WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Gabriel Lichacz, Patryk Motyka

Analiza statystyczna przepływów w kontekście systemów złożonych dla przykładowego ruchu przy pomocy pakietu R

Rzeszów, 2021

1. Spis treści

2. Wstęp

Celem badań była analiza ruchu sieciowego Information Security Talent Search z 7 marca 2015 roku. W części teoretycznej omówiono podstawowe zagadnienia związane ze statystycznymi miarami. W części badawczej dokonano analizy przepływu ruchu w sieci komputerowej na podstawie liczby pakietów w czasie. Wykonano analizę wykładnika Hurst'a, Korelacji, Kurtozy, współczynnika skośności, Giniego i zmienności.

3. Teoria

3.1. Wykładnik Hursta

Jest charakterystyką nieliniowych szeregów czasowych. Wykładnik przyjmuje wartości z przedziału [0, 1]. Wartość współczynnika H=0,5 oznacza losową serię czasową. Jeżeli H<0,5, mamy do czynienia z serią antypersystentną, która charakteryzuje się tym, że wartości górne są poprzedzone wartościami dolnymi i na odwrót. Natomiast H>0,5 oznacza serię persystentą, taką która posiada trend wzmacniający, co oznacza, że następna wartość jest zbliżona do poprzedzającej.

3.2. Kurtoza

Jest jedną z miar spłaszczenia rozkładu wartości cechy, czyli koncentracji wyników wokół średniej wartości. Dodatnia wartość kurtozy (rozkład leptokurtyczny) wskazuje na istnienie wielu wartości bliskich średniej, natomiast ujemna wartość kurtozy (rozkład platykurtyczny) wskazuje na większe rozproszenie wyników.

$$\text{Kurt} = \frac{\mu_4}{\sigma^4},$$

gdzie: μ_4 - czwarty moment centralny, σ - odchylenie standardowe.

3.3. Współczynnik skośności

Miara symetrii lub asymetrii rozkładu. Gdy wartość skośności wynosi 0, rozkład jest idealnie symetryczny. Wartości ujemne wskazują na rozkład lewoskrętny a dodatnie na prawoskrętny.

$$A_d = \frac{\mu - d}{s}$$
$$A_m = 3\frac{\mu - m}{s}$$
$$A_Q = \frac{Q_1 + Q_3 - 2m}{2Q} = \frac{Q_1 + Q_3 - 2m}{Q_3 - Q_1},$$

gdzie μ - średnia arytmetyczna, m - mediana, d - dominanta, s - odchylenie standardowe, Q_1, Q_2 - pierwszy i trzeci kwartyl, Q - odchylenie ćwiartkowe.

3.4. Współczynnik Giniego

Stosowana w statystyce miara koncentracji rozkładu zmiennej losowej.

$$G(y) = \frac{\sum_{i=1}^{n}(2i - n - 1)y_i}{n^2\overline{y}},$$

gdzie: . $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

3.5. Współczynnik zmienność

Powszechnie stosowana miara zróżnicowania cechy. Wysoka wartość współczynnika oznacza duże zróżnicowanie cechy i świadczy o niejednorodności badanej populacji, niska wartość świadczy o małej zmienności cechy i jednorodności badanej populacji.

$$V = \frac{s}{\overline{x}}, \quad \overline{x} \neq 0,$$

gdzie: s - odchylenie standardowe, $\bar{x}$ - średnia arytmetyczna z próby

3.6. Korelacja (współzależność)

Współczynnik korelacji liniowej Pearsona – współczynnik określający poziom zależności liniowej między zmiennymi. Można spotkać kilka rodzajów klasyfikacji korelacji. My użyliśmy klasyfikacji:

$0{,}0 \leq |r| \leq 0{,}1$ – brak korelacji

$0{,}1 < |r| \leq 0{,}3$ - korelacja słaba

$0{,}3 < |r| \leq 0{,}5$ - korelacja umiarkowana

$0{,}5 < |r| \leq 0{,}7$ - korelacja silna

$0{,}7 < |r| \leq 0{,}9$ - korelacja bardzo silna

$0{,}9 < |r| < 1{,}0$ - korelacja niemal pełna

$|r| = 1$ - korelacja pełna

Ujemna korelacja oznacza, że wzrost jednej zmiennej oznacza spadek drugiej.

3.7. Regresja liniowa

Zakłada, że zależność pomiędzy zmienną objaśnianą a objaśniająca jest zależnością liniową. Jak w analizie korelacji, jeżeli jedna wartość wzrasta to druga wzrasta (dodatnia korelacji) lub spada (korelacja ujemna). Model regresji liniowej ma ogólną postać kombinacji liniowej wyrazów:

$$Y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_n\beta_n + \varepsilon$$

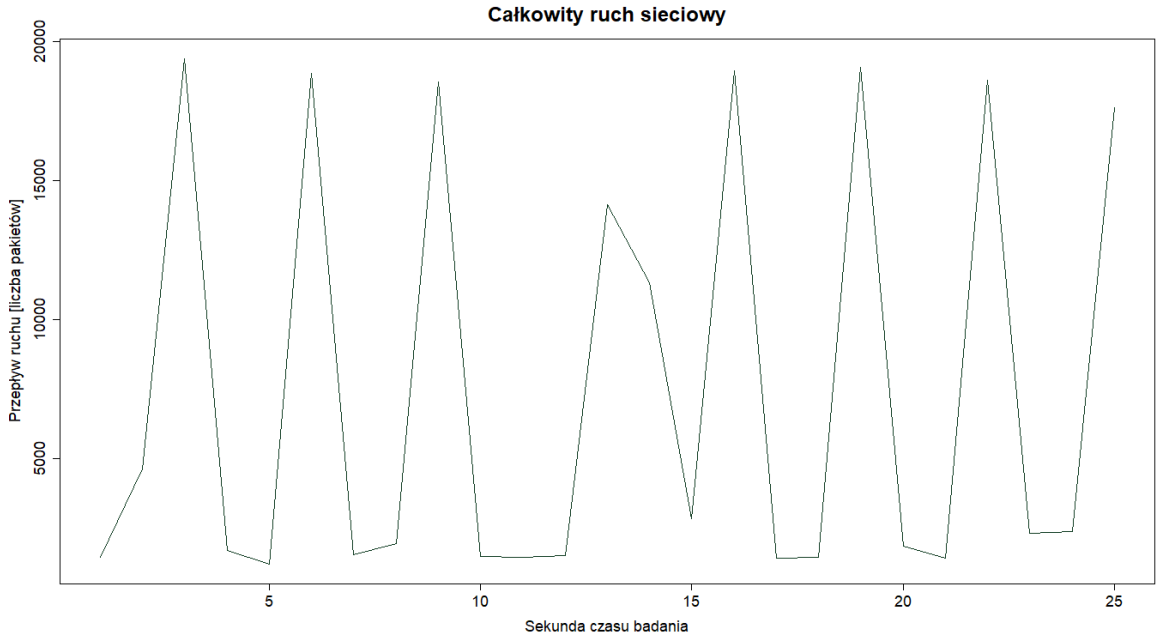
3.8. Gęstość

Jest to miara probabilistyczna określona na σ-ciele podzbiorów borelowskich pewnej przestrzeni metrycznej. Rozkład prawdopodobieństwa określa prawdopodobieństwo przyjęcia każdej możliwej wartości przez zmienną losową (jeśli jest ona dyskretna), lub jej prawdopodobieństwo znalezienia się w konkretnym przedziale (jeśli jest ciągła).

