Politechnika Warszawska

WYDZIAŁ MATEMATYKI
I NAUK INFORMACYJNYCH

Praca dyplomowa magisterska

na kierunku Matematyka

w specjalności Matematyka w Ubezpieczeniach i Finansach

Modelowanie ubezpieczeń w obszarze Cyberbezpieczeństwa

Dariusz Wójcik

Numer albumu 284150

promotor

Dr. Anna Cena

WARSZAWA 2021

Streszczenie

Modelowanie Ubezpieczeń w Obszarze Cyberbezpieczeństwa

Cyberbezpieczeństwem określamy procedury ochrony systemów komputerowych (sieci informatycznych, urządzeń i programów), a przede wszystkim – poufnych danych przed zagrożeniami wynikającymi z potencjalnych ataków (np. działań naruszających poufność, integralność, dostępność i autentyczność przetwarzanych danych). Duże przedsiębiorstwa i korporacje, ale też instytucje państwowe w coraz większym stopniu narażone są na cyber zagrożenia. Dlatego też możemy zaobserwować nie tylko coraz większą liczbę regulacji prawnych dotyczących cyberbezpieczeństwa, m. in. w kontekście ochrony danych, lecz także szybki rozwój sektora ubezpieczeń od cyber zagrożeń. Ze względu na dość nietypowe cechy ryzyka w tym przypadku, modelowanie i analiza tego obszaru jest nie tylko ciekawym, ale też wciąż rozwijanym tematem badawczym. Celem mojej pracy był przegląd oraz usystematyzowanie aktualnego stanu wiedzy z zakresu ubezpieczeń w obszarze cyberbezpieczeństwa, ze szczególnym uwzględnieniem zaproponowanych w literaturze technik modelowania.

Słowa kluczowe: cyberbezpieczeństwo, cyber ryzyko, cyber zagrożenie, ubezpieczenia, proces Markowa, kopuły, modelowanie, wycena, symulacja, model ACD, model ARMA, model GARCH, składki, modele bibliometrczne

Abstract

Cybersecurity Insurance Modeling

Cyber security refers to procedures and techniques design to protect computer systems (e.g. computer networks, devices and programs) and primarily confidential personal data from threats of potential attacks (e.g., actions that compromise the confidentiality, integrity, availability, and authenticity of processed data). Large enterprises and corporations, as well as government institutions, are increasingly exposed to cyber threats. Hence, we can observe not only an increasing number of legal regulations concerning cyber security, among others in the context of data protection, but also the rapid development of the cybersecurity insurance sector. However, due to the rather unusual risk characteristics in this case, the modeling and analysis in this domain is not only an interesting but still developing research topic. The purpose of my thesis was to review and systematize the „state of the art" in cybersecurity insurance business, with particular emphasis on modeling techniques proposed in the literature.

Keywords: cyber security, cyber risk, cyber threats, insurances, Markov process, copulas, modeling, pricing, simulation, ACD model, ARMA model, GARCH model, premiums, bibliometric models

Załącznik nr 1 do zarządzenia nr 109 /2021
Rektora PW z dnia 9 listopada 2021 r.

„załącznik nr 5 do zarządzenia nr 42 /2020 Rektora PW

Politechnika Warszawska

Warszawa 14.12.2021 r.
miejscowość i data

Dariusz Wójcik
imię i nazwisko studenta

284150
numer albumu

Matematyka
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.
Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2021 r., poz. 1062) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Wójcik
czytelny podpis studenta ".

Załącznik nr 3 do zarządzenia nr 109 /2021
Rektora PW z dnia 9 listopada 2021 r.

„załącznik nr 9 do zarządzenia nr 42 /2020
Rektora PW

Politechnika Warszawska

Warszawa, 15.12.2021
miejscowość i data

Dariusz Wójcik
imię i nazwisko studenta
284150
numer albumu
MiNI, Matematyka
Wydział i kierunek studiów

Oświadczenie studenta w przedmiocie udzielenia licencji

Politechnice Warszawskiej

Oświadczam, że jako autor/~~współautor~~* pracy dyplomowej pt. Modelowanie ubezpieczeń w obszarze Cyberbezpieczeństwa udzielam/~~nie udzielam~~* Politechnice Warszawskiej nieodpłatnej licencji na niewyłączne, nieograniczone w czasie, umieszczenie pracy dyplomowej w elektronicznych bazach danych oraz udostępnianie pracy dyplomowej w zamkniętym systemie bibliotecznym Politechniki Warszawskiej osobom zainteresowanym.

Licencja na udostępnienie pracy dyplomowej nie obejmuje wyrażenia zgody na wykorzystywanie pracy dyplomowej na żadnym innym polu eksploatacji, w szczególności kopiowania pracy dyplomowej w całości lub w części, utrwalania w innej formie czy zwielokrotniania.

Wójcik
czytelny podpis studenta

* niepotrzebne skreślić ".

Spis treści

Wstęp

Początki ubezpieczeń w obszarze cyberbezpieczeństwa (jak pisze M. Camillo [2]) przypadają na koniec lat 70-tych ubiegłego wieku. W latach 80-tych powstawały pierwsze ubezpieczenia dla najbogatszych firm finansowych. Pozostawały one jednak w rynkowej niszy. Zainteresowanie cyber ubezpieczeniem wzrosło dopiero po atakach terrorystycznych 11 września 2001r. Zauważono wtedy, że świat wirtualny nie przypomina świata rzeczywistego. Prawdziwy rozwój obszaru cyberbezpieczeństwa rozpoczął się tak na prawdę dopiero w latach 2002-2003. Wtedy to w Kalifornii w USA wprowadzono pierwsze prawo, które nakazywało pisemne informowanie osób ubezpieczonych w przypadku ewentualnego wycieku ich danych. W Europie podobne prawa wprowadzono dopiero w roku 2009. Dodatkowo finansowe rady nadzorcze zaczęły nakładać kary na firmy za różne naruszenia i wycieki danych osobowych (np. w 2018r. wprowadzono rozporządzenie o ochronie danych osobowych RODO[1]).

Przejdę teraz do sformułowania problemu związanego z obszarem cyberbezpieczeństwa. Cyber ryzyko to szeroko stosowany termin z co najmniej kilkoma definicjami. Pojęcie to może być różnie definiowane w zależności od kraju czy firmy ubezpieczeniowej. Posiadanie jasnego, wszechstronnego i wspólnego zestawu definicji cyber ryzyka umożliwiłoby bardziej zorganizowany i ukierunkowany dialog między branżą, organami nadzorczymi i decydentami w celu ułatwienia opracowywania solidnych rozwiązań w zakresie cyberbezpieczeństwa. Dlatego też największe grupy ubezpieczeniowe co roku proszone są przez różne urzędy nadzoru o przedstawienie definicji cyber ryzyka. Na podstawie odpowiedzi respondentów podanych w jednej z ankiet z 2019r. (przeprowadzonej przez European Insurance and Occupational Pensions Authority (EIOPA) [5] na grupie 41 największych ubezpieczycieli z 12 najbardziej rozwiniętych krajów Europy) widać, że połowa z nich jako cyber ryzyko wskazywała definicję podaną przez Financial Stability Board (FSB) [21]. Kilka grup w ogóle nie ma skonkretyzowanej definicji cyber ryzyka, choć zgodnie z ich deklaracją pracują nad jej ustaleniem. Ponadto definicje wielu grup znacznie się od siebie

[1]RODO, czyli ogólne rozporządzenie o ochronie danych osobowych, to rozporządzenie unijne, które zawiera przepisy dot. ochrony osób fizycznych w związku z przetwarzaniem danych osobowych, a także przepisy dot. wolnego przepływu danych osobowych. Ma ono pozwolić obywatelom UE na lepszą kontrolę nad ich danymi oraz stanowi unowocześnienie przepisów umożliwiających firmom ograniczanie biurokracji.

różniły. Duża część firm wybrała definicję sformułowaną przez International Association of Insurance Supervisors IAIS [7], która zawęża to pojęcie i ma najwięcej cech wspólnych z definicjami różnych firm ubezpieczeniowych. Widać więc, że ten sektor ubezpieczeń nie jest jeszcze w pełni dostosowany jeśli chodzi o koncepcyjne definiowanie cyber zagrożeń.

W swojej pracy będę korzystał z definicji podanej przez IAIS, ponieważ stanowi przecięcie większości definicji podawanych przez firmy ubezpieczeniowe.
Według IAIS [7] *cyber ryzyko* to wszelkie zagrożenia wynikające z korzystania z danych elektronicznych, w tym narzędzi technologicznych takich jak Internet i sieci telekomunikacyjne oraz z przesyłania tych danych. Obejmuje też szkody fizyczne, które mogą być spowodowane incydentami związanymi z cyberbezpieczeństwem (np. uszkodzeniem serwera), oszustwami popełnionymi w wyniku niewłaściwego korzystania z danych, wszelką odpowiedzialnością wynikającą z przechowywania danych oraz dostępnością, integralnością i poufnością informacji elektronicznych – zarówno to w odniesieniu do osób fizycznych, do firm jak i rządów.

Analogicznie do pojęcia cyber ryzyka, istnieją różne definicje cyberbezpieczeństwa w zależności od kraju lub grupy ubezpieczeniowej. Tu również będę wykorzystywał tę, którą podał IAIS [7]. Według IAIS cyberbezpieczeństwo odnosi się do:

- strategii, wytycznych i standardów obejmujących ograniczanie zagrożeń;
- zmniejszania podatności na zagrożenia;
- zaangażowania międzynarodowego;
- reagowania na incydenty;
- wytrzymałości i działań naprawczych systemów;
- zasad dotyczących bezpieczeństwa działalności ubezpieczyciela.

Na podstawie ankiet przeprowadzanych w SOA (Society of Actuaries [6]) dotyczących nowych wschodzących ryzyk już w 2011r. ryzyko cyber zagrożeń, według respondentów, było trzecim największym wschodzącym ryzykiem (38% ankietowanych wskazało właśnie to ryzyko), w latach 2012 - 2013 było drugim (odpowiednio 40% i 47%), a w latach 2014 - 2018 wskazywano go już jako najgroźniejsze (56% w 2018r). Od roku 2019 (zob. [19] i [20]) cyber ryzyko znajduje się wśród 4 uznawanych przez respondentów za najgroźniejsze. Cyber ryzyko zostało przysłonięte przez ryzyko pandemii wywołanej wirusem SARS-CoV-2. Oczywistym jest więc, że firmy poszukują przed nim jakiegoś zabezpieczenia. Oszacowano, że roczna składka ubezpieczeniowa z tytułu ryzyka cyberbezpieczeństwa wyniosła 3.25 mld USD w 2016r. w porównaniu do 2,75 mld USD w 2015r. (zob. [1]).

Ubezpieczenie od *cyber ryzyka* ma na celu ograniczenie strat spowodowanych różnymi incydentami cybernetycznymi np. naruszeniami ochrony danych (przez ujawnienie informacji poufnych), przerwami w działalności firmy lub różnego rodzaju uszkodzeniami sieci. Mając przegląd najpowszechniejszych cyber zagrożeń dla sektora ubezpieczeń, możemy pomóc ubezpieczycielom w określaniu działań i środków zapobiegawczych w celu zminimalizowania, kontrolowania i monitorowania ich skutków.

Pod względem częstotliwości występowania i wysokości kosztów pokrycia szkód jako najgroźniejsze z incydentów cybernetycznych (ponownie na podstawie ankiety dla EIOPA - European Insurance and Occupational Pensions Authority [5], z 2019r.) możemy wyróżnić kolejno:

- eksfiltrację (wykradzenie) danych, czyli utratę poufnych danych przez firmę na rzecz osób nieuprawnionych, którzy naruszają prywatność klientów, pracowników lub kontrahentów;
- naruszenie służbowej poczty e-mail lub tzw. "oszustwo na CEO". Podczas tych ataków cyberprzestępca podszywa się pod CEO firmy i wysyła maile do pracowników, w których pyta o wrażliwe dane;
- infekcje złośliwym oprogramowaniem (z ang. *malicious software*, inaczej malware). Są to najczęściej:
  - wirusy (kopiująca się aplikacja, która modyfikuje pliki systemowe i stopniowo pogarsza działanie systemu, aż do jego całkowitego unieruchomienia);
  - trojany (pozwalają na zdalny dostęp do komputera);
  - *ransomware* (oprogramowanie szantażujące, które blokuje pliki i żąda okupu w zamian za przywrócenie do nich dostępu);
  - *adware* (oprogramowanie, które natrętnie wyświetla uciążliwe reklamy);
  - *spyware* (oprogramowanie szpiegujące, które śledzi aktywność użytkownika);
  - *keylogger* (oprogramowanie odczytujące naciskane klawisze i umożliwiające kradzież haseł);
- ataki DDoS. To rodzaj ataku, w którym wiele zainfekowanych systemów będących zainfekowanych trojanem jest wykorzystywanych do ataku na pojedynczy system, co powoduje przeciążenie i daje hakerom swego rodzaju furtkę do wykradzenia danych;
- "*zero – day exploit*". To cyberatak mający miejsce tego samego dnia, w którym wykryto luki w zabezpieczeniach, jeszcze przed jej załataniem przez twórcę oprogramowania;

- kradzież transakcji finansowych, czyli nieautoryzowany transfer środków pieniężnych za pośrednictwem zaufanych sieci w celu wyprowadzenia pieniędzy z konta i niemożności ich odzyskania;

- phishing. Ten cyberatak zazwyczaj kieruje użytkownika do fałszywej strony internetowej, podszywającej się pod prawdziwą organizację, na której proszony jest o podanie danych osobowych, takich jak hasło, numer karty kredytowej lub numery kont bankowych.

Cyber ryzyko znacznie różni się od ryzyk tradycyjnych. Tym, co je odróżnia jest to, że zasoby technologii informacyjno-komunikacyjnych (z ang. *Information and Communication Technologies* w skrócie ICT) są ze sobą połączone w sieci, więc analiza ryzyka i związanych z nim potencjalnych strat być może będzie musiała uwzględniać topologię sieci. Ponadto, jeśli jakieś zasoby zostaną skradzione z jednego źródła, np. komputera, to ten komputer może stać się zagrożeniem dla kolejnych źródeł. Również rozwijająca się w szybkim tempie technologia może zwiększyć narażenie gospodarstw domowych na zagrożenia cybernetyczne, np. *Internet of Things* (IoT). Ideą IoT jest połączenie urządzeń codziennego użytku za pośrednictwem Internetu, które umożliwia użytkownikowi zdalne sterowanie różnymi urządzeniami elektronicznymi. Podczas gdy IoT może potencjalnie ułatwić zapobieganie stratom, np. poprzez wcześniejsze wykrywanie pewnych zagrożeń (np. oznak pożaru), to duża ilość danych i ich wzajemne powiązania tworzą potężny ekosystem, który może stać się atrakcyjny dla hakerów zainteresowanych dostępem do wrażliwych i poufnych informacji. Ponadto cyberataki mogą również materializować się jako szkody fizyczne, spowodowane na przykład przez urządzenia zaprogramowane zdalnie (celowo lub nie) w celu spowodowania awarii. Tę logikę można rozszerzyć na infrastrukturę, życie i środki transportu.

