioru poświęconemu aktywom giełdowym.

Porównanie modeli odbędzie się przy pomocy trzech miar:

1. precyzji
$$P = \frac{|S \cap G|}{|S|}$$
2. odzysku (z ang. *recall*)
$$R = \frac{|S \cap G|}{|G|}$$
3. miary $F - 1$
$$F = 2\frac{PR}{P + R}$$
będącej średnią harmoniczną P i R

przy czym S jest zbiorem poprawnie rozpoznanych anomalii, a G jest zbiorem wszystkich anomalii w zbiorze danych.

Ponadto zbadane zostaną różne nastawy parametrów sterujących użytych metod jak:

- sposób dokonywania dekompozycji sygnału w metodzie STL
  - zastąpienie składowej trendu medianą [10]

  - użycie różnych metod usunięcia szumu, tj. np. SMA oraz PEWMA

- wpływ szerokości okna na działanie metody MAD
- wpływ wielkości kolejki strojenia (*calibration queue*), długości zbioru referencyjnego oraz parametru metody NN na działanie algorytmu CAD k-NN.

References

[1] B. G. Malkiel, *A Random Walk Down Wall Street*. Norton, New York, 1973.

[2] G. Campani, "Sapiens: A brief history of humankind di yuval noah harari, harvill secker, london, 2014," *Comparative Cultural Studies - European and Latin American Perspectives*, vol. 2, pp. 113–114, Jul. 2018.

[3] "High frequency trading," 2020.

[4] "Wow token prices," 2015-2020.

[5] A. Blázquez-García, A. Conde, U. Mori, and J. Lozano, "A review on outlier/anomaly detection in time series data," 02 2020.

[6] Q. Wen, J. Gao, X. Song, L. Sun, H. Xu, and S. Zhu, "Robuststl: A robust seasonal-trend decomposition algorithm for long time series," 2018.

[7] K. Carter and W. Streilein, "Probabilistic reasoning for streaming anomaly detection," pp. 377–380, 08 2012.

[8] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess (with discussion)," 1990.

[9] V. Ishimtsev, I. Nazarov, A. Bernstein, and E. Burnaev, "Conformal k-nn anomaly detector for univariate data streams," 2017.

[10] M. Gander, M. Felderer, B. Katt, A. Tolbaru, R. Breu, and A. Moschitti, "Anomaly detection in the cloud: Detecting security incidents via machine learning," 01 2013.