4. Badanie ruchu w sieci komputerowej

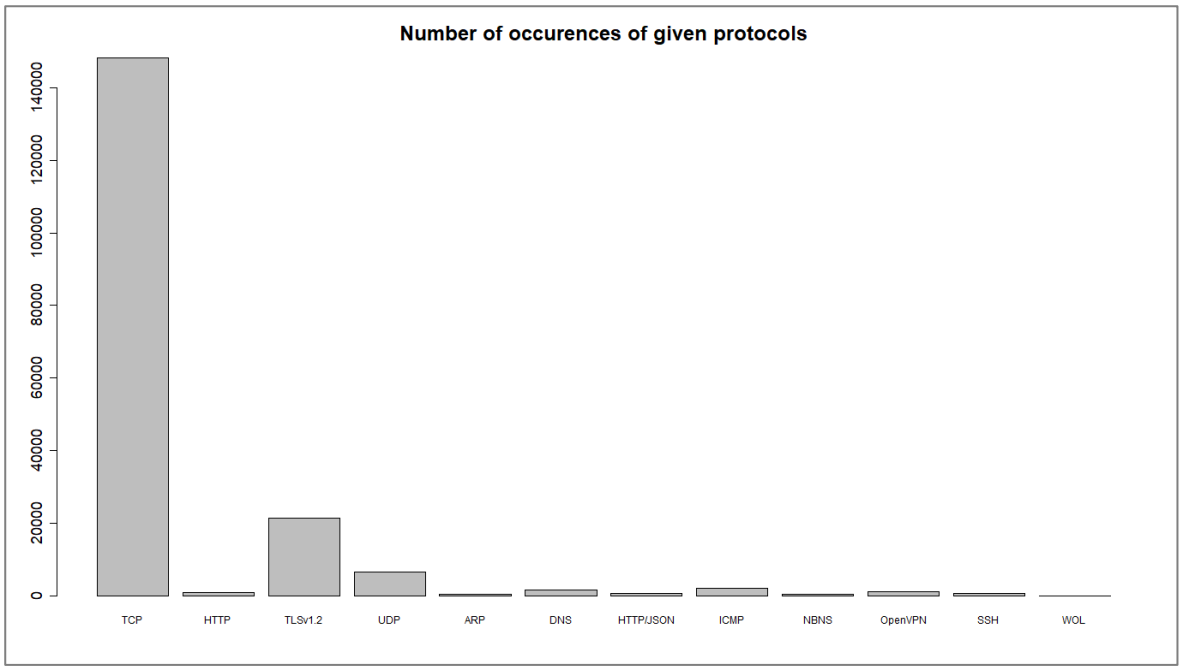
4.1. Analiza przepływu całego ruchu sieciowego

Przepływ całego zapisanego i poddanego analizie ruchu sieciowego:



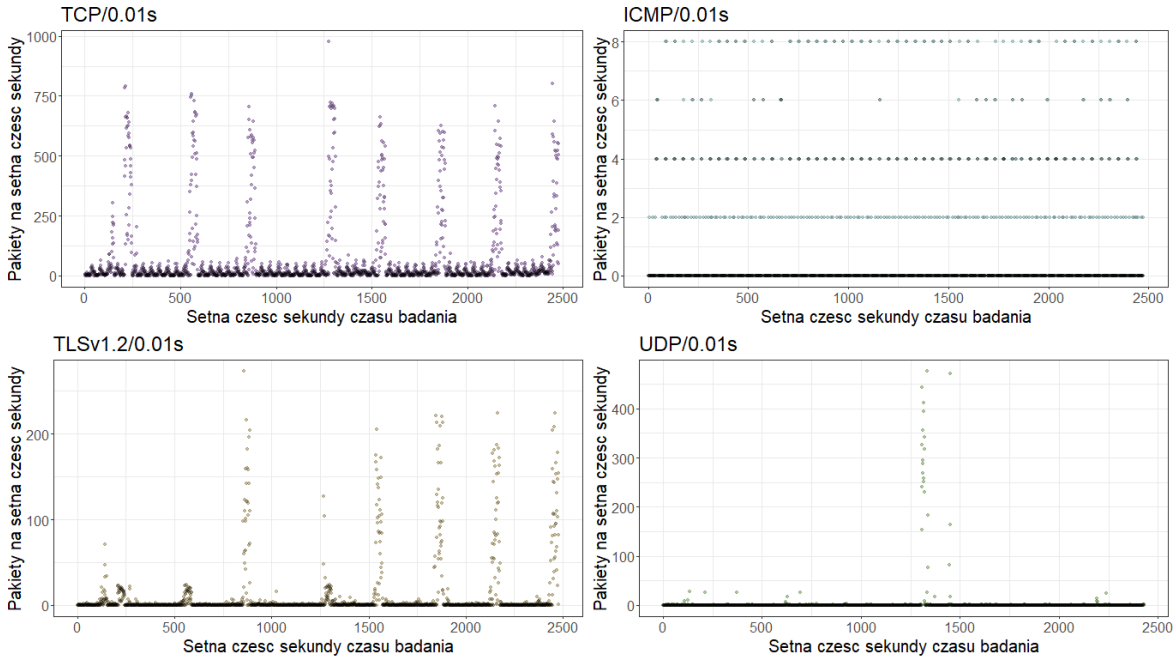
Rys. 4.1 Całkowity ruch sieciowy

Najmniejsza zanotowana liczba pakietów przesłanych w ciągu jednej sekundy wyniosła 1 216, a największa 19 373. Dla całego ruchu sieciowego zmiana wyniosła 18 157 pakietów. Ruch jest dość zmienny, co potwierdzają kolejne miary statystyczne.



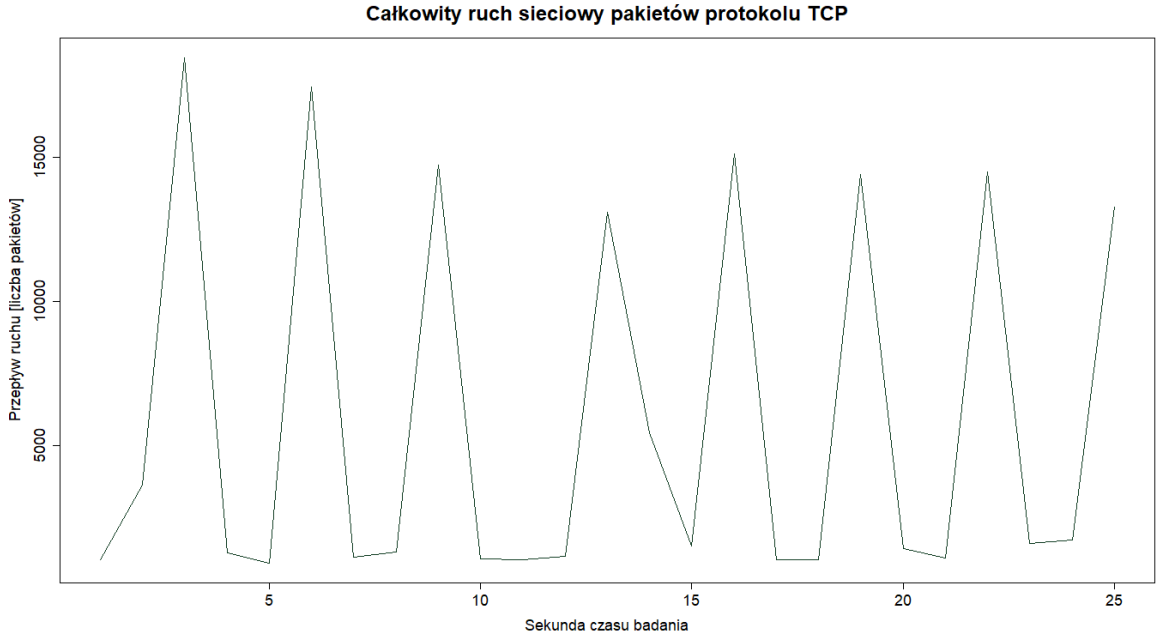
Rys. 3.2 Udział wszystkich protokołów sieciowych w ruchu.

Większość protokołów sieciowych nie miała większego wpływu na przepływ ruchu sieciowego - poniżej zestawienie najbardziej wpływowych:



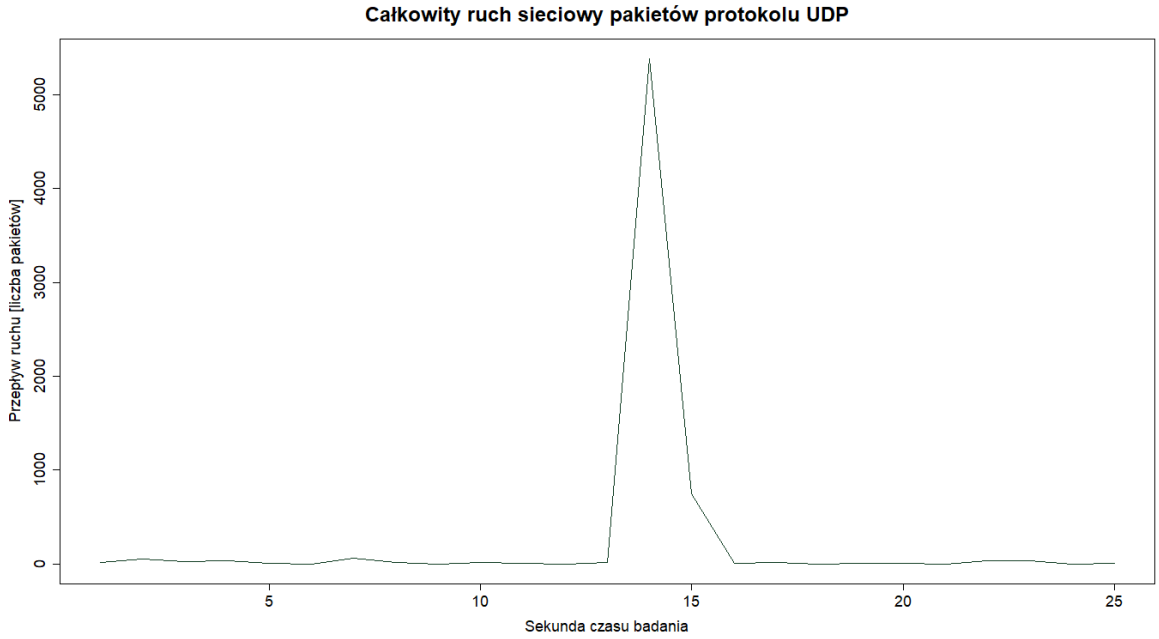
Rys. 3.3 Ilość pakietów przesłanych przez dane protokoły w czasie 0,01 sekundy