Wycena zwykłego produktu ubezpieczeniowego często opiera się na tablicach aktuarialnych skonstruowanych na podstawie danych historycznych. Dla omawianego produktu nie ma dostępnych jeszcze zbyt wielu danych historycznych, ponieważ jest to stosunkowo nowy temat. Dodatkowo pozyskiwanie takich danych jest utrudnione, gdyż firmy, które padły ofiarą ataku cybernetycznego niechętnie upubliczniają te informacje. Zwyczajnie boją się o utratę reputacji lub o spadki cen akcji. Dlatego też wycena odpowiedniej składki jest dużym wyzwaniem. Na oferowanie produktu ubezpieczeniowego od cyber ryzyka decyduje się coraz więcej ubezpieczycieli, jednak mają oni skłonność do zawyżania składek, a ewentualne pokrycia są stosunkowo ograniczone. Dodatkowo z wyżej wymienionych ankiet możemy się dowiedzieć, że ubezpieczyciele nie mają jednolitego sposobu sprzedaży polis cyberbezpieczeństwa. Najpopularniejsze sposoby sprzedaży polis to za pośrednictwem agentów, brokerów i sprzedaż bezpośrednia. Co ciekawe, nikt nie wspomniał o wykorzystaniu Internetu jako sposobu sprzedaży. Może to odzwierciedlać

stosunkowo skomplikowany i niestandardowy charakter produktów cybernetycznych.
Wymagana więc może być zwiększona pomoc reasekuratorów. Działania mające na celu gromadzenie danych, jak i opracowanie nowych modeli pomogą rynkowi reasekuracji, który chce zwiększyć swoje wsparcie dla rynku cyberbezpieczeństwa. Im dokładniej ubezpieczyciele będą w stanie mierzyć i monitorować ryzyko, tym większą ochronę reasekuracyjną będą mogli uzyskać.

W przypadku braku tradycyjnych lub alternatywnych mechanizmów reasekuracji, koszty poważniejszych incydentów cybernetycznych będą ponoszone przez rządy, przedsiębiorstwa oraz ich akcjonariuszy. Jednak coraz bardziej prawdopodobne jest to, że akcjonariusze i organy nadzorujące wymagać będą zakupu kompleksowego ubezpieczenia cybernetycznego po to, by zlikwidować luki między stratami gospodarczymi a ubezpieczonymi stratami cybernetycznymi.

Z uwagi na to, że problem cyber ubezpieczeń jest nowością, staje się on problemem nie tylko teoretycznym, ale też praktycznym. Teoria modelowania cyber ryzyka i problem wyceny odpowiedniej składki musi bowiem nadążać za praktyką.

Celem mojej pracy jest usystematyzowanie przeglądu zaproponowanych w literaturze technik modelowania, wyceny ryzyka cyber zagrożeń i krytyczna analiza zaproponowanych rozwiązań, ze szczególnym uwzględnieniem wybranych modeli.
Układ pracy: w 1. rozdziale zamieściłem opis ubezpieczeń cybernetycznych na rynku polskim i amerykańskim, a także krótkie podsumowanie aktualnego stanu wiedzy z tego obszaru. W 2. rozdziale opisałem w skrócie wykorzystane przeze mnie metody modelowania. W rozdziale trzecim przedstawiłem wykorzystane dane rzeczywiste, wyróżniłem i opisałem najważniejsze definicje, twierdzenia oraz testy statystyczne, szczegółowo opisałem wykorzystane przeze mnie metody modelowania, oraz przedstawiłem uzyskane wyniki analizy i wyciągnąłem z nich wnioski. W 4. rozdziale krótko opisałem teorię modeli bibliometrycznych, a także przedstawiłem i opisałem otrzymane wyniki analizy, uzyskane dzięki zastosowaniu wybranych modeli. Ponadto w dodatku A. umieściłem tabelki z wynikami dot. analiz z poprzednich rozdziałów. Rozdział 5 zawiera podsumowanie pracy i wskazanie dalszych kierunków badań.

1. Przegląd ubezpieczeń

Rozdział ten rozpocznę od przypomnienia dwóch podstawowych definicji, a mianowicie cyber ryzyka oraz ubezpieczenia od cyber ryzyka. Według IAIS [7] *cyber ryzyko* to wszelkie zagrożenia wynikające z korzystania z danych elektronicznych, w tym narzędzi technologicznych takich jak Internet i sieci telekomunikacyjne oraz z przesyłania tych danych. Obejmuje też szkody fizyczne, które mogą być spowodowane incydentami związanymi z cyberbezpieczeństwem (np. uszkodzeniem serwera), oszustwami popełnionymi w wyniku niewłaściwego korzystania z danych, wszelką odpowiedzialnością wynikającą z przechowywania danych oraz dostępnością, integralnością i poufnością informacji elektronicznych – zarówno to w odniesieniu do osób fizycznych, do firm jak i rządów. Natomiast ubezpieczenie od *cyber ryzyka* ma na celu ograniczenie strat spowodowanych różnymi incydentami cybernetycznymi np. naruszeniami ochrony danych (przez ujawnienie informacji poufnych), przerwami w działalności firmy lub różnego rodzaju uszkodzeniami sieci.

Ubezpieczenie dotyczące ryzyka cybernetycznego, pomimo bycia dużą niszą, oferowane jest już od kilku do kilkunastu lat, w zależności od firmy, na świecie, a od niedawna także w Polsce. Na świecie istnieje wiele firm ubezpieczeniowych oferujących różne rodzaje produktów cyber ubezpieczeniowych. W Polsce natomiast nie jest ono jeszcze zbyt popularne. Na razie oferuje je tylko kilka firm ubezpieczeniowych. Największą z nich jest PZU S.A. W internecie znaleźć można znacznie więcej ofert niż tylko te wymienione w poniższym podrozdziale, jednakże większość z nich mają tylko jedną placówkę w Polsce, a niektóre nie mają żadnej - ubezpieczenie można wykupić tylko online. Można się zatem zastanawiać nad tym, czy zakup ubezpieczenia w nieznanej firmie jest bezpieczny.

Według Polskiej Izby Ubezpieczeń (PIU) ubezpieczenia dzielą się na dwie grupy: ubezpieczenia na życie oraz ubezpieczenia majątkowe i niektóre ubezpieczenia osobowe. Produkt ubezpieczeniowy dotyczący cyber ryzyka jest jednym z rodzajów ubezpieczeń majątkowych. W Polsce, zgodnie z art. 821 kodeksu cywilnego, przedmiotem ubezpieczenia majątkowego może być każdy interes majątkowy, który nie jest sprzeczny z prawem i daje się ocenić w wartości pieniężnej. Należą do nich np. ubezpieczenie nieruchomości, ubezpieczenia komunikacyjne – takie jak OC czy AC, ubezpieczenie turystyczne i wspomniane już wyżej cyberubezpieczenie. Oferowane jest ono głównie dla firm średnich i dużych, jednak znaleźć można też oferty dla małych przedsiębiorstw.

W następnym podrozdziale postaram się wymienić największe i najbardziej popularne firmy ubezpieczeniowe w Polsce, a w kolejnym - kilka firm oferujących omawiany produkt na rynku amerykańskim. Spróbuję przedstawić w skrócie ich oferty, wymienić rodzaje cyber ubezpieczeń (jeśli oferowane są różne rodzaje) oraz wypunktować najważniejsze cechy danej firmy na podstawie informacji dostępnych na stronach internetowych wybranych firm.

1.1. Polski rynek ubezpieczeń

1.1.1. PZU S.A.

PZU S.A. oferuje produkt ubezpieczeniowy, którego pełna nazwa to "Ubezpieczenie od ryzyk cybernetycznych i związanych z RODO". Ubezpieczenie to chroni np. przed skutkami ataków hakerskich i konsekwencjami naruszenia przepisów dotyczących prywatności, w tym RODO. Grupa PZU wśród najpoważniejszych zagrożeń cybernetycznych wymienia:

- Złośliwe oprogramowanie (*malware*): np. wirusy komputerowe lub *ransomware*, które mogą spowodować niezaplanowane przerwy w działaniu komputerów i infrastruktury IT;
- Phishing, mogący doprowadzić do kradzieży danych;
- Atak blokujący dostęp (atak *DoS* i *DDoS*);
- Naruszenie bezpieczeństwa danych, np. kradzież lub wyciek danych, zgubienie dokumentów, spowodowane nieuczciwością lub zwykłą nieuwagą pracownika.

Jak zatem widać, produkt ten chroni firmę przed skutkami ataków cybernetycznych oraz przed zobowiązaniami wynikającymi z naruszenia przepisów dotyczących prywatności, np. rozporządzenia o ochronie danych (RODO). Zapewnia on pokrycie:

- kosztów roszczeń (np. naruszenia prywatności, w tym danych osobowych, naruszenia praw autorskich, zniesławienia, piractwa, przywłaszczenia lub kradzieży koncepcji, albo przekazania wirusa dalej);
- kosztów związanych z naprawą szkód (np. koszty obsługi prawnej, odzyskania danych, naprawy wizerunku, ochrony dobrego imienia, monitorowania transakcji lub związane z cybernetycznym wymuszeniem);
- kosztów kar i oceny PCI (*Payment Card Industry* - czyli norma zapewniająca wysoki i spójny poziom bezpieczeństwa wszędzie tam, gdzie przetwarzane są dane posiadaczy kart

płatniczych), które obowiązują wszystkie firmy akceptujące płatności kartami kredytowymi i debetowymi;

- kosztów utraty zysków przedsiębiorstwa i wydatków niezbędnych do podtrzymania działalności firmy (np. koszty i wydatki na uniknięcie lub zmniejszenie rozmiarów skutków awarii, koszty związane z pogorszeniem bezpieczeństwa sieci informatycznej, zakłóceniem działalności).

PZU S.A. jako zalety swojego produktu podaje:

- finansowe zabezpieczenie w przypadku naruszenia danych osobowych;
- uzupełnienie zabezpieczeń systemów (wiadomo bowiem, że nawet najbardziej zaawansowane zabezpieczenia nie zapewniają bezpieczeństwa w 100%);
- uzupełnienie standardowej ochrony ubezpieczeniowej ubezpieczającej się firmy;
- wsparcie międzynarodowych partnerów i ich lokalną pomoc;
- ochronę dla każdej firmy i branży (niezależnie od jej wielkości).

1.1.2. Findia Insurance

Findia Insurance, z siedzibą w Brukseli (Belgia), oferuje produkt ubezpieczeniowy o krótkiej nazwie "Ubezpieczenie cyber". Ubezpieczenie to zapewnienia ochronę ubezpieczonemu, który narażony jest na szkody spowodowane użytkowaniem komputerów, systemów informatycznych, smartfonów, tabletów oraz innych urządzeń. Ubezpieczone szkody dotyczą strat wyrządzonych bezpośrednio ubezpieczonemu (np. przerwa w działaniu systemów lub utrata danych), jak również strat wyrządzonych osobom trzecim (np. wyciek danych osobowych, zainfekowanie systemów lub komputerów podmiotów trzecich).
Zadaniem ubezpieczenia jest:

- zapewnienie ubezpieczonemu pomocy ekspertów, którzy mają za zadanie przywrócić system i urządzenia do stanu przed awarią (atakiem);
- sfinansowanie pomocy prawnej wtedy, gdy roszczenia wysuną osoby trzecie (np. osoby fizyczne, firmy, organy państwowe).

Zakres ubezpieczenia oferowany przez firmę Findia Insurance składa się z dwóch sekcji.

1. Ubezpieczenie przed szkodami w cyberprzestrzeni, która zawiera ochronę przed utratą i uszkodzeniem danych i sieci, przerwy w działalności biznesowej, ubezpieczenie od kradzieży, wymuszeń, ataku hakerskiego na linie telefoniczne, kosztów notyfikacji, skutków oszustwa podszywania się pod inną osobę oraz kosztów ochrony reputacji.

2. Ubezpieczenie cyber od roszczeń osób trzecich tzn. odpowiedzialność medialna (np. zniesławienia, zdyskredytowania produktu lub nieumyślnego naruszenia własności intelektualnej), naruszenie prywatności i utrata dokumentów, odpowiedzialność z tytułu naruszenia poufności informacji, odpowiedzialność z tytułu bezpieczeństwa sieci, koszty ograniczenia roszczenia, postępowania regulacyjne i kary oraz odpowiedzialność z tytułu naruszenia bezpieczeństwa płatności.

Dodatkowo firma oferuje trzy pakiety ubezpieczeniowe tj. Findia Cyber SMART, Findia Cyber oraz Findia Cyber Enterprise. Na stronie ubezpieczyciela możemy zobaczyć infografikę przedstawiającą porównanie tych 3 pakietów. Dowiemy się z niej, że szybkość uzyskania polisy jest najwyższa w pierwszym pakiecie, zakres ochrony jest taki sam w każdym z nich, premia za bezpieczeństwo cyber (bezpieczne działanie firmy) jest najwyższa w 2. i 3. pakiecie i właśnie one otrzymały rekomendację Findii. Jeśli chodzi o sposób oceny ryzyka, to w 1. pakiecie wystarczy oświadczenie, w 2. wymagane jest wypełnienie formularza w systemie, a w 3. potrzebna jest już indywidualna ocena ryzyka. Wsparcie brokera dostępne jest tylko w 2. i 3. pakiecie. Pakiet 1. jest dla klientów z obrotem rocznym do 50 mln zł, a 2. i 3. powyżej 50 mln zł.

1.1.3. Colonnade Insurance S.A.

Firma Colonnade Insurance S.A., zarejestrowana w Luksemburgu, oferuje produkt ubezpieczeniowy o nazwie "Ubezpieczenie ryzyk cybernetycznych". Oferowane tu ubezpieczenie chroni przedsiębiorstwo, ale też ogranicza szkody w związku z roszczeniami odszkodowawczymi spowodowanymi utratą lub ujawnieniem danych. Zatem ubezpieczenie ryzyk cybernetycznych ma za zadanie ograniczyć dotkliwe skutki naruszenia bezpieczeństwa danych.