4.1.1. Analiza przepływu wszystkich pakietów protokołu TCP



Rys. 3.4 Całkowity ruch sieciowy pakietów protokołu TCP

4.1.2. Analiza przepływu wszystkich pakietów protokołu UDP

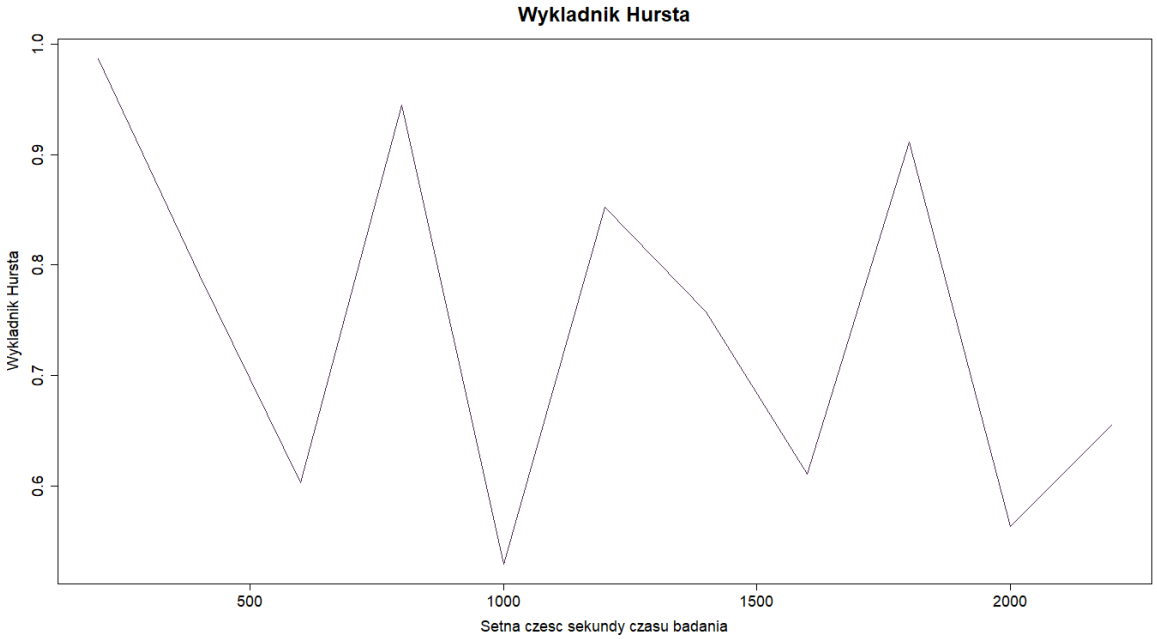


Rys. 3.5 Całkowity ruch sieciowy pakietów protokołu TCP

Protokół UDP wykorzystywany był ekstensywnie tylko w krótkim okresie czasowym, przez resztę ruchu nie miał większego wpływu.

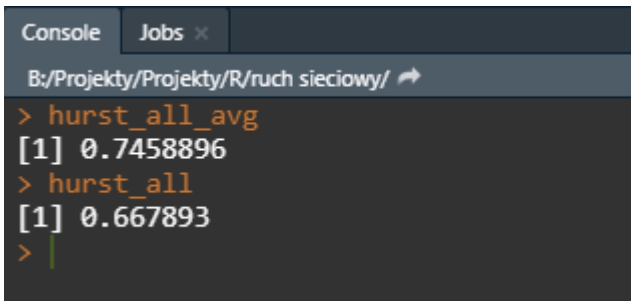
4.2. Analiza serii czasowych z wykorzystaniem wykładnika Hurst'a dla całego ruchu sieciowego

Dane zostały podzielone na okna czasowe co 1/100 sekundy. Współczynnik obliczony co 200 okien czasowych (2 sekundy). Pozwoliło to na uzyskanie realnej wartości dla mniejszych przedziałów. Minimalna wartość wyniosła 0.5293177 a maksymalna 0.9866478.



Rys. 3.6 Wykładnik Hursta dla okien czasowych całego ruchu sieciowego

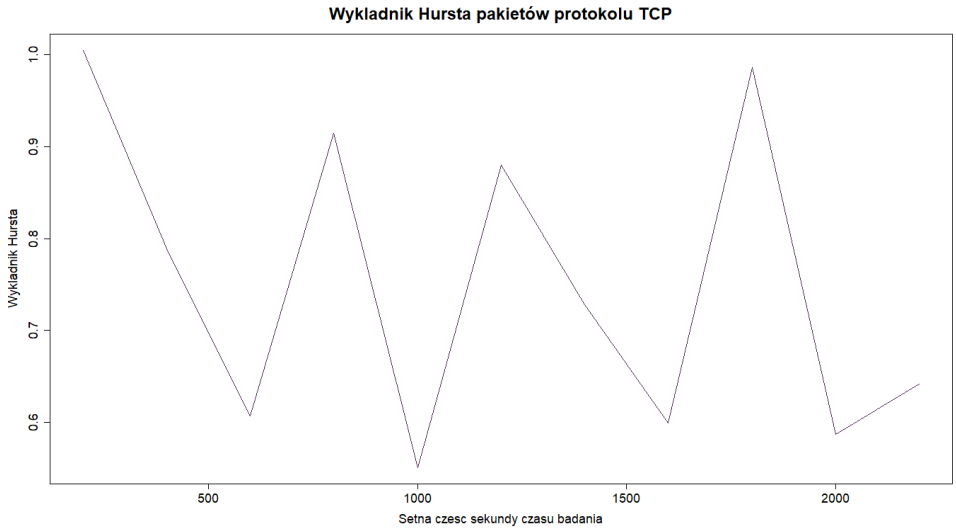
Otrzymane wyniki świadczą o tym, że ruch jest persystentny czyli o rosnącym trendzie ruchu, ponieważ wartości wykładnika wahają się od 0,5 do prawie 1.



Rys. 3.7 Średnia wartość wykładnika Hursta z okien czasowych oraz wartość wykładnika liczona z całego przedziału

Wartość wykładnika Hursta policzona z całości ruchu sieciowego wynosi około 0,67, więc nie odbiega znacząco od wartości liczonych na oknach czasowych, których średnia wynosi około 0,74.

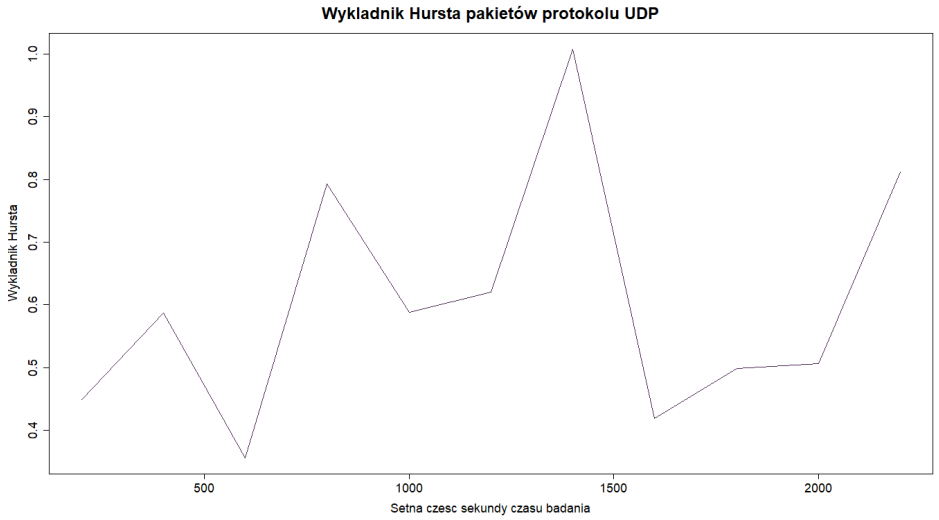
4.2.1. Analiza serii czasowych z wykorzystaniem wykładnika Hurst'a dla pakietów protokołu TCP



Rys. 3.8 Wykładnik Hursta dla pakietów protokołu TCP

Wyniki otrzymane dla pakietów protokołu TCP są stosunkowo podobne do tych otrzymanych dla całego ruchu sieciowego, ponieważ jest to protokół dominujący. Zachowuje się trend rosnący, wartość wykładnika Hursta wynosi dla każdego okna ponad 0.5.

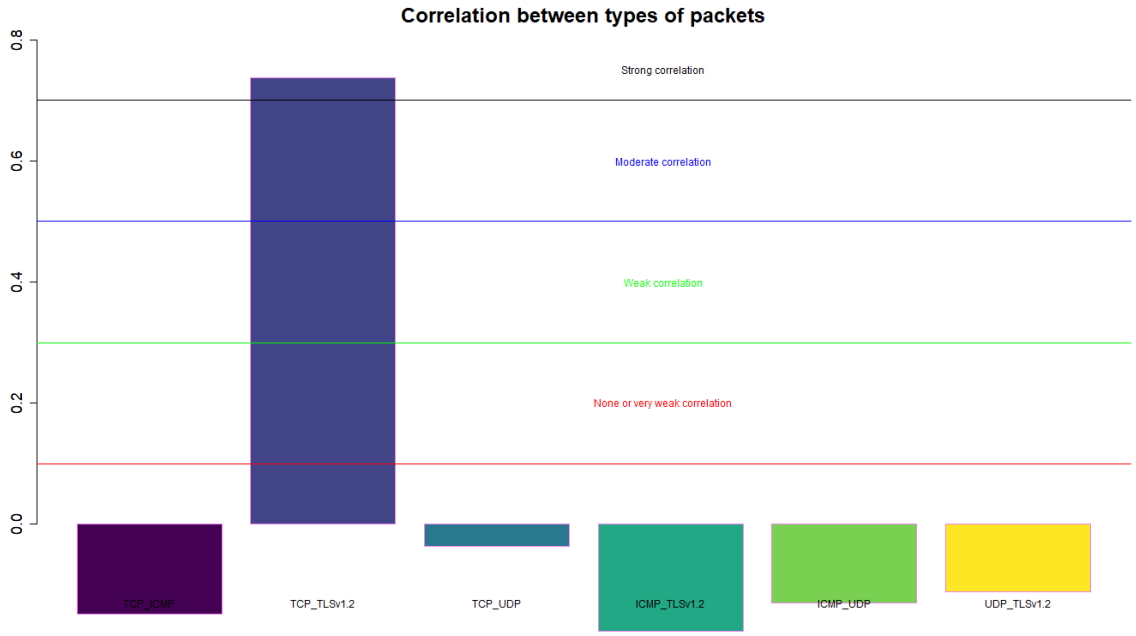
4.2.2. Analiza serii czasowych z wykorzystaniem wykładnika Hurst'a dla pakietów protokołu UDP



Rys. 3.9 Wykładnik Hursta dla pakietów protokołu UDP

Wartości wykładnika Hursta dla pakietów protokołu oscylują od 0,3 do prawie, więc nie można jednoznacznie określić, że trend jest rosnący lub malejący. Spowodowane jest to nierównomiernością ruchu tego protokołu.

4.3. Wartości współczynników korelacji prostych między poszczególnymi typami protokołów sieciowych

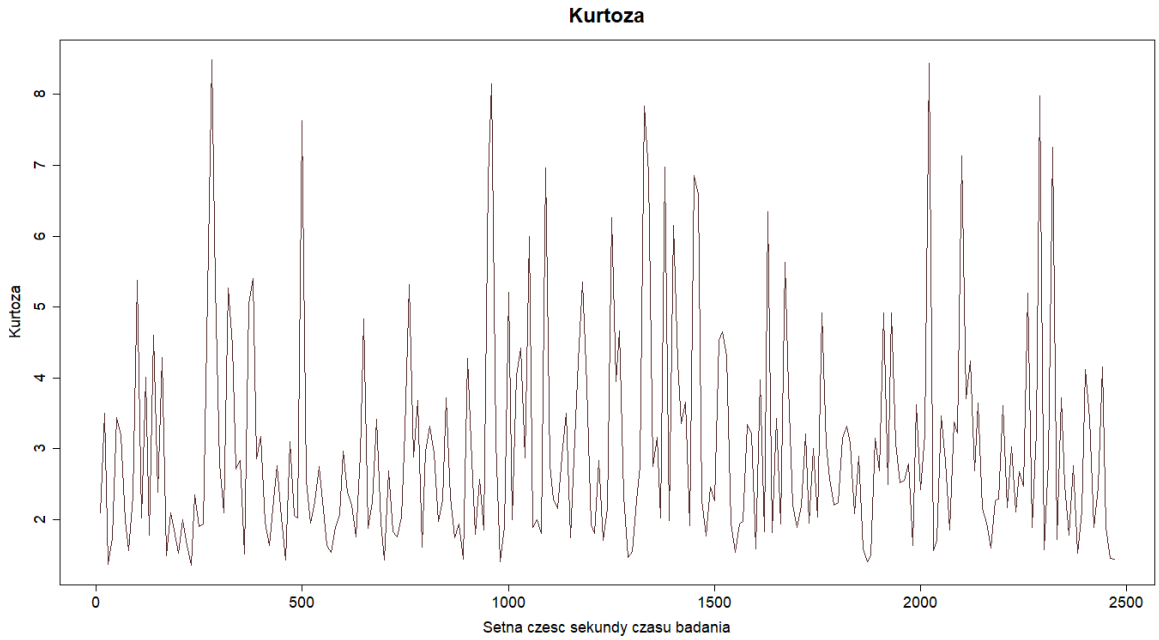


Rys. 3.10 Korelacje między rodzajami pakietów

Do liczenia korelacji użyty został współczynnik Pearsona.

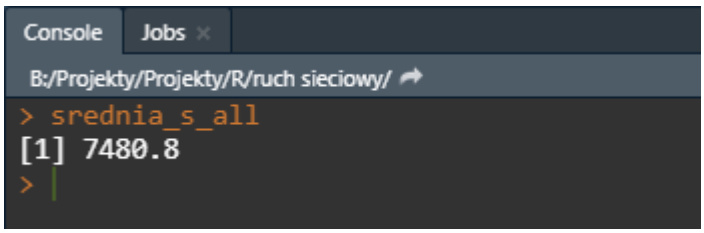
Największe korelacje zachodzą pomiędzy pakietami wykorzystującymi protokoły TCP i TLSv1.2. Zależność między resztą jest ujemna, ale jej wartość bezwzględna nie przekracza 0,1 co oznacza, że korelacja nie zachodzi.

4.4. Analiza serii czasowych z wykorzystaniem Kurtozy dla całego ruchu sieciowego



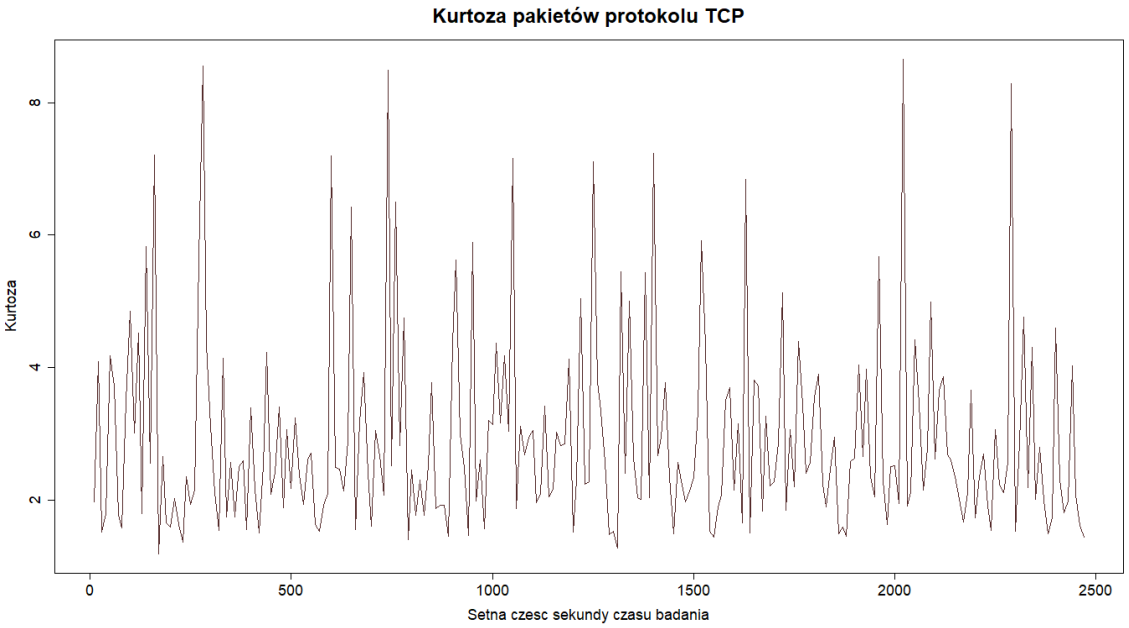
Rys. 3.11 Kurtoza dla okien czasowych całego ruchu sieciowego

Dane zostały podzielone na okna czasowe co 1/100 sekundy Minimalna wartość kurtozy wyniosła 1.362668 a maksymalna 8.488389. Rozkład jest leptokurtyczny, ponieważ wartości kurtozy są zawsze dodatnie. Oznacza to, że wartości są blisko średniej. Wartość średnia ilości pakietów na sekundę wynosi około 7480,8.