Ubezpieczenie pokrywa też koszty związane z obroną prawną, ujawnieniem informacji handlowej, odszkodowaniem w przypadku utraty danych osobowych, korzystaniem z usług różnych podwykonawców i z bezpieczeństwem sieci; kwoty odpowiadające wysokości kar administracyjnych za naruszenie danych; honoraria dla prawników specjalistów od informatyki śledczej oraz konsultantów PR; wynagrodzenia związane z zawiadomieniem osoby, której dotyczą dane oraz koszty związane z odzyskaniem danych elektronicznych.

Zatem swoją ochroną firma obejmuje:

- odpowiedzialność za dane osobowe i informacje handlowe;

- zarządzanie kryzysowe (pokrycie kosztów zarządzania kryzysowego związanego z atakiem cybernetycznym);
- koszty śledztwa informatycznego i koszty powiadomień osób;
- postępowania administracyjne (koszty porad prawnych związanych z postępowaniem);
- dane elektroniczne (koszty odzyskania lub odtworzenia danych elektronicznych).

Do zakresu ochrony ubezpieczeniowej mogą też zostać dodane zakłócenia w działaniu sieci oraz działalność multimedialna i próba szantażu.

Colonnade Insurance proponuje jeszcze dodatkowy "Wariant RODO". Obejmuje on

- ochronę ubezpieczeniową "CYBER GUARD" w dwóch wariantach: szerszy obejmuje wszystkie ryzyka cybernetyczne, węższy natomiast obejmuje tylko ryzyko związane z RODO;
- ubezpieczenie "CYBER GUARD RODO": obejmuje ono koszty zawiadomienia osób, których dane dotyczą, koszty postępowań i kar administracyjnych.

Colonnade Insurance oszacowało, że koszty związane z wyciekiem danych w firmie średniej wielkości mogą sięgać nawet 500 000 zł.
Zarządzanie ryzykiem cybernetycznym zależy od wielkości przedsiębiorstwa. Większe firmy chętniej wybierają ubezpieczenie CYBER GUARD w pełnym zakresie, mniejsze natomiast - CYBER GUARD RODO.

1.2. Amerykański rynek ubezpieczeń (USA)

Rynek cyber ubezpieczeń w USA jest oczywiście znacznie bardziej rozwinięty niż rynek w Polsce, zatem klienci mają dużo większy wybór ofert cyber ubezpieczeń. W Internecie znaleźć można wiele stron opisujących w skrócie oferty poszczególnych ubezpieczycieli, a także ich rankingi (patrz np. [17] lub [18]). W tym podrozdziale przedstawię kilka z wyżej sklasyfikowanych firm ubezpieczeniowych i opiszę ich oferty, podobnie jak na rynku polskim.

1.2.1. AXA XL Insurance

Firma AXA XL Insurance (informacje pochodzą z oferty udostępnionej na stronie firmy [14]) oferuje dwa produkty ubezpieczeniowe cyber. Pierwszy z nich "Cyber Insurance: International Coverage" przeznaczony na rynki Europejskie, Ameryki Południowej, Azji oraz Australii oraz

drugi – "Cyber Insurance for North America - CyberRiskConnect" oferowany firmom znajdującym się na terenie USA i Kanady. W tym poddziale zajmę się tylko rynkiem USA, zatem opiszę tylko drugi z wymienionych produktów ubezpieczeniowych.

Polisa dla produktu "Cyber Insurance for North America - CyberRiskConnect" obejmuje rozszerzony zakres ochrony przed nowo pojawiającymi się zagrożeniami cybernetycznymi. Zapewnia szereg usług, które pomagają zapobiegać cyberatakom zanim hakerzy zdążą namierzyć firmę. AXA XL współpracuje z wiodącymi usługodawcami reagowania na wycieki danych, aby pomóc radzić sobie we wrażliwych sytuacjach. Dodatkowo firma chwali się, że już od 20 lat zajmuje się incydentami naruszenia danych i zabezpiecza cyber narażenia klientów.

Polisa "CyberRiskConnect" zapewnia rozszerzony zakres i warunki ochrony przed pojawiającymi się nowymi zagrożeniami technologicznymi, utratą danych oraz prywatnością. Dokładniej, firma oferuje kompleksowe ubezpieczenie związane z: odzyskiwaniem danych; odpowiedzialnością za prywatność i bezpieczeństwo; przerwą w działaniu systemu i dodatkowymi kosztami z tym związanymi; różnego rodzaju wymuszeniami i oprogramowaniem wymuszającym okup; reagowaniem na naruszenia danych; zarządzaniem kryzysowym oraz kosztami ochrony prywatności i pokryciem ewentualnych grzywien. Istnieje także możliwość rozszerzenia ubezpieczenia o pokrycie kosztów związanych z grzywnami i karami PCI (Payment Card Industry), awarią systemu, przerwą w biznesie, całkowitą niezdolnością urządzeń do działania oraz utratą reputacji.

Jako branże docelowe firma wskazuje kolejno: branżę cyber (np. instytucje finansowe, służba zdrowia, uczelnie wyższe lub przedsiębiorstwa energetyczne), firmy technologiczne (np. deweloperzy oprogramowania i producenci sprzętu komputerowego, dostawcy usług w chmurze lub konsultanci IT), firmy telekomunikacyjne (przewodowe i bezprzewodowe, VOIP lub dostawcy telewizji kablowej i satelitarnej) oraz firmy internetowe (firmy zajmujące się mediami społecznościowymi, usługi hostingowe lub projektanci stron internetowych).

Na stronie ubezpieczyciela możemy także znaleźć pełną listę partnerów pomagających reagować na naruszenia ochrony danych. Podzieleni oni są na cztery grupy: informatyka śledcza, PR, radcy prawni oraz firmy zajmujące się monitorowaniem i powiadamianiem o naruszeniu danych.

1.2.2. Zurich Insurance Group Ltd.

Kolejną firmą ubezpieczeniową działającą na rynku ubezpieczeniowym w USA, którą postaram się w skrócie opisać jest szwajcarska firma Zurich Insurance Group Ltd. Oferuje ona produkt o nazwie "Zurich Cyber Solution: Security and Privacy Liability Policy", którego głównym celem jest ochrona istotnych danych i zasobów cyfrowych. Według niej poważny i potencjalnie szkodliwy atak cybernetyczny jest bardziej kwestią "kiedy", a nie "czy" w przypadku większo-

ści organizacji. Z tego powodu proponuje kompleksowe ubezpieczenie pomagające chronić przed cyber zagrożeniami.

Dokładniej, ubezpieczyciel pokrywa: koszty naruszenia prywatności, stratę dochodów związaną z działalnością gospodarczą, koszty wymiany zasobów cyfrowych, płatności za wymuszenia cybernetyczne, koszty związane z awarią systemu, koszty nagłych awarii.
Ponadto ubezpieczyciel współpracuje też z innymi firmami pomagającymi reagować na różnego rodzaju naruszenia. Firmy współpracujące z Zurich Insurance Group Ltd pokrywają:

- koszty związane z bezpieczeństwem i prywatnością;
- koszty postępowania sądowego oraz grzywny i kary cywilne związane z PCI.

Według Zurich Insurance Group Ltd. efektywne bezpieczeństwo cybernetyczne wymaga systemu zarządzania bezpieczeństwem zbudowanego na trzech filarach: ludzie, procesy i technologia. Ich planem jest wsparcie organizacji w dostosowaniu się do tych trzech filarów i pomoc w opracowaniu i utrzymaniu skutecznego programu bezpieczeństwa cybernetycznego.
W pierwszym filarze, "ludzie", chodzi o szkolenia dotyczące świadomości użytkowników (np. czym jest phishing, jak tworzyć silne hasła lub jakich maili lepiej nie otwierać), edukację zarządu i kadry kierowniczej oraz przekazywanie wytycznych dotyczących bezpieczeństwa. Kolejnym filarem są "procesy", czyli strategia bezpieczeństwa cybernetycznego; opracowywanie zasad i procedur zarządzania zasobami, słabymi punktami i poprawkami; ocena ryzyka; zarządzanie dostawcami oraz reagowanie na incydenty i powrót do stanu sprzed awarii. Ostatni filar, "technologia", to zalecenia dotyczące szeregu specjalistycznych rozwiązań technologicznych z wiodącymi zewnętrznymi dostawcami zabezpieczeń i konsultantami.

W przypadku ewentualnego ataku firma:

- zlokalizuje i "usunie" źródło ataku, awarii lub naruszenia danych;
- pomoże chronić firmę przed dalszymi atakami;
- oceni straty finansowe;
- pomoże chronić markę i reputację zatrudniając konsultantów PR;
- zapewni wsparcie prawne.

1.2.3. The Travelers Indemnity Company

Kolejna firma to The Travelers Indemnity Company założona w 1864r. przez Jamesa G. Battersona. Podobnie jak Zurich Insurance, Travelers uważa, że pytanie czy jakaś firma doświadczy ataku cybernetycznego nie ma sensu. Sensowne staje się pytanie, kiedy go doświadczy.

Firma proponuje produkt ubezpieczeniowy o nazwie "Cyber Liability Insurance". Oferowane tu cyber ubezpieczenie od odpowiedzialności cywilnej to polisa ubezpieczeniowa, która zapewnia firmom kombinację opcji ochrony, aby pomóc chronić firmę przed naruszeniami danych i innymi problemami związanymi z cyberbezpieczeństwem. Posiadacze polis ubezpieczeniowych Travellers mogą również uzyskać dostęp do narzędzi i zasobów do zarządzania oraz ograniczania ryzyka cybernetycznego przed i po ataku.

Przechodząc do szczegółów, cyber ubezpieczenie od odpowiedzialności cywilnej pokrywa koszty związane z naruszeniem danych i cyberatakami przeciwko firmie. Mogą one obejmować:

- utracone dochody z powodu jakiegoś zdarzenia cybernetycznego;
- koszty związane z powiadamianiem klientów dotkniętych naruszeniem danych osobowych;
- koszty odzyskiwania danych, które uległy uszkodzeniu;
- koszty naprawy uszkodzonych systemów komputerowych.

Ubezpieczenie może być kluczowym zabezpieczeniem przed finansowymi konsekwencjami cyberataku.

Travellers oferują indywidualne rozwiązania ubezpieczeniowe w zależności od poziomu ryzyka ubezpieczającej się firmy z opcją pokrycia między innymi kosztów dochodzenia kryminalistycznego, postępowania sądowego, zarządzania kryzysowego, wydatków na obronę, przerwania działalności ubezpieczonego oraz cyber wymuszeń.

Firma oferuje wybór pomiędzy czterema różnymi wariantami cyber ubezpieczenia "Cyber Liability Insurance".

1. "CyberRisk dla wielu branż i firm" - obejmuje szeroki zakres odpowiedzialności cybernetycznej dostosowany do potrzeb małych firm, firm z listy Fortune 500 oraz każdej organizacji pośredniej, w tym instytucji finansowych i organizacji non-profit;
2. "CyberRisk Tech dla firm technologicznych" - obejmuje szeroki zakres odpowiedzialności cybernetycznej oraz błędów i zaniechań opracowany z myślą o spełnieniu złożonych potrzeb firm technologicznych;
3. "CyberRisk dla podmiotów publicznych" - obejmuje szeroki zakres odpowiedzialności cybernetycznej mający na celu zaspokojenie potrzeb podmiotów publicznych, takich jak gminy i powiaty, władze tranzytowe i inne organizacje sektora publicznego;
4. "CyberRisk Essentials dla małych firm" - jest specjalnie dostosowany do ochrony małych firm przed zagrożeniami cybernetycznymi.

1.3. Aktualny stan wiedzy w obszarze cyber ubezpieczeń

Od momentu powstania rynku cyber ubezpieczeń rok 2020 znacząco różnił się od każdego poprzedniego. Doświadczyliśmy globalnej pandemii wywołanej wirusem SARS-CoV-2, która spowodowała bardzo duże bezrobocie, bardzo wysoki współczynnik śmiertelności na całym świecie, a nawet zamieszki i protesty na ulicach. Gospodarka wielu, szczególnie mniej rozwiniętych państw, coraz bardziej upada, a inflacja rośnie. Ten temat jest niezmiernie ważny dla rynku cyber ubezpieczeń. Przed pandemią większość prac, czynności i innych działań wykonywanych było w świecie rzeczywistym. Podczas pandemii staraliśmy się przenieść jak najwięcej z nich do świata wirtualnego. Nauka stacjonarna przeszła na naukę zdalną. Spotkania zarządu lub pracownicze w firmach stały się spotkaniami online, z wykorzystaniem zestawu narzędzi i usług służących współpracy zespołowej przez Internet. Robienie zakupów stało się zamawianiem ich przez dedykowane aplikacje internetowe. Dla wielu ludzi nawet zwykłe spotkania towarzyskie czy wydarzenia kulturalne odbywają się online. Ze wszystkim powyższym idzie w parze podwyższone ryzyko cybernetyczne. Zwiększony ruch sieciowy automatycznie generuje niebezpieczne zagrożenia dla firm, ale też zwykłych użytkowników. Stąd także świat cybernetyczny doświadczył znaczących zmian, które będą miały wpływ na nadchodzące lata, co pozostawia pole do dalszych prac nad modelowaniem cyber ubezpieczeń.

Na podstawie artykułu Thomasa Rippa [10] z 29 stycznia 2021r. opisującego wydarzenia z ostatnich lat, możemy wyciągnąć wnioski:

1. Wykorzystanie oprogramowania wymuszającego okup (*ransomware*) nie zmniejszyło się.

   Co więcej, jest coraz gorzej. Na rynku nastąpił znaczny wzrost liczby ataków i stały się one bardziej dotkliwe. Żądane kwoty okupu są coraz wyższe. Według Ponemon Institute średni koszt jednego ataku *ransomware* wynosił 4,44 mln dolarów. Oprócz *ransomware*, także przerwy w działalności biznesowej i koszty związane z przywracaniem do działania systemów były istotnie większe.