Rys. 3.12 Wartość średnia ilości pakietów na sekundę

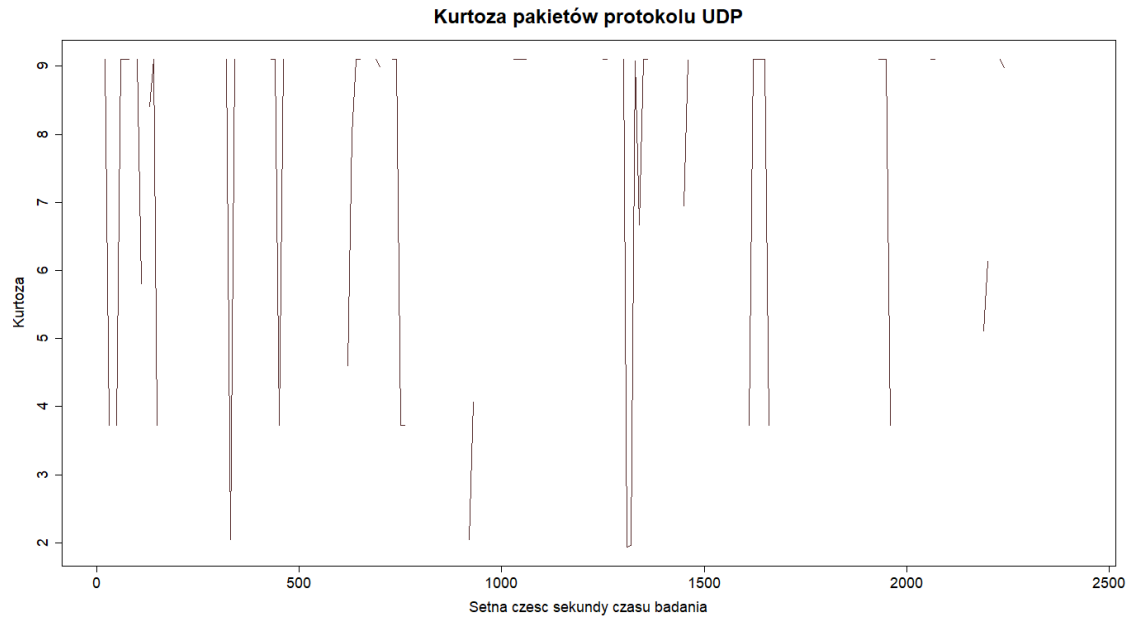
4.4.1. Analiza serii czasowych z wykorzystaniem Kurtozy pakietów protokołu TCP



Rys. 3.13 Kurtoza dla pakietów protokołu TCP

W przypadku pakietów protokołu TCP rozkład także jest leptokurtyczny, ponieważ wartości kurtozy są zawsze dodatnie. Oznacza to, że wartości są blisko średniej. Wartości kurtozy w tym przypadku są bardzo podobne do wartości z całego ruchu.

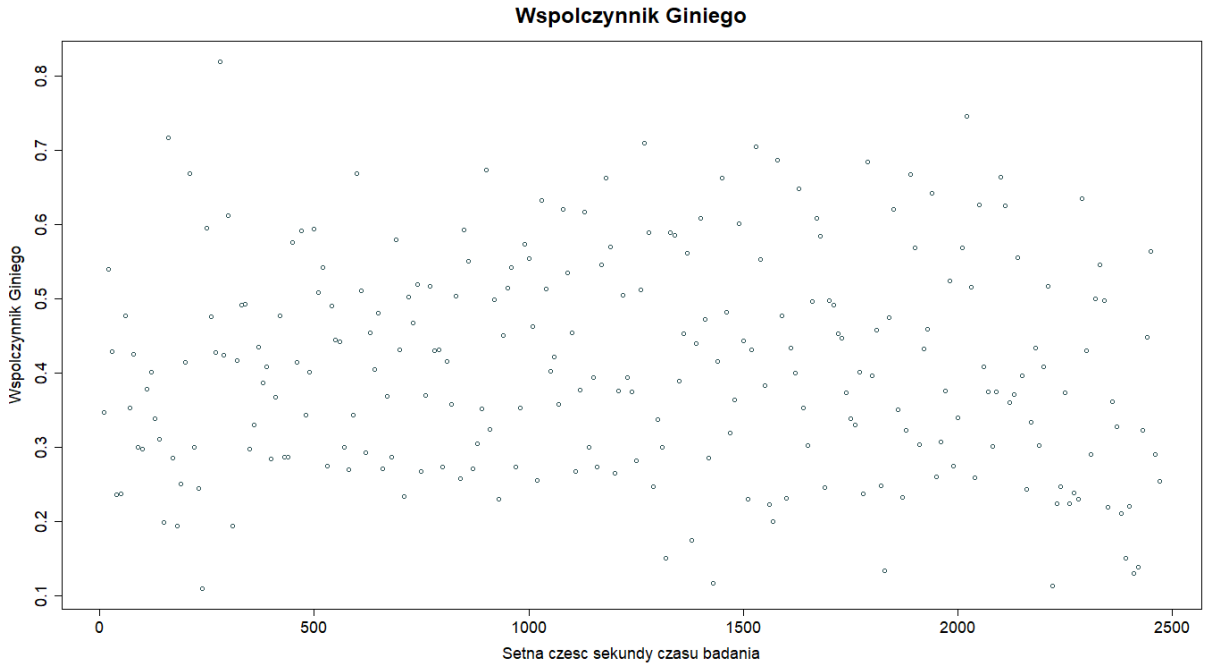
4.4.2. Analiza serii czasowych z wykorzystaniem Kurtozy pakietów protokołu UDP



Rys. 3.14 Kurtoza dla pakietów protokołu UDP

Dla pakietów protokołu UDP możemy dojść do tego samego wniosku, co dla TCP, mamy doczynienia z rozkładem leptokurtycznym - brak wartości poniżej zera. Wartości nie są jednak równomierne ze względu na krótkie występowanie protokołu UDP w całym ruchu.

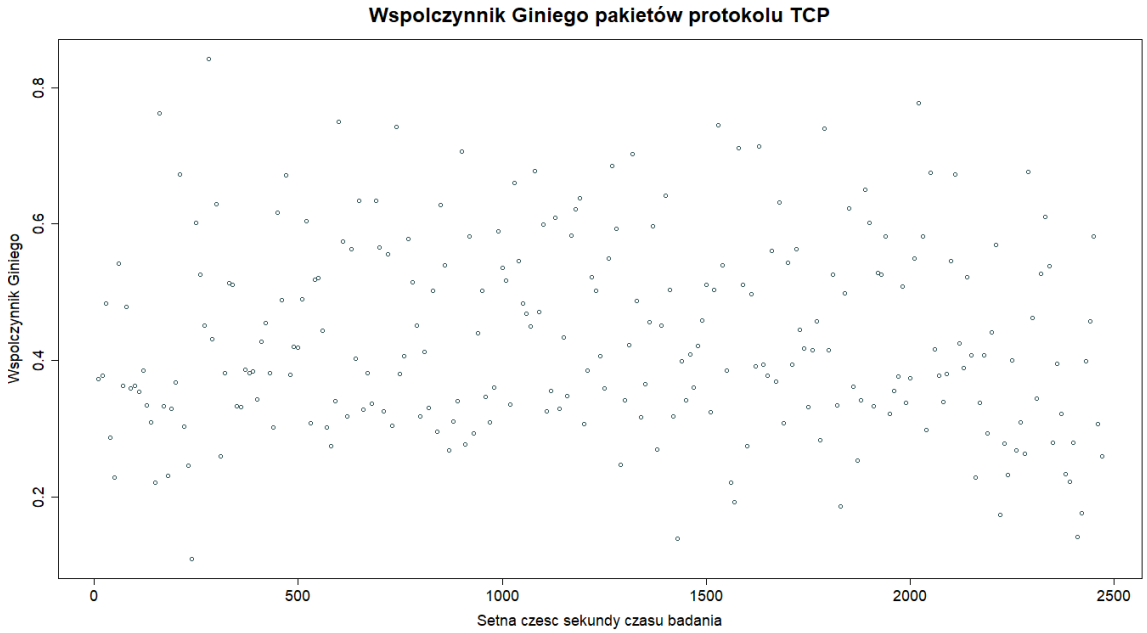
4.5. Współczynnik Giniego dla całego ruchu sieciowego



Rys. 3.15 Wartość współczynnika Giniego dla okien czasowych całego ruchu sieciowego

Dane zostały podzielone na okna czasowe co 1/100 sekundy. Wartości współczynnika Giniego dla wybranych okien czasowych różnią się znacząco co wskazuje na dużą nierównomierność rozkładu. Maksymalna wartość wynosi 0.8187839 a minimalna 0.1102941.

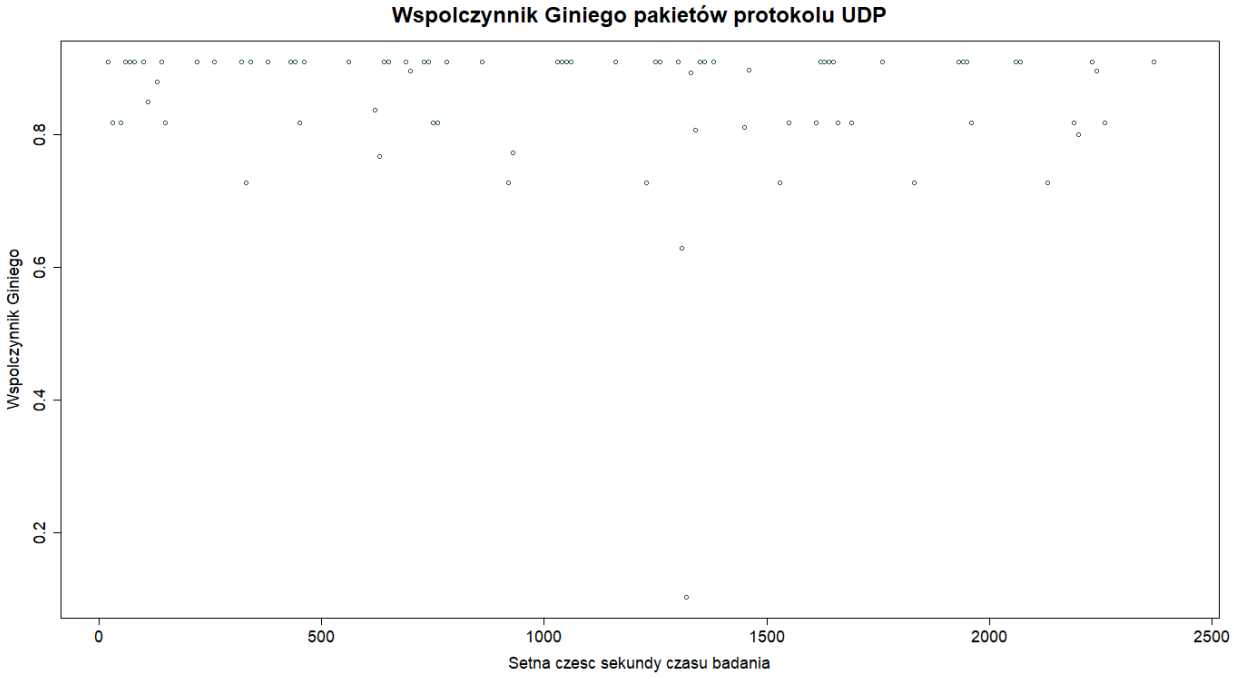
4.5.1. Współczynnik Giniego dla pakietów protokołu TCP



Rys. 3.16 Wartość współczynnika Giniego dla pakietów protokołu TCP

Wartości współczynnika Giniego dla pakietów TCP sytuacja nie różni się znacząco od wartości dla całego ruchu - ruch jest nierównomierny.

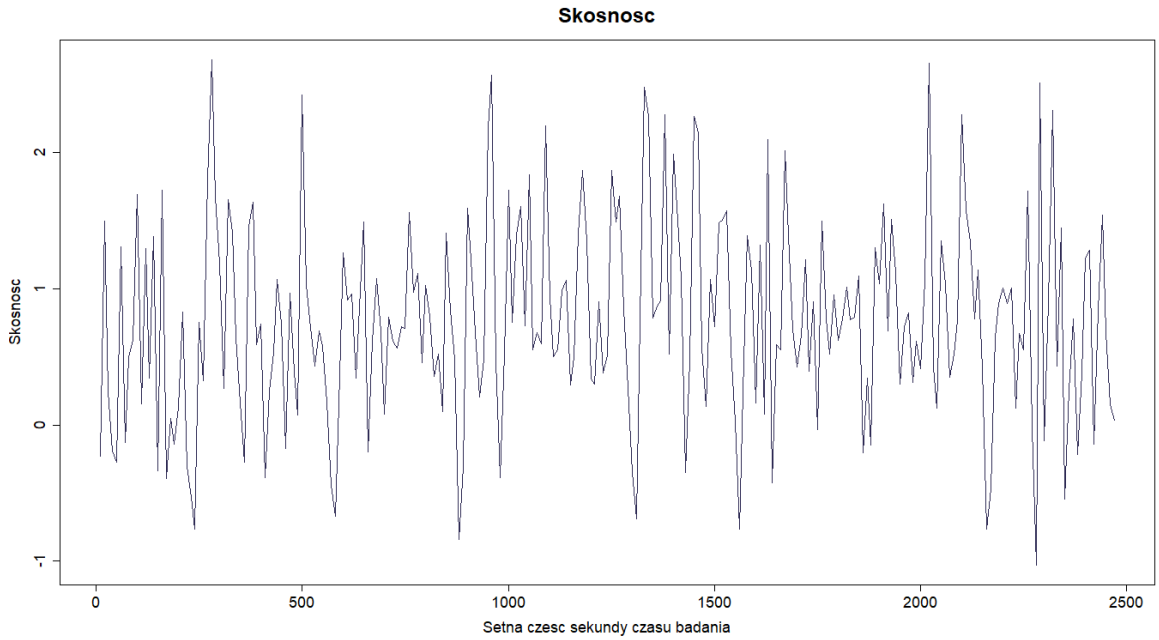
4.5.2. Współczynnik Giniego dla pakietów protokołu UDP



Rys. 3.17 Wartość współczynnika Giniego dla pakietów protokołu UDP

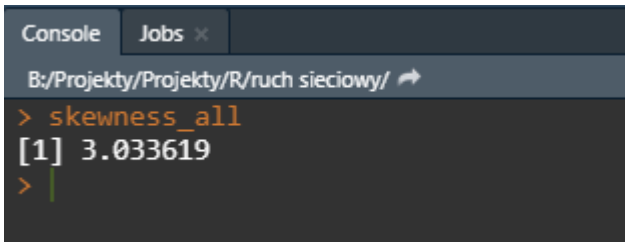
Dla pakietów UDP ruch jest bardzo nierównomierny, wartości dochodzą do nawet 0.9090909.