2. Bezpieczeństwo sieci stało się bardziej skomplikowane.

   Gwałtowny wzrost liczby pracowników pracujących zdalnie zagroził sieciom firmowym oraz utrudnił administratorom sieci zarządzanie infrastrukturą firmy. Niedawno firmy zajmowały się głównie zabezpieczaniem swoich danych, a teraz zostało to powiązane z zadaniem zwalczania ataków *ransomware*, które czyhają na pracowników pracujących w swoich domach. Sprawy wcale nie ułatwia pandemia, która wymusza cięcia kosztów także dla kadry zarządzającej IT.

3. Rządy odegrały większą rolę w egzekwowaniu i regulowaniu cyber zagrożeń.

   W 2018r. w Europie wprowadzono nowe rozporządzenie o ochronie danych osobowych tj. RODO, aby stworzyć w niej standaryzację zapewniając większą kontrolę danych osób przez osoby fizyczne, nakładając grzywny i kary na firmy, które je naruszyły. Nie chodzi tu tylko o sankcje, ale raczej o odpowiedzialność organizacji za ochronę swoich danych i niezwłoczne zgłaszanie ewentualnych naruszeń.

4. Cyber ryzyko jest ryzykiem globalnym.

   Firmy potrzebują klarowności, aby skutecznie działać. Stąd pojawiało się wiele wezwań do wprowadzenia globalnych standardów i lepszych uregulowań prawnych. Jednakże ponieważ świat cybernetyczny stale się zmienia, to trudno go uregulować. Wprowadzone dziś rozporządzenia mogą kompletnie nie uwzględniać nowych problemów, które pojawią się np. za 2 lata. Poza tym komisje nadzoru w różnych krajach mogą działać trochę inaczej, a na dodatek mogą one mieć własne przepisy i ramy czasowe raportowania, co też utrudnia globalne uregulowania.

5. Cyberbezpieczeństwo pozostaje niedocenianym ryzykiem dla wielu firm.

   Większość firm nadal nie wykupuje żadnej ochrony cybernetycznej. Cyber ubezpieczenie jest często postrzegane jako "miło je mieć", a cyber straty to coś, czego doświadczają inne firmy. Obecnie koszt zakupu polisy cybernetycznej to pieniądze, których menedżerowie ds. ryzyka mogą nie mieć w obliczu rosnących kosztów innych potrzebnych im programów ubezpieczeniowych. Apetyt na ubezpieczenia cybernetyczne wzrósł tylko umiarkowanie, podczas gdy cyberataki są coraz bardziej popularne, a żądane kwoty okupu *ransomware* są coraz wyższe.

2. Modelowanie cyberubezpieczeń

Problem modelowania w obszarze cyberubezpieczeń możemy ogólnie podzielić na dwie grupy:

- Modelowanie zmiennych określających rozmiar naruszeń oraz czasy między naruszeniami;
- Modelowanie całej sieci.

W kolejnych podrozdziałach omówię teoretyczne podstawy każdego z tych podejść oraz wskażę ich wady i zalety.

2.1. Modelowanie rozkładami prawdopodobieństwa

W niniejszym podrozdziale omówię podejście modelowania kluczowych wspomnianych wyżej zmiennych, tj.:

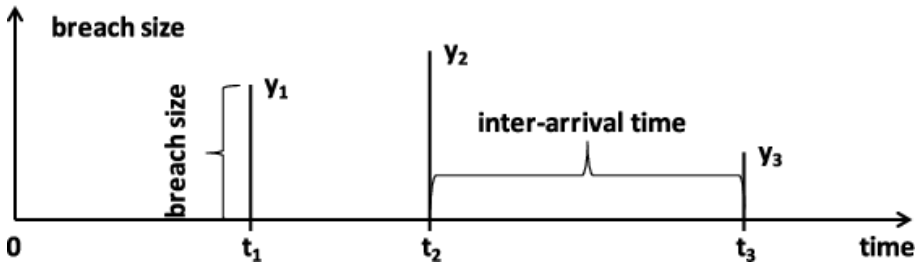
- *breach sizes* czyli rozmiarów naruszeń lub dotkliwości;
- *interarrival times* czyli czasów między zdarzeniami (np. atakami) lub po prostu częstotliwości,

przy użyciu znanych rozkładów prawdopodobieństwa. Takie podejście spotykane jest w literaturze, patrz np. Wheatley i inni [22], Eling i Loperfido [23] lub Edwards i inni [59].

Zacznę od wprowadzenia podstawowych definicji, które będę wykorzystywał do końca swojej pracy. Jedna z nich przybliży, czym są tzw. czasy między zdarzeniami, o których wspomniałem powyżej.

Definicja 2.1. Niech t oznacza czas używany do opisywania danego modelu, ciąg $\{(t_i, y_{t_i})\}_{i\geqslant 0}$ oznacza i-ty incydent występujący w czasie t_i z rozmiarem naruszenia (*breach sizes*) y_{t_i} oraz niech $d_i = t_i - t_{i-1}$ oznaczają czasy między zdarzeniami (*interarrival times*).

Powyższe definicje obrazuje rysunek 2.1.



Rysunek 2.1: Ilustracyjny opis zmiennych *breach sizes* oraz *interarrival times* (zob. [1])

Praca Elinga i Loperfido [23] skupiona jest wokół empirycznej weryfikacji zastosowania rozkładów prawdopodobieństwa w modelowaniu cyberubezpieczeń. Dokładniej, zmienna *interarrival times* modelowana była przez rozkłady dyskretne, takie jak rozkład Poissona i rozkład ujemny dwumianowy (przedstawione na końcu tablicy 2.2). Natomiast zmienna *breach sizes*, czyli rozmiary naruszeń, modelowana była przez wybrane rozkłady ciągłe (przedstawione w tablicy 2.1 oraz 2.2)[1].

W celu zbadania zgodności założonych rozkładów teoretycznych z rozkładem empirycznym autorzy wykorzystali testy zgodności Kołmogorowa - Smirnowa i Andersona - Darlinga oraz kryteria informacyjne AIC i BIC. Podejście wykorzystujące modelowanie przy użyciu rozkładów prawdopodobieństwa, choć jest łatwe w implementacji, a także bardzo intuicyjne w interpretacji, nie zawsze będzie prowadzić do jednoznacznych wniosków. Stąd, coraz częściej w literaturze znaleźć można propozycję innych technik mających zastosowanie w tym problemie.

[1] O kolumnie *Estymacja* powiem więcej w kolejnym rozdziale.

Tablica 2.1: Analizowane rozkłady (patrz [28])

Nazwa	**Parametry**	**Gęstość**	**Dystrybuanta**	**Estymacja**
Wykładniczy	$\lambda > 0$	$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)}$	$F(x) = 1 - e^{-\lambda x}$	Na podst. $EX = \frac{1}{\lambda}$.
Gamma	$k > 0$, $\theta > 0$	$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$	$F(x) = \frac{1}{\Gamma(k)} \gamma(k, \frac{x}{\theta})$	Na podst. W.O. i wariancji. Po przekształceniach $k = \frac{(\mathbb{E}(X))^2}{Var(X)}$ oraz $\theta = \frac{Var(X)}{\mathbb{E}(X)}$.
Log - Normalny	$\mu \in \mathbb{R}$, $\sigma > 0$	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\Big(-\frac{(\ln x - \mu)^2}{2\sigma^2}\Big)$	$F(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\Big(\frac{\ln x - \mu}{\sigma\sqrt{2}}\Big)$	Z W.O. i mediany. Po przekształceniach: $\mu = \ln(\text{median}(X))$, $\sigma = \sqrt{\Big\lvert 2 \ln\big(\frac{\mathbb{E}(X)}{\text{median}(X)}\big)\Big\rvert}$.
Normalny	$\mu \in \mathbb{R}$, $\sigma > 0$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$	$F(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\Big(\frac{x-\mu}{\sigma\sqrt{2}}\Big)$	Z W.O. μ i wariancji σ^2.
Weibulla	$\lambda \in (0,\infty)$, $k \in (0,\infty)$	$f(x) = \frac{k}{\lambda}\Big(\frac{x}{\lambda}\Big)^{k-1} e^{-(x/\lambda)^k} \mathbb{1}_{[0,\infty)}$	$F(x) = 1 - e^{-(x/\lambda)^k} \mathbb{1}_{[0,\infty)}$	Z funkcji eweibull(), z biblioteki EnvStats.
Skośny - Normalny	$\xi \in \mathbb{R}$, $\omega > 0$, $\alpha \in \mathbb{R}$	$f(x) = \frac{2}{\omega\sqrt{2\pi}} \cdot \exp\Big(-\frac{(x-\xi)^2}{2\omega^2}\Big) \cdot \int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$	$F(x) = \Phi(\frac{x-\xi}{\omega}) - 2T(\frac{x-\xi}{\omega}, \alpha)$, gdzie $T(h,\alpha)$ - funkcja T Owena	Z funkcji snormFit(), z biblioteki fGarch.

Tablica 2.2: Analizowane rozkłady (patrz [28])

Nazwa	**Parametry**	**Gęstość**	**Dystrybuanta**	**Estymacja**
Log - Logistyczny	$\alpha > 0$, $\beta > 0$	$f(x) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2}$	$F(x) = \frac{1}{1+(x/\alpha)^{-\beta}}$	Z funkcji fitdist(), z biblioteki fitdistrplus.
Skośny - T - Studenta	$\mu > 0$, $\sigma > 0$, $\lambda > 0$, $q > 0$	$f(x) = \frac{\Gamma\left(q+\frac{1}{2}\right)}{\nu\sigma\sqrt{\pi q}\Gamma(q)\left(1+\frac{\lvert x-\mu+m\rvert^2}{q(\nu\sigma)^2(\lambda\,\mathrm{sign}(x-\mu+m)+1)^2}\right)^{q+\frac{1}{2}}}$, gdzie $m = \frac{2\nu\sigma\lambda q^{1/2}\Gamma(q-\frac{1}{2})}{\pi^{1/2}\Gamma(q)}$, $\nu = q^{-1/2}((3\lambda^2+1)(\frac{1}{2q-2}) - \frac{4\lambda^2}{\pi}(\frac{\Gamma(q-\frac{1}{2})}{\Gamma(q)})^2)^{-1/2}$	$F(x) = \int_{\mathbb{R}} f(x)dx$ (brak zamkniętej formuły analitycznej)	Z funkcji sstdFit(), z biblioteki fGarch.
Uogólniony Pareto (GPD)	$\mu \in \mathbb{R}$, $\sigma > 0$, $\xi \in \mathbb{R}$	$f(x) = \frac{1}{\sigma}(1+\xi\frac{x-\mu}{\sigma})^{-(\frac{1}{\xi}+1)}$	$F(x) = 1-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}$	Z funkcji gpdRangeFit(), z biblioteki texmex.
Ujemny Dwumianowy	$n > 0$, $p \in [0,1]$	$\mathbb{P}(X = x) = \frac{\Gamma(x+n)}{x!\Gamma(n)}p^n(1-p)^x$	$\mathbb{P}(X \leqslant x) = \sum_{i=0}^{x}\binom{n+i-1}{i}\cdot p^n(1-p)^i$	Z W.O. i wariancji. Po przekształceniach: $n = \frac{(\mathbb{E}(X))^2}{Var(X)-\mathbb{E}(X)}$ oraz $p = \frac{\mathbb{E}(X)}{Var(X)}$.
Poisson	$\lambda > 0$	$\mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$	$\mathbb{P}(X \leqslant x) = \sum_{i=0}^{x} e^{-\lambda}\frac{\lambda^i}{i!}$	na podst. W.O. $\mathbb{E}(X) = \lambda$.

Eksploracja danych

We wspomnianej przeze mnie pracy autorzy przeprowadzili także analizę eksploracyjną, która miała na celu wykrycie wcześniej nieznanych cech danych. Zdecydowali się wykorzystać metodę skalowania wielowymiarowego (ang. *MultiDimensional Scaling*, w skrócie MDS). Polega ona (patrz Ćwik, Mielniczuk [67]) na takiej reprezentacji danych w przestrzeni niskowymiarowej, że odległości między reprezentantami względnie dobrze odzwierciedlają odległości między odpowiadającymi obiektami. Metoda skalowania metrycznego stosowana jest wtedy, gdy odległości spełniają warunek trójkąta [2]. Dokładniej, metoda ta wykorzystuje macierz odległości między poszczególnymi elementami, tj. macierz $D \in R_+^{n\times n}$ taką, że $d_{ij} = d(x_i, x_j)$, gdzie d oznacza funkcję odległości np. metrykę euklidesową.

Szczegółowy opis metody MDS, wraz z jego potencjalnymi zastosowaniami można znaleźć np. w [55] lub [66]. Autorzy wykorzystali implementację metody udostępnionej przy pomocy funkcji cmdscale(), w języku R. Przeprowadzona w [23] analiza miała na celu zbadanie różnic między:

- typami organizacji, które narażone są na naruszenia bezpieczeństwa danych;
- wielorakimi typami ataków.

Ostatecznie, z wniosków [23] wypływających z wykorzystania metody MDS dostajemy informację, iż różne rodzaje naruszeń danych muszą być modelowane jako odrębne kategorie ryzyka. Wykorzystano dane, które opisane są szczegółowo w rozdz. 3.1. Biorąc pod uwagę wyniki otrzymane przez autorów artykułu, w dalszej części swojej pracy będę uwzględniał podział na różne kategorie naruszeń i organizacji.

2.2. Modelowanie z wykorzystaniem procesów stochastycznych

Innym, popularnym w literaturze podejściem do modelowania, jest wykorzystanie procesów stochastycznych. Na przykład w pracach [1, 9, 24, 29, 60] lub [61] stwierdzono, że modelowanie zmiennych *breach sizes* oraz *interarrival times* przy użyciu rozkładów wiąże się m. in. z problemem:

(a) weryfikacji na podstawie konkretnej próbki (nie jest to opisane wzorami);

(b) wykorzystywania testów statystycznych o różnej mocy, itp.

[2] Gdy odległości są euklidesowe, MDS i analiza składowych głównych (PCA) prowadzą do tych samych wyników, zatem pierwsza z nich uogólnia drugą.

Co więcej, w pracy [24] zauważono, że istnieje autokorelacja pomiędzy *iterarrival times*. Z tego względu poszczególne próbki nie są niezależne i nie powinny być modelowane przez rozkład. Dalej, w pracach [24] oraz [61] dodatnia zależność między *breach sizes* oraz *interarrival times* była opisana przy użyciu wyselekcjonowanej kopuły.
Dla przypomnienia, dysponujemy wartościami:

- t_i - czas w którym wystąpił incydent (t_0 to moment rozpoczęcia obserwacji);
- y_{t_i} - wielkość naruszenia w czasie t_i;
- $\{(t_i, y_{t_i})\}_{0 \leqslant i \leqslant n}$ - ciąg par, gdzie t_i i y_{t_i} dane powyżej;
- $d_i = t_i - t_{i-1}$ - *interarrival times*;

Dodatkowo, niech $C(\cdot)$ oznacza wybraną kopułę używaną do modelowania zależności.
Bazując na artykułach [24, 60, 62, 63] zaproponuję teraz ogólny schemat algorytmu służącego do predykcji szeregów czasowych d_i i y_{t_i}, obliczenia wartości narażonej na ryzyko (VaR), a także do zliczenia całkowitej liczby przekroczeń *backtestingu*. Formalne definicje wspomnianych pojęć przedstawię w dalszej części pracy. Zauważmy, że kolejne punkty poniższego schematu są szczegółowo omówione w odpowiadających im częściach rozdz. 3.3.2 i 3.3.3.

Schemat

1. Dopasowanie odpowiedniego modelu i jego testowanie dla:
   (a) zmiennej *interarrival times*;
   (b) zmiennej zlogarytmowanej *breach sizes*;
2. Określenie zależności pomiędzy obydwiema zmiennymi oraz:
   (a) w przypadku zależności:
      - wybór odpowiedniej kopuły przy użyciu dwuwymiarowych residuów dla wybranych modeli na podstawie kryteriów;
   (b) w przypadku niezależności:
      - przejście do kolejnego podpunktu;
3. Wyznaczenie predykcji i jej testowanie dla:
   (a) zmiennej *interarrival times*;
   (b) zmiennej zlogarytmowanej *breach sizes*;

4. Obliczenie wartości narażonej na ryzyko (VaR), znalezienie jej predykcji oraz przeprowadzenie testów dla:

   (a) zmiennej *interarrival times*;

   (b) zmiennej zlogarytmowanej *breach sizes*.

2.3. Modele dla ryzyka cyberbezpieczeństwa

W przeciwieństwie do opisanych powyżej rozwiązań, w pracy [1] zaproponowano modelowanie całej sieci. Dokładniej, załóżmy, że sieć komputerowa będzie opisana przy pomocy nieskierowanego grafu $\Gamma = (V, E)$, gdzie V to zbiór węzłów, natomiast E to zbiór krawędzi. Dalej, graf Γ reprezentuje strukturę sieci, która narażona jest na ataki cybernetyczne (np. rozprzestrzenianie się złośliwego oprogramowania). Zauważmy, że krawędź $(u, v) \in E$ oznacza, że węzły u i v mogą się wzajemnie atakować (graf nieskierowany). Ogólnie grafem Γ może być graf o dowolnej strukturze, np. graf pełny (tzn. dowolny węzeł $u \in V$ może zaatakować dowolny $v \in V$). Oznaczmy przez $A = (a_{vu})$ macierz sąsiedztwa grafu Γ, gdzie $a_{vu} = 1$, gdy $(u, v) \in E$ oraz $a_{vu} = 0$, gdy $(u, v) \notin E$. Zauważmy, że $a_{vv} = 0$, zatem macierz sąsiedztwa A jest macierzą symetryczną z zerami na przekątnej. Następnie oznaczmy przez $\deg(v)$ stopień węzła (tzn. liczbę węzłów, z którymi dany węzeł jest połączony bezpośrednio) v oraz przez $N = |V|$ całkowitą liczbę węzłów.

Jak zauważono w pracy [1] węzeł $v \in V$ w dowolnym czasie $t = 0, 1, ...$ może być albo:

- bezpieczny (jednakże podatny na ataki), lub
- zarażony (i sam może zarażać inne węzły).

Stan tej sieci w czasie t można więc zapisać w postaci:

$$(I_1(t), I_2(t), ..., I_N(t)),$$

gdzie $I_v(t) = 1$ oznacza, że węzeł v został zarażony w czasie t, natomiast $I_v(t) = 0$ oznacza, że węzeł v jest bezpieczny w czasie t. Następnie wektor prawdopodobieństw zarażenia oznaczamy przez:

$$\mathbf{p}^T(t) = (p_1(t), ..., p_N(t)),$$

gdzie $p_j(t) = \mathbb{P}(I_j(t) = 1)$, dla $j = 1, ..., N$. Dalej, dla węzłów rozpatrujemy następujące dwa zagrożenia:

(i) Zagrożenia z zewnątrz sieci tzn. gdy węzeł v zostanie zarażony wskutek ataku, lub gdy użytkownik odwiedzi zarażoną witrynę,

(ii) Zagrożenia wewnątrz sieci tzn. jeśli węzeł v jest zarażony, to zaatakuje on swoich sąsiadów.

Zakładamy także, że jeśli węzeł v jest zarażony, to zostanie on naprawiony lub oczyszczony, aby powrócił do stanu bezpiecznego. Infekcja wywołuje następujące dwa typy strat:

(a) strata wywołana przez infekcję jak np. wykradzenie lub uszkodzenie danych, ujawnienie poufnych danych, koszty prawne lub koszty ewentualnego postępowania sądowego,

(b) strata wywołana przez naprawę zainfekowanego węzła i jego powrót do stanu bezpiecznego.

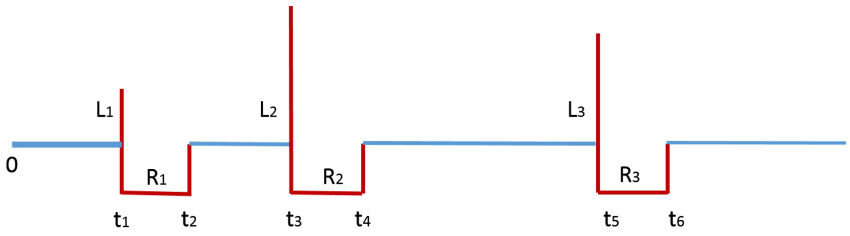
Stratę typu (a) modelujemy funkcją kosztów $\eta_v(L_{v,1})$, gdzie $L_{v,1}$ oznacza utratę danych (np. uszkodzenie danych). Strata typu (b) jest związana z długością spowolnienia obsługi (lub naprawy). Modelujemy ją funkcją kosztów $C_v(R_{v,1})$, gdzie $R_{v,1}$ oznacza długość spowolnienia obsługi. Zatem dla węzła v skumulowana strata do czasu t może być przedstawiona w następujący sposób:

$$s_v(t) = \sum_{i=1}^{M_v(t)} \left[\eta_v(L_{v,i}) + C_v(R_{v,i})\right],$$

gdzie $M_v(t)$ jest całkowitą liczbą infekcji węzła v do czasu t, $\eta_v(\cdot)$ oznacza koszty z powodu infekcji, oraz $C_v(\cdot)$ oznacza funkcję kosztów związaną z długością czasu $R_{v,i}$ spowolnienia obsługi. Całkowita strata towarzystwa ubezpieczeniowego w czasie $(0, t]$ ma wtedy postać:

$$S(t) = \sum_{v=1}^{N} s_v(t) = \sum_{v=1}^{N} \sum_{i=1}^{M_v(t)} \left[\eta_v(L_{v,i}) + C_v(R_{v,i})\right]. \tag{2.1}$$

Przykład 2.2. Przyjrzyjmy się teraz przykładowi zaprezentowanemu przez M. Xu, L. Hua [1]. Przeanalizujmy rysunek 2.2[3].



Rysunek 2.2: Ryzyko cyberbezpieczeństwa dla węzła v (zob. [1])

Widzimy, że węzeł v jest bezpieczny w czasie $T = 0$, a pierwsze zainfekowanie pojawia się w czasie $T = t_1$. Następnie, w czasie $T = t_2$ węzeł v jest ponownie bezpieczny i znowu podatny

[3]Zwróćmy uwagę na rys. 2.2 w kontekście rys. 2.1. Zauważmy, że L_i odpowiadają stracie *breach size*. Z kolei $t_3 - t_1$, $t_5 - t_3$, itd. odpowiadają *interarrival times*. Innymi słowy, czas między zdarzeniami to suma czasu spowolnienia obsługi R_i oraz pozostałego czasu do następnego zdarzenia, oznaczonego na rysunku linią niebieską.

na ataki, a w czasach t_3 oraz t_5 po raz kolejny został zainfekowany. W tym przypadku $M_v(t) = 3$, więc skumulowana strata do czasu t ma następującą postać

$$s_v(t) = \eta_v(L_1) + C_v(R_1) + \eta_v(L_2) + C_v(R_2) + \eta_v(L_3) + C_v(R_3),$$

gdzie $\eta_v(L_i)$ - koszty z powodu infekcji oraz $C_v(R_i)$ - koszty związane z długością czasu spowolnienia obsługi, dla $i = 1, 2, 3$.

2.3.1. Modelowanie sieci $I(t) = (I_1(t), I_2(t), ..., I_N(t))$

W tym podrozdziale rozpatrzę różne modele służące do modelowania ryzyka cyberbezpieczeństwa. Zacznę jednak od wprowadzenia niezbędnej definicji i twierdzenia oraz obrazującego je przykładu (patrz J. Jakubowski, R. Sztencel [42]).

Definicja 2.3 (proces Poissona). Rodzinę nieujemnych zmiennych losowych $\{N_t : t \geqslant 0\}$ na $(\Omega, \mathcal{F}, \mathbb{P})$ nazywamy procesem Poissona o intensywności λ gdy:

(a) $\mathbb{P}(N_0 = 0) = 1$ (mówimy, że ten proces startuje z zera),

(b) Dla $0 < t_1 < t_2 < ... < t_k$ przyrosty $N(t_1), N(t_2) - N(t_1), ..., N(t_k) - N(t_{k-1})$ są niezależne,

(c) poszczególne przyrosty mają rozkład Poissona tzn. $\mathbb{P}(N_t - N_s = n) = e^{-\lambda(t-s)} \frac{(\lambda(t-s))^n}{n!}$, dla $n = 0, 1, 2, ...$ i $s \in [0, t)$.

Twierdzenie 2.4 (charakteryzacja procesu Poissona). Niech $\{N_t : t \geqslant 0\}$ to proces o stacjonarnych przyrostach niezależnych, oraz $N_0 = 0$ - p.n. Wtedy następujące warunki są równoważne:

(a) N jest procesem Poissona o intensywności λ,

(b) $\forall_{t \geqslant 0}\ \mathbb{P}(N_{t+h} - N_t = 1) = \lambda h + o(h)$ (tzn. skoki mają wartość równą 1) oraz $\mathbb{P}(N_{t+h} - N_t > 1) = o(\lambda)$, na przedziale $(t, t + h]$.

Przykład 2.5 (konstrukcja procesu Poissona). Rozpatrzmy zdarzenie losowe, np. wezwania telefoniczne. Załóżmy, że przeszłość nie ma wpływu na przyszłość. Można przypuszczać, że czas oczekiwania na pierwszy sukces (sygnał) będzie zmienną losową z własnością braku pamięci, czyli zmienną losową o rozkładzie wykładniczym. Po każdym sukcesie wszystko zaczyna się od nowa. Kolejne czasy oczekiwania X_k są więc niezależnymi zmiennymi losowymi o jednakowym rozkładzie wykładniczym. Definiujemy zmienną losową liczącą sygnały, które pojawiły się do

momentu t:

$$N_t = \begin{cases} 0 & \text{dla } X_1 > t, \\ \sup\{n : X_1 + ... + X_n \leqslant t\} & \text{dla } X_1 \leqslant t. \end{cases}$$

Rodzinę $\{N_t : t \geqslant 0\}$ nazywa się procesem Poissona.

Model Markowa

W modelu[4] tym zakładamy, że proces regeneracji dowolnego zainfekowanego węzła v jest procesem Poissona z intensywnością δ_v. Z kolei proces zainfekowania węzła wewnątrz sieci także jest procesem Poissona, ale z intensywnością β. Dodatkowo zakładamy, że każdy węzeł może być też zainfekowany z zewnątrz sieci z intensywnością Poissona ε_v. Zatem oprócz możliwości zarażenia się węzłów wewnątrz sieci, mogą być one zarażone także z zewnątrz. Stąd proces infekcji jest procesem Poissona z intensywnością: $\beta \sum_{j=1}^{n} a_{vj} I_j(t) + \varepsilon_v$. Przyjmujemy, że procesy regeneracji i infekcji są niezależne.

Inne sposoby modelowania

W przypadku modelowania bez wykorzystania modelu Markowa zakładamy, że dla dowolnego węzła v istnieje D_v zainfekowanych sąsiadów. Stąd v może być zaatakowany przez swoich sąsiadów wewnątrz sieci, przy czym czasy do zainfekowania przez danego sąsiada są zmiennymi losowymi $Y_{v_1}, ..., Y_{v_{D_v}}$ o takim samym rozkładzie F. Z kolei czas do infekcji spowodowanej przez zagrożenia spoza sieci, modelowany jest przy pomocy zmiennej losowej Z_v o dystrybuancie G_v. W związku z tym czas do zainfekowania węzła v wyraża się wzorem:

$$T_v = \min\{Y_{v_1}, ..., Y_{v_{D_v}}, Z_v\}.$$

Zakładamy też, że jeśli węzeł v jest zainfekowany, to przestaje on być celem ataków aż do momentu jego naprawy. Wtedy ponownie jest narażony na atak. Czas regeneracji zainfekowanego węzła v oznaczmy przez R_v. Zauważmy, że:

$$D_v = \sum_{j=1}^{N} a_{vj} I_j,$$

gdzie I_j oznacza stan węzła j oraz:

$$p_j = \mathbb{P}(I_j = 1).$$

Stąd mamy:

$$\mathbb{E}[D_v] = \sum_{j=1}^{N} a_{vj} p_j.$$

[4]W literaturze model ten znany jest także jako ε-SIS model (patrz np. [8]).