4.6. Analiza serii czasowych z wykorzystaniem współczynnika skośności dla całego ruchu sieciowego



Rys. 3.18 Skośność dla okien czasowych całego ruchu sieciowego

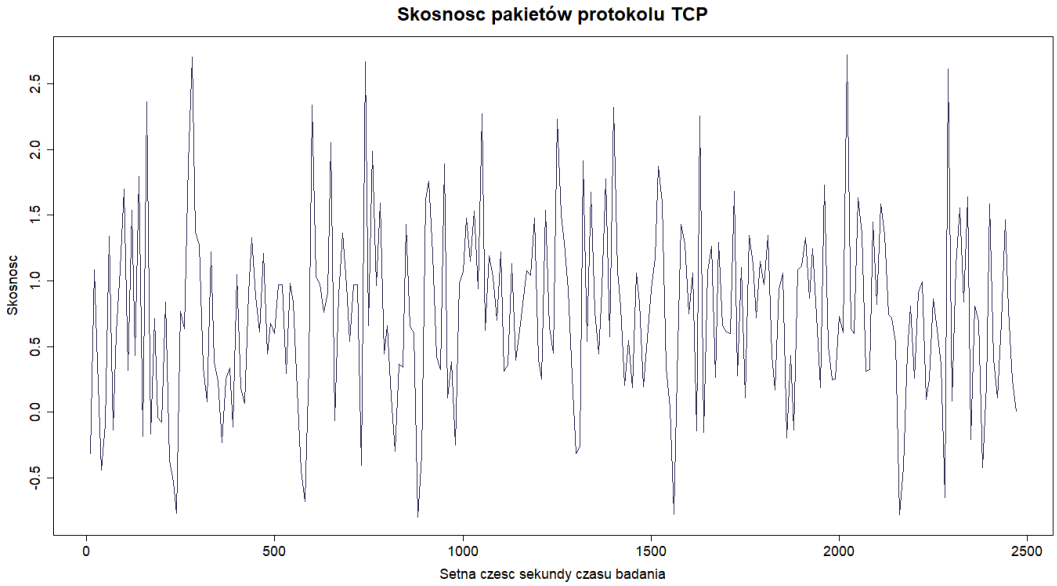
Dane zostały podzielone na okna czasowe co 1/100 sekundy. Minimalna wartość współczynnika skośności wyniosła -1.02565753 a maksymalna 2.68260827. Co oznacza, że na różnych oknach czasowych rozkład jest lewoskrętny bądź prawoskrętny.



Rys. 3.19 Wartość skośności dla całego przedziału

Cały rozkład jednakże jest prawoskrętny.

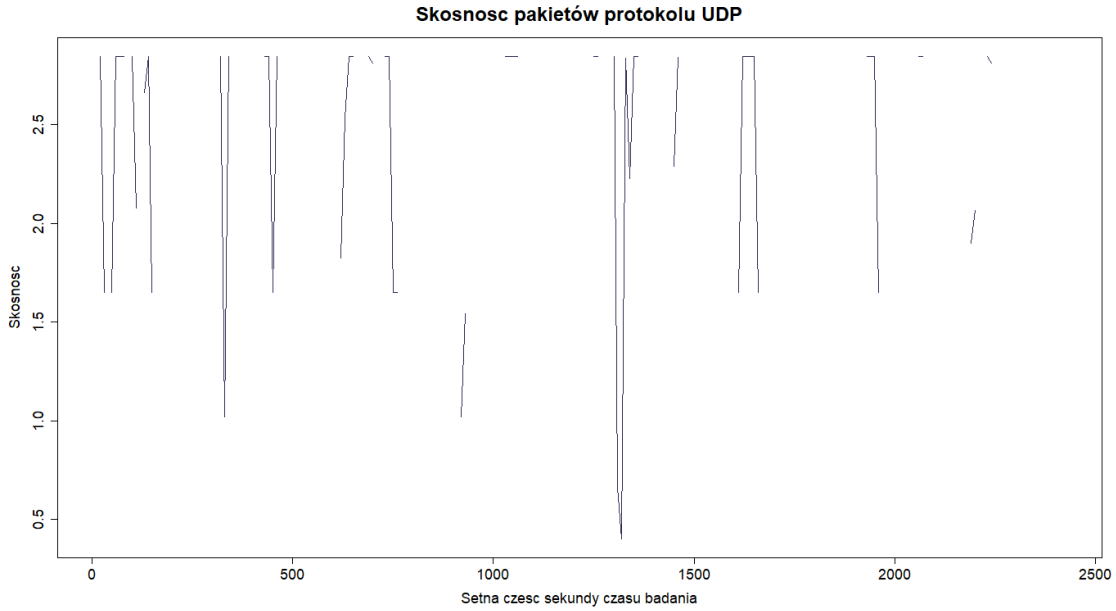
4.6.1. Analiza serii czasowych z wykorzystaniem współczynnika skośności dla pakietów protokołu TCP



Rys. 3.20 Skośność dla protokołów pakietu TCP

Na różnych oknach czasowych rozkład pakietów TCP jest lewoskrętny bądź prawoskrętny, ponieważ wartości oscylują między -0,8 a 2,7.

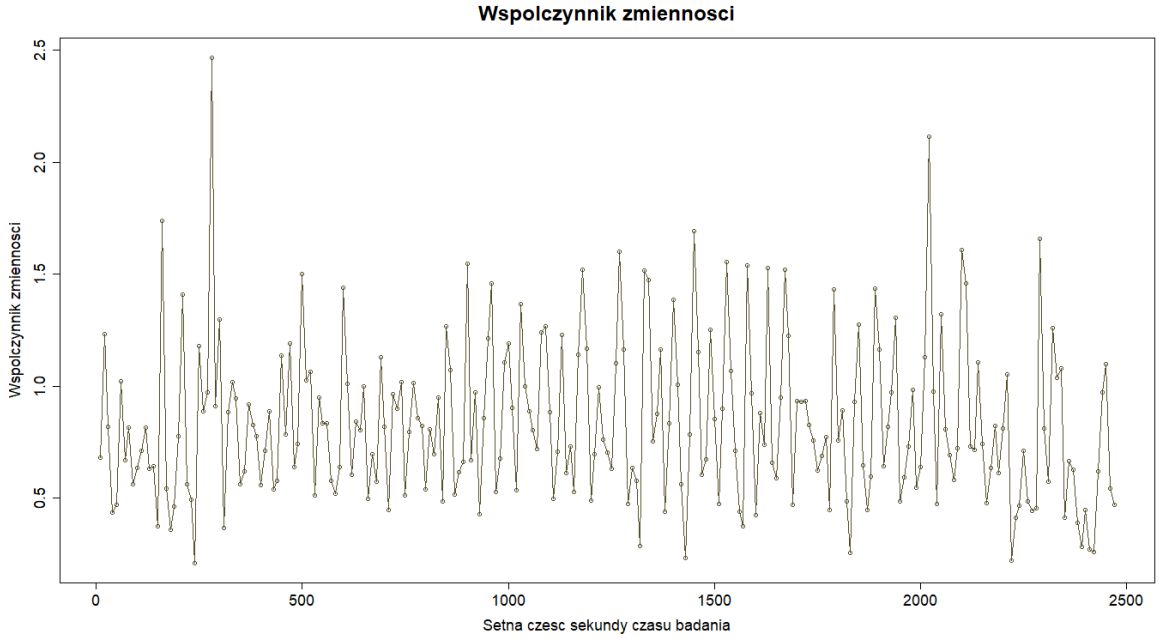
4.6.2. Analiza serii czasowych z wykorzystaniem współczynnika skośności dla pakietów protokołu UDP



Rys. 3.21 Skośność dla protokołów pakietu UDP

Rozkład pakietów UDP na wykorzystywanych oknach czasowych jest prawoskrętny - wartości wahają się między 0,4 a 2,8

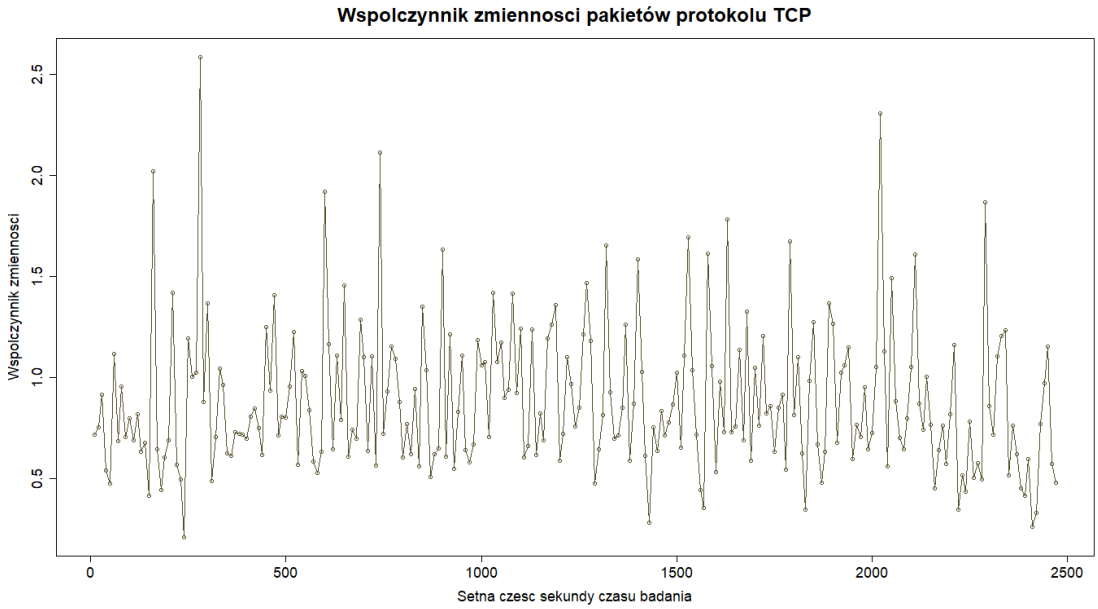
4.7. Analiza serii czasowych z wykorzystaniem współczynnika zmienności dla całego ruchu sieciowego



Rys. 3.22 Współczynnik zmienności dla okien czasowych całego ruchu sieciowego

Dane zostały podzielone na okna czasowe co 1/100 sekundy. Minimalna wartość współczynnika zmienności wyniosła 0.2123284 a maksymalna 2.4658894. Najmniejsze obliczone wartości znajdują się poniżej 25% co oznacza, że zmienność jest niska na odpowiednich dla nich oknach czasowych. Jednakże dla całego ruchu wahanie współczynnika jest zbyt wielkie by ruch określić stabilnym. Jest on w większości przypadków silnie lub bardzo silnie zmienny (większość wartości znajduje się powyżej 50%).