W pracy [1] pokazano, że $\mathbb{E}[T_v]$ (czyli średni czas do zainfekowania węzła v) możemy przedstawić w postaci:

$$\begin{aligned}\mathbb{E}[T_v] &= \mathbb{E}[\min\{Y_{v_1},...,Y_{v_{D_v}},Z_v\}] = \mathbb{E}[\mathbb{E}[\min\{Y_{v_1},...,Y_{v_{D_v}},Z_v\}|D_v]] = \\ &= \sum_{d_v=0}^{\deg(v)} \mathbb{P}(D_v = d_v) \int_0^\infty \overline{H}_{d_v}(x,...,x)\overline{G}_v(x)dx,\end{aligned} \tag{2.2}$$

gdzie $\overline{H}_{d_v}$ jest funkcją przeżycia rozkładu łącznego $\{Y_{v_1},...,Y_{v_{d_v}}\}$, tj.:

$$\overline{H}_{d_v}(x,...,x) = \mathbb{P}(Y_{v_1} > x,...,Y_{v_{d_v}} > x),$$

dla $d_v \geqslant 1$, $\overline{H}_0 \equiv 1$ oraz $\overline{G}_v$ jest funkcją przeżycia dla zmiennej Z_v:

$$\overline{G}_v(x) = \mathbb{P}(Z_v > x).$$

Zauważmy, że aby wyznaczyć $\mathbb{E}[T_v]$ musimy zamodelować rozkład łączny $\{Y_{v_1},...,Y_{v_{d_v}}\}$ (por. (2.2)). W tym celu wykorzystamy kopuły (por. np. [1] lub [43]) pozwalające na skuteczne uchwycenie wielowymiarowej zależności między zmiennymi.

Wprowadzę teraz pojęcie kopuły, omówię przydatne własności i wymienię kilka potrzebnych klas.

Kopuły

Kopuły to podejście, które ma za zadanie połączyć rozkłady brzegowe i strukturę zależności (patrz [43]).

Definicja 2.6. n - wymiarową kopułą nazywamy funkcję $C : [0,1]^n \to [0,1]$ o następujących właściwościach:

(i) $C(u_1,...,u_n)$ jest rosnącą funkcją dla każdego komponentu u_i,

(ii) $\forall i \in 1,...,n : \quad C(1,...,1,u_i,1,...,1) = u_i$,

(iii) $\forall u_j \in [0,1]$, gdzie $j = 1,...,n$ i $j \neq i : \quad C(u_1,...,u_{i-1},0,u_{i+1},...,u_n = 0$,

(iv) Niech $0 \leqslant a_j \leqslant b_j \leqslant 1$ dla każdego $j = 1,...,n$ zachodzi:

$$\sum_{i_1=1}^{2} ... \sum_{i_n=1}^{2} (-1)^{i_1+...+i_n} C(u_{i_1},...,u_{i_n}) \geqslant 0$$

gdzie dla każdego $j = 1,..,n : \; u_{i_j} = a_j$ dla $i_j = 1$ i $u_{i_j} = b_j$ dla $i_j = 2$.

Niech $X_1,...,X_n$ będzie ciągiem zmiennych losowych o dystrybuantach odpowiednio $F_1,...,F_n$. Rozważmy rozkład łączny $F(x_1,...,x_n) = \mathbb{P}(X_1 \leqslant x_1,...,X_n \leqslant x_n)$. Twierdzenie Sklara (por. [43, 60]), przedstawione poniżej, pozwala dla dowolnej zmiennej wielowymiarowej o rozkładzie łącznym F wyznaczyć kopulę C.

Twierdzenie 2.7 (Sklar). Niech F będzie dystrybuantą wielowymiarowej zmiennej losowej na $\mathbb{R}^n$ z dystrybuantami brzegowymi $F_1, ..., F_n$. Wtedy istnieje kopuła C taka, że:

$$F(x_1, ..., x_n) = C\Big(F_1(x_1), ..., F_n(x_n)\Big).$$

Kopuła C jest jednoznacznie określona na przestrzeni $\text{Ran}F_1 \times ... \times \text{Ran}F_n$, gdzie $\text{Ran}F_i = \{F_i(x_i) : x_i \in \mathbb{R}\}$. Ponadto, jeśli F_i jest ciągła to $\text{Ran}F_i = [0, 1]$.

To znaczy nieznany, wielowymiarowy, łączny rozkład może być przybliżony najlepiej dopasowaną kopułą oraz odpowiednimi rozkładami brzegowymi. Poniżej przedstawię jeszcze pomocniczą uwagę wynikającą z twierdzenia Sklara (zob np. [43]).

Uwaga 2.8 (Sklar). Jeśli $F_1, ..., F_n$ są ściśle rosnące i ciągłe, to kopuła C może być opisana przy pomocy dystrybuanty F wielowymiarowej zmiennej losowej i funkcji odwrotnych dystrybuant brzegowych $F_1^{-1}, ..., F_d^{-1}$:

$$C(u_1, ..., u_d) = F(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)).$$

Rozważmy teraz następujące przykłady rodzin kopuł.

Przykład 2.9 (Kopuła Gaussa). Kopułę Gaussa definiujemy za pomocą wielowymiarowego rozkładu normalnego. Dokładniej:

$$C^{Ga}(u_1, ..., u_n) = \Phi_\Sigma\Big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)\Big),$$

gdzie Φ^{-1} jest dystrybuantą odwrotną standardowego rozkładu normalnego, a Φ_Σ jest dystrybuantą łączną wielowymiarowego rozkładu normalnego o średniej równej zero oraz macierzy kowariancji Σ równej macierzy korelacji. Dla uproszczenia przyjmujemy, że macierz korelacji ma postać:

$$\Sigma = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \dots & \rho_{1,n} \\ \rho_{2,1} & 1 & \rho_{2,3} & \dots & \rho_{2,n} \\ \rho_{3,1} & \rho_{3,2} & 1 & \dots & \rho_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \rho_{n,3} & \dots & 1 \end{bmatrix}, \tag{2.3}$$

gdzie $\rho_{i,j}$ to współczynnik korelacji pomiędzy dwiema zmiennymi. W tym przypadku kopułę Gaussa możemy zapisać w postaci:

$$C^{Ga}(u_1, ..., u_n) = \Phi_\rho\Big(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)\Big) \tag{2.4}$$

Przykład 2.10 (Klasa kopuł Archimedesa). Kolejnym przykładem jest klasa kopuł Archimedesa, do których należą między innymi kopuła Gumbela, Claytona i Franka.

Definicja 2.11. Kopułą Archimedesa nazywamy kopułę następującej formy:

$$C^{Ar}(u_1, ..., u_n) = \psi\Big(\psi^{-1}(u_1), ..., \psi^{-1}(u_n)\Big),$$

gdzie $\psi(\cdot)$ jest funkcją generującą spełniającą następujące warunki:

(i) $\psi : [0, +\infty) \to [0, 1]$,

(ii) $\psi(0) = 1$, $\lim\limits_{x \to +\infty} \psi(x) = 0$,

(iii) $\psi(\cdot)$ jest funkcją ciągłą, nierosnącą na całej dziedzinie, a dodatkowo jest ściśle malejącą na przedziale $[0, \inf\{u : \psi(x) = 0\})$.

Szczególnymi przypadkami kopuły Archimedesa są[5]:

1. Kopuła Gumbela:

$$C_\theta^{Gu}(u_1, u_2) = \exp[-((-\log u_1)^\theta + (-\log u_2)^\theta)^{1/\theta}], \;\; 1 \leqslant \theta < \infty;$$

2. Kopuła Claytona (która modeluje dodatnią zależność, a zwłaszcza zależność dolno ogonową):

$$C_\theta^{Cl}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \;\; 0 < \theta < \infty;$$

3. Kopuła Franka:

$$C_\theta^{Fr}(u_1, u_2) = -\frac{1}{\theta} \log\Big(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1) - 1}{e^{-\theta} - 1}\Big), \;\; \theta \in \mathbb{R} \setminus \{0\};$$

4. Kopuła Joe'a:

$$C_\theta^{Jo}(u_1, u_2) = 1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta}, \;\; 1 \leqslant \theta < \infty;$$

5. Kopuła BB6 (Joe - Gumbel) [75]:

$$C_{\theta,\delta}^{BB6}(u_1, u_2) = 1 - (1 - \exp(-[(-\log(1-u_1)^\theta)))^\delta + (-\log(1-(1-u_2)^\theta))^\delta]^{1/\delta})^{1/\theta}, \;\; 1 \leqslant \theta, 1 \leqslant \delta;$$

6. Kopuła BB8 (Joe - Frank) [75]:

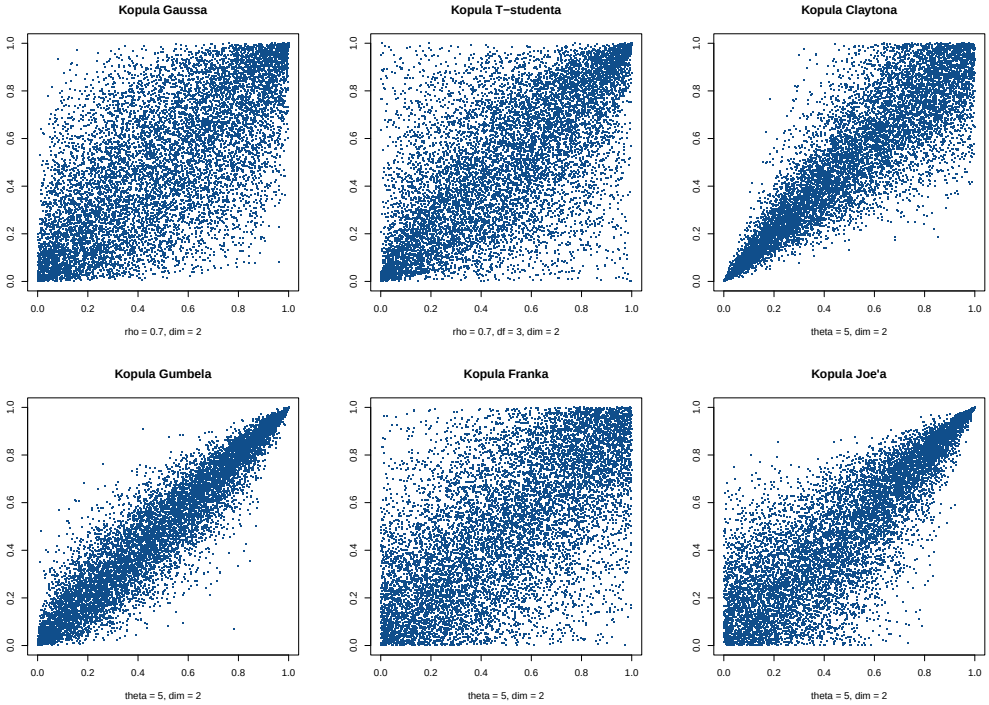
$$C_{\theta,\delta}^{BB8}(u_1, u_2) = \frac{1}{\theta}(1 - [1 - \frac{1}{1-(1-\delta)^\theta}(1 - (1-\delta u_1)^\theta)(1-\delta u_2)^\theta)]^{1/\delta}), \;\; 1 \leqslant \theta, 0 < \delta \leqslant 1;$$

7. Kopuła Tawn [75]:

$$C_\theta^{Ta}(u_1, u_2) = u_1 u_2 \exp\Big(\theta \frac{(\log u_1)(\log u_2)}{\log u_1 + \log u_2}\Big), \;\; 0 \leqslant \theta \leqslant 1.$$

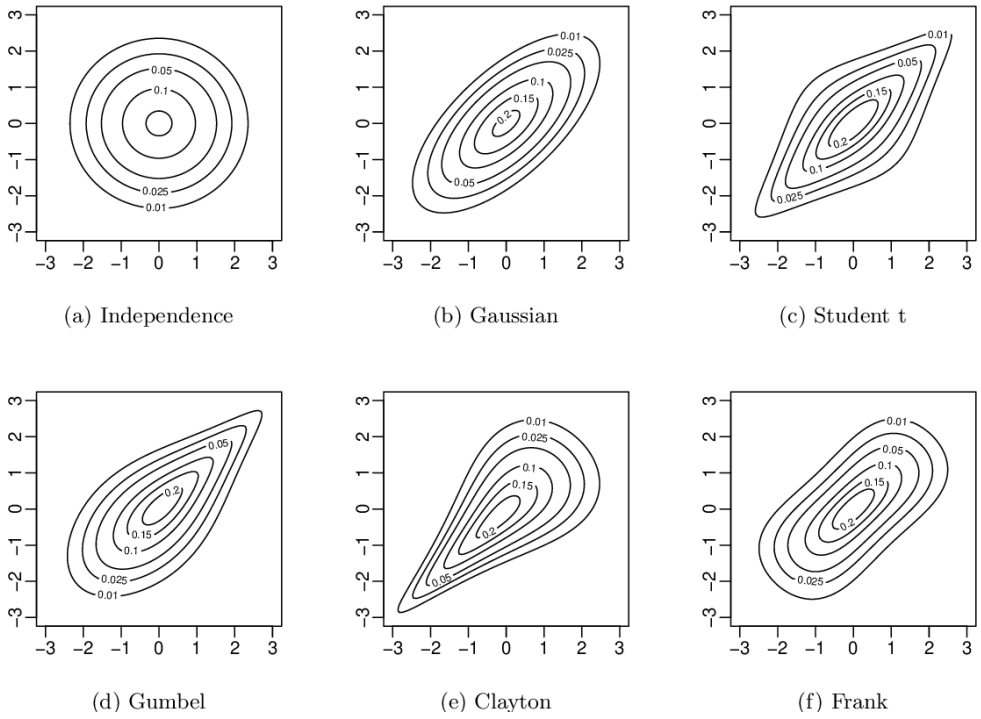
[5]Dla uproszczenia zapisu podaję ich wersje dwuwymiarowe

Następnie przyjrzyjmy się rys. 2.3, który przedstawia wykresy rozproszenia wygenerowanych próbek z sześciu podstawowych typów kopuł dwuwymiarowych odpowiednio: Gaussa, T – studenta, Claytona, Gumbela, Franka i Joe'a. Wykres został wykonany przy pomocy pakietu copula w języku R. Wartości parametrów dla tych kopuł to $0.7 \in (-1, 1)$ w przypadku Gaussa i T – studenta oraz 5 w przypadku Gumbela, Claytona, Franka i Joe'a. Dodatkowo liczba stopni swobody dla kopuły T – studenta została ustalona i jest równa 3. Rysunek ten pomaga w zrozumieniu korelacji pomiędzy różnymi zmiennymi.



Rysunek 2.3: Porównanie kopuł (inspirowany [43]).

Popatrzmy teraz na rys. 2.4 przedstawiający wykresy konturowe dla gęstości wybranych kopuł (patrz [64]).



(a) Independence (b) Gaussian (c) Student t

(d) Gumbel (e) Clayton (f) Frank

Rysunek 2.4: Wykresy konturowe gęstości dla wybranych kopuł (wykresy zaczerpnięte z [64]).

Widzimy, że gdy zmienne losowe są od siebie niezależne, to kontury są okręgami (patrz 2.4(a)). Dalej, (patrz 2.4(b)) przedstawiona jest kopuła Gaussa, której kontury są elipsami. Kolejny wykres (patrz 2.4(c)) przedstawia kopułę T-studenta. W tym wypadku kontury przypominają diamenty – wynika to z faktu, że w przeciwieństwie do kopuły Gaussa, kopuła T-studenta wykazuje zależność w ogonach (*tail dependence*). Można to ponownie zobaczyć na rys. 2.4(d) i 2.4(e), gdzie ukazane są kopuły Gumbela i Claytona. Kopuła Gumbela jest asymetryczna i charakteryzuje się zależnością tylko górno - ogonową. Widać to po kolczastym kształcie w prawym górnym rogu i bardziej płaskim w lewym dolnym rogu. Dla kopuli Claytona jest na odwrót. Na końcu, kopuła Franka (patrz 2.4(f)) nie ma zależności ogonowej i ma lżejsze ogony niż Gaussa, co odpowiada bardziej płaskiemu kształtowi konturów.

Ogólnie, kopuły mają zastosowanie np. w analizie danych finansowych, medycynie oraz przetwarzaniu sygnałów.

Na koniec warto wspomnieć, że w swojej pracy wykorzystałem implementacje kopuł w pakiecie VineCopula (będącym rozwinięciem zarchiwizowanego już pakietu CDVine) języka R, por. [65].

Modelowanie zależności czasów infekcji

Wrócę teraz do reprezentacji rozkładu łącznego $\{Y_{v_1}, ..., Y_{v_{d_v}}\}$.
Funkcję przeżycia $\overline{H}_{d_v}$ możemy przedstawić w następującej postaci:

$$\overline{H}_{d_v}(x, ..., x) = C\Big(\overline{F}_1(x), ..., \overline{F}_{v_{d_v}}(x)\Big), \tag{2.5}$$

gdzie C to wybrana kopuła $(Y_{v_1}, ..., Y_{v_d})$.

Uwaga 2.12 (Prawdopodobieństwo infekcji węzła v). W pracy M. Xu, L. Hua [1] wyznaczono górne oszacowanie prawdopodobieństwa infekcji węzła v. Ponieważ szczegółowy dowód stwierdzenia 2.13 zawarty jest w wyżej wymienionej publikacji, pomijam go w swojej pracy.

Stwierdzenie 2.13. Jeśli dla $Y_1, ..., Y_{v_{d_v}}$ i dla każdego $d_v \geqslant 1$ zachodzi nierówność:

$$\mathbb{P}(Y_1 \leqslant t_1, ..., Y_{v_{d_v}} \leqslant t_d) \geqslant \mathbb{P}(Y_1 \leqslant t_1) \cdot ... \cdot \mathbb{P}(Y_{v_{d_v}} \leqslant t_d),$$

to granica górna dla prawdopodobieństwa infekcji węzła v dana jest wzorem:

$$p_v \leqslant \frac{\mathbb{E}[R_v]}{\mathbb{E}[R_v] + \sum_{d_v=0}^{\deg(v)} \mathbb{P}(D_v = d_v) \int_0^\infty \overline{F}^{d_v}(x)\overline{G}_v(x)dx}. \tag{2.6}$$

W praktyce, p_v możemy przybliżyć następującym wzorem:

$$p_v^* = \frac{\mathbb{E}[R_v]}{\mathbb{E}[R_v] + \int_0^\infty \overline{F}^{\sum_{j=1}^N a_{vj}p_j^*}(x)\overline{G}_v(x)dx}, \tag{2.7}$$

przypomnijmy, że $A = (a_{vu})$ to macierz sąsiedztwa grafu Γ, gdzie $a_{vu} = 1$, gdy $(u, v) \in E$ oraz $a_{vu} = 0$, gdy $(u, v) \notin E$.

Rozważę teraz przykłady dla różnych rozkładów czasu do infekcji.

1. **Wykładniczy:** W tym wypadku zakładamy, że procesy zainfekowania i regeneracji mają rozkład wykładniczy następującej postaci:

   - proces zainfekowania wewnątrz sieci: $\overline{F}(x) = e^{-\beta x}$, oraz $\overline{G}_v(x) = e^{-\epsilon_v x}$ z zewnątrz,
   - proces regeneracji: $\overline{S}_v(x) = \mathbb{P}(R_v > x) = e^{-\beta_v x}$.

   Wtedy:

   $$p_v^* = \frac{\beta \sum_{j=1}^N a_{vj}p_j^* + \varepsilon_v}{\delta_v + \beta \sum_{j=1}^N a_{vj}p_j^* + \varepsilon_v} \tag{2.8}$$

2. **Weibulla:** W tym wypadku zakładamy, że procesy zainfekowania i regeneracji mają rozkład Weibulla następującej postaci:

   - proces zainfekowania wewnątrz sieci: $\overline{F}(x) = e^{-(\beta x)^{\alpha_1}}$,
     oraz $\overline{G}_v(x) = e^{-(\epsilon_v x)^{\alpha_2}}$ z zewnątrz;
   - proces regeneracji: $\overline{S}_v(x) = \mathbb{P}(R_v > x) = e^{-(\delta_v x)^{\alpha_3}}$.

   Wtedy:

   $$p_v^* = \frac{\Gamma\left(1 + \frac{1}{\alpha_3}\right)}{\Gamma\left(1 + \frac{1}{\alpha_3}\right) + \delta_v \phi\left(\varepsilon_v, \beta, \alpha_1, \alpha_2, \mathbf{p}^*\right)}, \tag{2.9}$$

gdzie $\phi\Big(\varepsilon_v, \beta, \alpha_1, \alpha_2, \mathbf{p}^*\Big) = \int_0^\infty \exp\Big(- (\varepsilon_v^{\alpha_2} x^{\alpha_2} + \beta^{\alpha_1} x^{\alpha_1} \sum_{j=1}^N a_{vj} p_j^*)\Big) dx.$

3. **Log - Normalny:** W tym wypadku zakładamy, że procesy zainfekowania i regeneracji mają rozkład Log - Normalny następującej postaci:

   - proces zainfekowania wewnątrz sieci: $f_{vj}(x) = \frac{1}{x\sigma_1\sqrt{(2\pi)}} \exp\Big(- \frac{\ln(x)-\mu_1}{2\sigma_1^2}\Big)$, oraz $g(x) = \frac{1}{x\sigma_2\sqrt{(2\pi)}} \exp\Big(- \frac{\ln(x)-\mu_2}{2\sigma_2^2}\Big)$ z zewnątrz;
   - proces regeneracji: $S_v(x) = \mathbb{P}(R_v \leqslant x) = \Phi\Big(\frac{\ln(x)-\mu_v}{\sigma_v}\Big)$.

   Wtedy:

$$p_v^* = \frac{\exp(\mu_v + \sigma_v^2/2)}{\exp(\mu_v + \sigma_v^2/2) + \Psi(\mu_1, \mu_2, \sigma_1, \sigma_2, \mathbf{p}^*)}, \tag{2.10}$$

   gdzie

$$\Psi\Big(\mu_1, \mu_2, \sigma_1, \sigma_2, \mathbf{p}^*\Big) = \int_0^\infty \Big[1 - \Phi\Big(\frac{\ln(x)-\mu_2}{\sigma_2}\Big)\Big]\Big[1 - \Phi\Big(\frac{\ln(x)-\mu_1}{\sigma_1}\Big)\Big]^{\sum_{j=1}^N a_{vj} p_j^*} dx.$$

Zauważmy, że wybór rozkładu dla procesów zainfekowania i regeneracji węzła v wpływa na prawdopodobieństwo infekcji p_v. W powyższych przykładach rozważyliśmy tylko przypadek gdy oba procesy modelowane są taką samą rodziną rozkładów prawdopodobieństwa. Niemniej, można rozważyć również sytuację, gdy reprezentujemy je różnymi rozkładami.

2.3.2. Symulacja i wycena

Poprawna wycena ryzyka ma wielkie znaczenie ekonomiczne (patrz W. Otto [72]). Pozwala na prawidłowe realizowanie rachunku opłacalności podczas podejmowania decyzji, których wyniki prowadzą zazwyczaj do ograniczenia lub zwiększenia danych zagrożeń, ale też do stworzenia nowych. W tym sensie wiadomość o cenie ryzyka jest bardzo użyteczna zarówno dla tych, którzy na dane ryzyko są podatni, oraz tych, którzy w tej lub innej postaci realizują przekazanie tego ryzyka innym.

Kwestia wyceny ryzyka ma swoją specyfikę pod względem komponentów składki ubezpieczeniowej. Oprócz kosztów wypłacanych świadczeń i odszkodowań, musi ona pokrywać również inne wydatki. Składka ta zawiera też narzuty na koszty akwizycji, administracji lub likwidacji szkód. Narzuty te nie są typowe dla specyficznego produktu jakim jest polisa. Zasady wyliczania składek mogą się trochę różnić zależnie od sektora ubezpieczeń.

W poniższym podrozdziale omówię schemat (z ang. *framework*) wyceny ryzyka cyberbezpieczeństwa oparty na symulacji.

Załóżmy, że dla węzła v jego początkowa wartość (np. informacje, których utrata może nastąpić w wyniku awarii) wynosi w_v. Dalej, niech strata L_v ma rozkład Beta o gęstości:

$$f_{L_v}(x) = \frac{1}{w_v^{a+b-1}} \frac{1}{\mathrm{B}(a,b)} x^{a-1}(w_v - x)^{b-1}, \quad 0 \leqslant x \leqslant w_v,$$

gdzie $a, b > 0$ to parametry kształtu oraz $\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$, dla $\mathrm{Re}(a) > 0, \mathrm{Re}(b) > 0$ to funkcja Beta.

Niech funkcje kosztów będą zdefiniowane następującymi wzorami (patrz [1]):

$$\eta_v(l_v) = c \cdot l_v, \quad C_v(r_v) = c_1 w_v + c_2 r_v, \tag{2.11}$$

gdzie c oznacza wagę, z jaką straty spowodowane infekcją wchodzą do całkowitej wartości starty, c_1 - wagę dla wartości początkowej (z ang. *initial value*) oraz c_2 - wagę dla strat wynikających z regeneracji węzła.

Następnie rozważę dwa sposoby ustalania składek ubezpieczeniowych, jednakże zanim do tego przejdę, wprowadzę kilka pojęć (patrz R. Szekli [68]). Zacznę od tego, co rozumiemy pod pojęciem składki (jest to dokładna definicja z powyższej publikacji).

Definicja 2.14 (Składka). Składka to opłata, którą podmiot narażony na ryzyko płaci ubezpieczycielowi za przejęcie części ryzyka związanego ze swoją działalnością. Składkę wyznacza ubezpieczyciel opierając się na zasadzie, że losowe roszczenia klientów powinny być skompensowane przez ustalone opłaty (składki).
Niech $X = X_1 + ... + X_n$, gdzie $X_1, ..., X_n$ to n niezależnych, jednakowo rozłożonych ryzyk. Wtedy wartość składki H powinna zależeć od dystrybuanty F_X zmiennej X, która oznacza ubezpieczaną wielkość. Ozn.: $H = H(X)$.

Najbardziej podstawową składką jest $H = \mathbb{E}(X)$, którą nazywamy składką netto. Ponieważ wycena szkody przez wartość oczekiwaną nie zapewni bezpieczeństwa w pokryciu wielkości szkody, skupię się na innych sposobach ustalania składek.
Jedną ze składek z ustalonym poziomem bezpieczeństwa jest składka odchylenia standardowego. Równa się ona składce netto, jednak powiększona jest o czynnik zależny od odchylenia standardowego.

Definicja 2.15 (Składka odchylenia standardowego). Składkę odchylenia standardowego definiuje się następującym wzorem:

$$H(X) = \mathbb{E}(X) + \lambda\sqrt{\mathrm{Var}(X)}, \tag{2.12}$$

gdzie $\lambda > 0$ to narzut bezpieczeństwa (z ang. *safety loading).*

Kolejnym sposobem liczenia składki ubezpieczeniowej, którą wykorzystam w swojej pracy, jest metoda oparta o funkcję użyteczności. Najpierw jednak określę, czym jest funkcja użyteczności.

Definicja 2.16 (Funkcja użyteczności). Funkcją użyteczności u nazywamy dowolną niemalejącą funkcję, która mierzy użyteczność losowej kwoty Y.
Bardziej formalnie, funkcją użyteczności nazywamy funkcję $u : \mathbb{X} \subset \mathbb{R}_+^n \to \mathbb{R}$, taką że relacja $\succeq$ określona wzorem: $\forall_{\overline{x},\overline{y}\in\mathbb{X}}\ \overline{x} \succeq \overline{y} \iff u(\overline{x}) \geqslant u(\overline{y})$ jest relacją słabej preferencji.
Relację $\succeq$ określoną na przestrzeni towarów $\mathbb{X}$ nazywamy relacją słabej preferencji jeśli:
$\overline{x} \succeq \overline{y} \ \wedge\ \overline{y} \succeq \overline{z} \ \Rightarrow\ \overline{x} \succeq \overline{z}$ oraz $\overline{x} \succeq \overline{y} \ \vee\ \overline{y} \succeq \overline{x}$, dla każdych $\overline{x}, \overline{y}, \overline{z} \in \mathbb{X}$.

Funkcja użyteczności nie jest zdefiniowana jednoznacznie.

Przykład 2.17. Przykłady funkcji użyteczności:

- $u(x_1, x_2) = x_1 x_2$;
- $u(x_1, x_2) = x_1 x_2 + 5$;
- $u(x_1, x_2) = 5x_1 x_2 - 200$,

dla $x_1, x_2 \in \mathbb{R}^+$.

Zatem jeśli kontrahent posiada jakiś kapitał w i stosuje funkcję użyteczności u, to wielkość składki, którą skłonny jest zapłacić, zależy od tej funkcji u. To prowadzi do poniższej definicji.

Definicja 2.18 (Składka funkcji użyteczności). Składką funkcji użyteczności nazywamy składkę H wyliczaną z następującego równania:

$$u(w_v) = \mathbb{E}[u(w_v - X + H(X))], \tag{2.13}$$

gdzie u jest rosnącą, wklęsłą funkcją użyteczności, a w_v to wartość początkowa (np. informacje, których utrata może nastąpić w wyniku awarii).