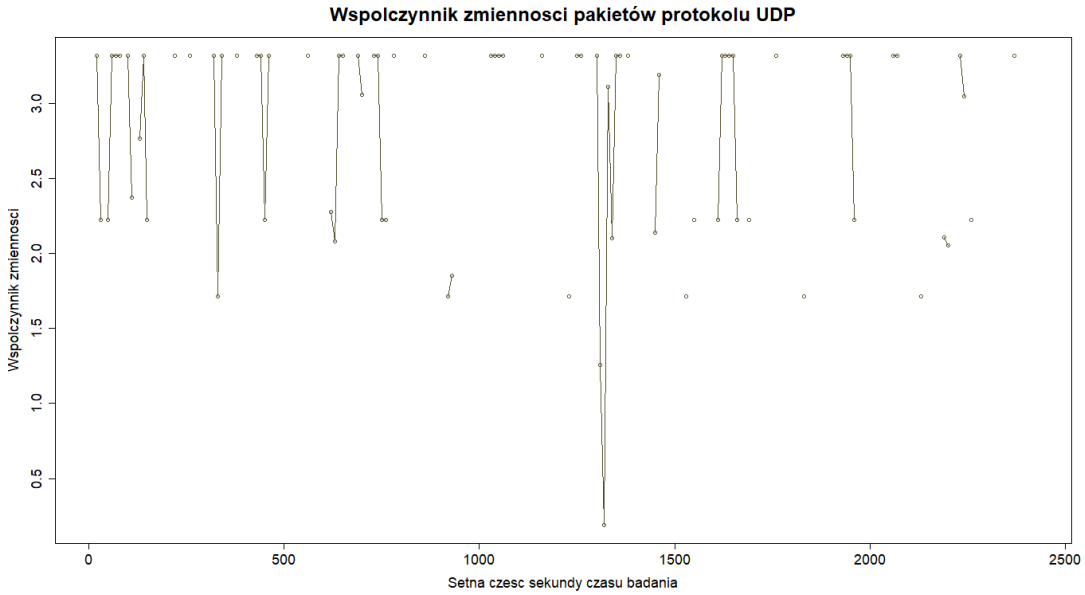
4.7.1. Analiza serii czasowych z wykorzystaniem współczynnika zmienności dla pakietów protokołu TCP



Rys. 3.23 Współczynnik zmienności dla pakietów protokołu TCP

Ruch TCP jest zmienny.

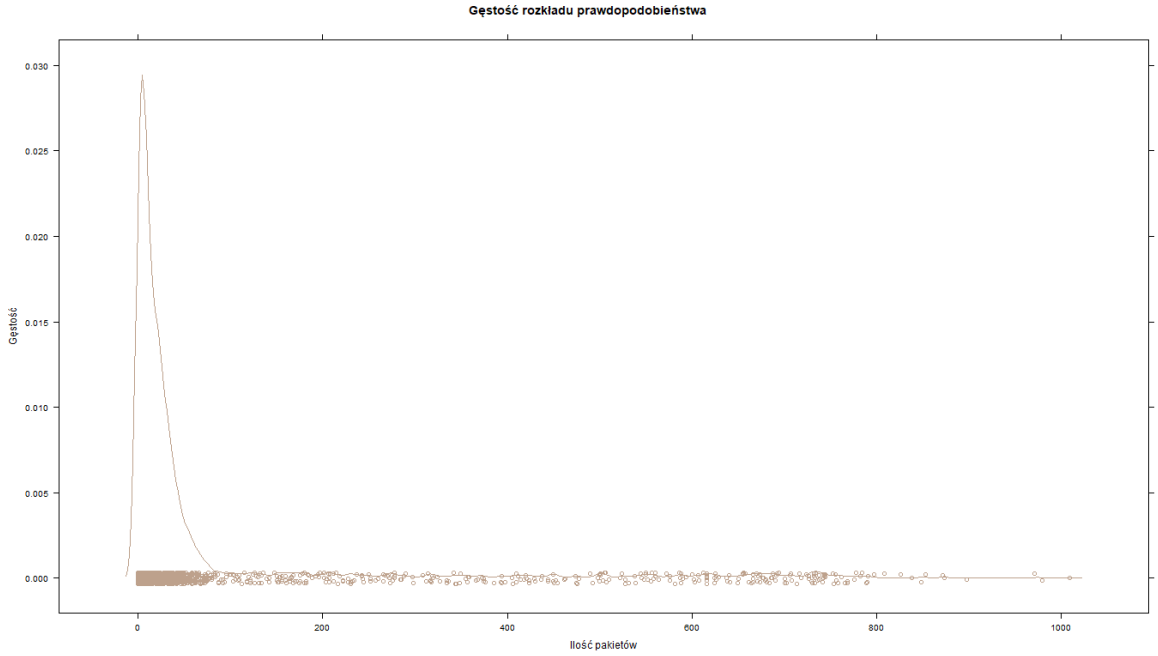
4.7.2. Analiza serii czasowych z wykorzystaniem współczynnika zmienności dla pakietów protokołu UDP



Rys. 3.24 Współczynnik zmienności dla pakietów protokołu TCP

Ruch UDP podobnie jak TCP jest zmienny. Wyjątkiem jest tutaj mały przedział czasowy, w którym protokół był ekstensywnie wykorzystywany - jest mało zmienny.

4.8. Gęstość rozkładu prawdopodobieństwa dla całego ruchu sieciowego



Rys. 3.25 Gęstość rozkładu prawdopodobieństwa pakietów dla całego ruchu sieciowego

Dane zostały podzielone na okna czasowe co 1/100 sekundy. Gęstość rozkładu prawdopodobieństwa pakietów liczona jest na podstawie ilości pakietów na setną część sekundy.

Zdecydowanie największe zagęszczenie pakietów widać w przedziale od 0 do 100 pakietów na setną część sekundy. Pojedyncze wartości pojawiają się nawet w dużej odległości od głównego skupiska, co oznacza, że w niektórych oknach czasowych sieć była wykorzystywana zdecydowanie bardziej niż zwykle.

5. Podsumowanie i wnioski końcowe

Wszystkie zastosowane miary statystyczne wskazują na wysoką zmienność badanego ruchu sieciowego. Jest on niestabilny. Dzięki tej wiedzy można łatwiej wykryć stanowiące zagrożenie anomalie w przyszłej analizie tegoż ruchu. Miałyby one odzwierciedlenie w miarach statystycznych.

Literatura

[1] M. Burdacki, P. Dymora, M. Mazurek, Analiza ruchu w sieci komputerowej w oparciu o modele multifraktalne (2017)

[2] Ł. Bil, Metody nieekstensywnej fizyki statystycznej w badaniu układów złożonych na przykładzie polskiej Giełdy Papierów Wartościowych (2015)

[3] A. Kiłyk, Z. Wilimowska, Wykorzystywanie wykładnika Hursta do prognozowania zmian cen na giełdzie papierów wartościowych (2015)

[4] https://pogotowiestatystyczne.pl/slowniczek/korelacja/

[5] http://zsi.tech.us.edu.pl/~nowak/smad/SMAD_w3.pdf

[6] https://www.naukowiec.org/wiedza/statystyka/analiza-regresji--idea_734.html

[7] http://zsi.tech.us.edu.pl/~nowak/odzw/korelacje.pdf

[8] http://home.agh.edu.pl/~adan/wyklady/rpis3.pdf

[9] http://www.mif.pg.gda.pl/kmd/magda/pdfy/wiacek.pdf

[10] https://www.statystyka-zadania.pl/wspolczynnik-asymetriiskosnosci/