Definicja 2.19. Mówimy, że decydent ma awersję do ryzyka, jeśli jego funkcja użyteczności jest wklęsła. Dla przypomnienia, funkcja u jest wklęsła, gdy: $u(w + h) - u(w) \geqslant u(w' + h) - u(w')$, dla każdego $h > 0$ i $w \leqslant w'$, tzn. gdy u ma przyrosty malejące.

W tym wypadku, korzystając z nierówności Jensena dla funkcji wklęsłych (tzn. $\mathbb{E}(u(X)) \leqslant u(\mathbb{E}(X))$), zachodzi $H \geqslant \mathbb{E}(X)$ (patrz [68]). Z tego wynika, że decydent, który ma awersję do ryzyka, skłonny jest do płacenia składki większej od wartości średniej ryzyka. Jednakże wielkość tej składki zależy od rodzaju funkcji u oraz kapitału w_v, który on posiada.

Autorzy analizowanego przeze mnie artykułu [1] rozpatrują funkcję użyteczności stałej relatywnej awersji do ryzyka (z ang. *constant relative risk averse utility function*), która pojawia się w literaturze dotyczącej cyberubezpieczeń i dana jest wzorem:

$$u(w) = \begin{cases} \frac{w^{1-\gamma}}{1-\gamma}, & \gamma \neq 1 > 0, \\ \log(w), & \gamma = 1, \end{cases}$$

gdzie γ to parametr oznaczający stopień awersji do ryzyka.

Modelowanie całej sieci wymaga:

1. dokładnej znajomości topologii sieci;
2. informacji na temat rozkładów procesu regeneracji i infekcji;
3. dostępu do bardzo szczegółowych danych.

Z tego względu jest to podejście bardzo trudne do odtworzenia i stanowi samo w sobie interesujące pytanie badawcze. Niestety nie mieści się ono w zakresie tej pracy, stąd pozostawiam je jako przyszły kierunek badań.

Przejdę teraz do przedstawienia algorytmu z publikacji [1], bazującego na 3000 symulacjach metodą Monte - Carlo.

Algorytm 1 pozwala nam w ciągu obowiązywania rocznej umowy śledzić zmiany stanu sieci i obliczać skumulowane straty dla każdego węzła v, w dowolnym czasie t.

Algorytm 1 Symulacja ryzyka cyberbezpieczeństwa dla umowy jednorocznej.

INPUT: Topologia sieci A, $T = 365$, rozkłady infekcji i regeneracji (wykładniczy, Weibulla lub Log - normalny), funkcje straty.

1: **for** $i = 1$ **to** 3000 **do**
2: **while** $t \leqslant T$ **do**
3: Wygeneruj losowe czasy regeneracji $r_1, ..., r_m$ zgodnie z rozkładem infekcji, gdzie m to liczba zainfekowanych węzłów w czasie t;
4: Dla każdego bezpiecznego węzła v losowo wygeneruj czasy infekcji $y_1, ..., y_{d_v}, z_v$, gdzie d_v to liczba zainfekowanych sąsiadów węzła v, oraz z_v to czas samoinfekcji (infekcji spowodowanej zagrożeniem z zewnątrz sieci);
5: Określ, które zdarzenie wystąpiło jako pierwsze, tzn. $t_1 = \min\{r_1, ..., r_m, y_1, ..., y_{d_v}, z_v\}$;
6: **if** Infekcja nastąpiła **then**
7: Zmień stan węzła z 0 na 1 i oblicz stratę;
8: **else**
9: Zmień stan węzła z 1 na 0 i oblicz stratę;
10: **end if**
11: $t \leftarrow t + t_1$
12: **return** t, stan węzłów oraz skumulowane straty dla każdego węzła do czasu t
13: **end while**
14: **end for**

OUTPUT: Stan sieci i skumulowane straty dla każdego węzła w każdym przypadku infekcji lub regeneracji.

3. Analizy empiryczne

W niniejszym rozdziale przeprowadzę empiryczną analizę omówionych przeze mnie sposobów modelowania cyberubezpieczeń na przykładzie danych rzeczywistych.

3.1. Dane rzeczywiste

Dane, które wykorzystałem do analizy w swojej pracy, pobrałem ze strony Privacy Rights Clearinghouse `https://privacyrights.org/data-breaches` [25] (w dalszej części pracy będę odnosił się do tego zbioru przy użyciu aliasu PRC). PRC to organizacja non-profit, której misją jest angażowanie, edukowanie i upoważnianie osób do ochrony ich prywatności. Dane dotyczą zdarzeń np. oszustw kredytowych, zhakowań, kradzieży danych itp. dla firm z najważniejszych sektorów biznesowych, które występowały w USA od 2005 roku. Zwróćmy uwagę, że ten zbiór danych jest na bieżąco aktualizowany. Miałem zatem dostęp do znacznie większej liczby rekordów niż autorzy w analizowanych przeze mnie publikacjach, które ukazały się kilka lat wcześniej. Warto także zaznaczyć, że nie mam dostępu do danych archiwalnych[1]. Dane do analizy w swojej pracy pobrałem w dniu 05.05.2021 r.

Tabela 3.1 zawiera podsumowanie zbioru danych. Poniżej przedstawiam szczegółowy opis analizowanego zbioru.

[1]Ze względu na aktualizacje zbioru danych oraz brak dostępu do danych archiwalnych, wyniki przeprowadzonych przeze mnie analiz nie muszą być dokładnie takie same jak w analizowanych publikacjach.

Zmienna	**Typ**[2]	**Znaczenie**
Date.Made.Public	character	data zgłoszenia naruszenia
Company	character	nazwa poszkodowanej firmy
City	character	miasto w USA, w którym doszło do naruszenia
State	character	stan w USA, w którym doszło do naruszenia
Type.of.breach	character	typy naruszeń
Type.of.organization	character	typy organizacji
Total.Records	character	rozmiar naruszenia (ang. *breach size*)
Description.of.incident	character	krótki opis zdarzenia
Information.Source	character	źródło
Source.URL	character	źródło internetowe
Year.of.Breach	integer	rok zgłoszenia naruszenia
Latitude	double	szerokość geograficzna
Longitude	double	długość geograficzna

Tablica 3.1: Podsumowanie analizowanego zbioru danych.

Do najważniejszych zmiennych zaliczamy:

- *Date.made.public* - data zgłoszenia naruszenia (breach) organowi rządowemu lub mediom;
- *Total.Records* - rozmiar naruszenia, czyli liczba / oszacowanie ilości utraconych danych, przez tę liczbę mierzone są szkody;
- *Company* – nazwa poszkodowanej firmy.

Dodatkowo mamy jeszcze podział danych na typy naruszeń i typy organizacji. Dokładniej:

- Typy naruszeń:
  - CARD: oszustwa związane z kartami debetowymi i kredytowymi, które nie są dokonywane przez włamanie (np. skimming);

[2]umieszczony typ zmiennej zgodny z notacją w języku R.

  - DISC: niezamierzone ujawnienie poufnych informacji na stronie lub przesłanie danych do niewłaściwych osób np. przez e-maila;
  - HACK: zhakowanie lub złośliwe oprogramowanie (np. oprogramowanie szpiegujące czyli spyware);
  - INSD: ktoś z uprawnionym dostępem, np. pracownik, celowo narusza informacje lub dane;
  - PHYS: zgubione lub skradzione dane nieelektroniczne, takie jak dokumenty papierowe;
  - PORT: zgubiony lub skradziony laptop, smartfon, karta SD, itp.;
  - STAT: zgubione lub skradzione stacjonarne urządzenie elektroniczne np. komputer lub serwer;
  - UNKN: nieznany typ naruszeń.

- Typy organizacji:
  - BSF: firmy – usługi finansowe i ubezpieczeniowe;
  - BSO: firmy – inne;
  - BSR: firmy – handel detaliczny;
  - EDU: instytucja naukowa;
  - GOV: rząd i wojsko;
  - MED: opieka zdrowotna – dostawcy usług medycznych;
  - NGO: organizacje non - profit;
  - UNKN: nieznany typ organizacji (brak rekordów).

Należy zaznaczyć, że analizowany zbiór danych może nie zawierać wszystkich danych dotyczących włamań, gdyż np. nie zostały one upublicznione. Dodatkowo, daty odpowiadające incydentom dotyczą terminu upublicznienia danych, a niekoniecznie terminów, w których te incydenty naprawdę nastąpiły. Jest to jednak jeden z największych publicznie dostępnych zbiorów danych (na podstawie źródeł umieszczonych w literaturze). Co więcej, jest on często wykorzystywany w publikacjach dotyczących cyberbezpieczeństwa oraz cyber ryzyka.

Przygotowanie i czyszczenie zbioru PRC

W celu przygotowania zbioru danych do analizy, wykorzystałem podejście przedstawione w analizowanych przeze mnie publikacjach. Dokładniej, proces przygotowania obejmuje:

- oczyszczenie danych tzn. usunięcie rekordów z brakami danych (czyli wartości 0 i NA dla *breach sizes*);
- wybranie zakresu analizowanych danych;
- przekształcenie zmiennych zgodnie z opisem zawartym w literaturze (np. losowe porządkowanie incydentów odpowiadających temu samemu dniu, a następnie wstawienie małego, losowego przedziału czasowego pomiędzy dwa kolejne incydenty, zapewniając jednocześnie, że incydenty te odpowiadają nadal temu samemu dniu por. [24]);

Na przygotowanych według powyższego opisu danych przeprowadzę analizę empiryczną, której celem jest między innymi weryfikacja sposobów modelowania zaproponowanych w literaturze oraz sprawdzenie, jak dodanie nowych lub uzupełnienie przeszłych (wcześniej nie znanych) naruszeń wpłynęło na modelowanie strat.

3.2. Metodologia badań

W niniejszym rozdziale opiszę metody, testy oraz kryteria wykorzystane do analizy w mojej pracy. Zacznę od przedstawienia teorii dotyczącej modelowania rozkładami, a w kolejnym podrozdziale opiszę teorię dotyczącą modelowania procesami.

3.2.1. Badanie zgodności rozkładów

Poniżej przedstawię podstawowe testy wykorzystane w celu weryfikacji zgodności rozkładu. W dalszej części będę korzystał z notacji przedstawionej w Koronacki, Mielniczuk [26].
Niech $(X_1, ..., X_n)$ będzie próbką i.i.d. (tzn. niezależne zmienne losowe o takim samym rozkładzie) z nieznanego rozkładu o dystrybuancie $F(x)$.

Test Kołmogorowa - Smirnowa

Test Kołmogorowa - Smirnowa (patrz [26, 69]) jest testem zgodności, który służy do sprawdzenia hipotezy, czy nieznana, ale ciągła dystrybuanta zmiennej losowej jest równa pewnej zakładanej dystrybuancie ciągłej. Stosuje się go do testowania hipotezy $H_0 : F(x) = F_0(x)$ przy hipotezie alternatywnej $H_1 : F(x) \neq F_0(x)$, gdzie F_0 to dystrybuanta teoretyczna, natomiast dystrybuanta empiryczna $F_n(x)$ wyznaczana jest z następującego wzoru:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i \leqslant x)}.$$

Statystyka testowa tego testu jest równa:

$$T_{KS} = \sup_{x \in \mathbb{R}} \{|F_n(x) - F_0(x)|\},$$

Małe wartości statystyki T_{KS} przemawiają za przyjęciem hipotezy zerowej.
W języku R test ten realizuje funkcja ks.test() z biblioteki stats.

Test Kołmogorowa - Smirnowa z poprawką Lillieforsa [56, 57]

Niestety, gdy dystrybuanta teoretyczna F_0 zależy od nieznanych parametrów, tzn. gdy pewne parametry rozkładu muszą być oszacowane na podstawie próby, to test Kołmogorowa - Smirnowa nie ma już zastosowania, gdyż rozkład statystyki testowej zależy od F_0. W takim wypadku należy stosować test K-S z poprawką Lillieforsa. Statystyka testowa tego testu jest równa:

$$T_{KS}^{L} = \sup_{x \in \mathbb{R}} \{|F^*(x) - F_n(x)|\},$$

gdzie $F_n(x)$ to odpowiednio skorygowana dystrybuanta empiryczna, natomiast $F^*(x)$ to dystrybuanta teoretyczna rozkładu:

(a) normalnego [56] ze średnią próbkową $\overline{X} = \mu$ i wariancją próbkową $s^2 = \sigma^2$;

(b) wykładniczego [57] ze średnią próbkową $\overline{X} = \frac{1}{\lambda}$;

(c) Weibulla [28, 58] ze średnią próbkową $\overline{X} = \lambda\Gamma(1 + 1/k)$ i wariancją próbkową $s^2 = \lambda^2\Big(\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2\Big)$;

(d) log - normalnego [28] ze średnią próbkową $\overline{X} = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ i wariancją próbkową $s^2 = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$.

Jeżeli wartość statystyki T_{KS}^{L} przekracza wartość krytyczną podaną w tabeli wartości krytycznych dla testu Lillieforsa (patrz Lilliefors [56, 57]), wtedy odrzuca się hipotezę, że obserwacje pochodzą z populacji (a) normalnej / (b) wykładniczej / (c) Weibulla / (d) log - normalnej. Wówczas wartości krytyczne dla statystyki T_{KS}^{L} uzyskuje się w wyniku obliczeń metodą Monte - Carlo.
W języku R test ten realizuje funkcja LcKS() z biblioteki KScorrect.

Test Kołmogorowa - Smirnowa w wersji dyskretnej

Celem testu Kołmogorowa - Smirnowa w wersji dyskretnej jest sprawdzenie, czy próbka $X = \{X_1, X_2, ..., X_n\}$ pochodzi z rozkładu dyskretnego o dystrybuancie $F_\theta(x)$, gdzie θ jest nieznanym parametrem. Autorzy artykułu [73] proponują wykorzystanie metody bootstrapowej Monte - Carlo w celu estymacji p - wartości dla tego testu. Poniższy algorytm przedstawia szczegóły:

Algorytm 2 Estymacja p - wartości testu Kołmogorowa - Smirnowa w wersji dyskretnej.

1: Dla $b = 1, ..., B$ wylosuj próbkę bootstrapowaną: $X^{(b)} = \{X_1^{(b)}, ..., X_n^{(b)}\}$.

2: Dla każdego b oblicz estymator największej wiarygodności (MLE) parametru θ, tj. $\hat{\theta}^{(b)}$.

3: Wygeneruj próbkę $X^{(c)} = \{X_1^{(c)}, ..., X_n^{(c)}\}$ taką, że $X^{(c)}$ jest *i.i.d.* z rozkładu $F_{\hat{\theta}^{(b)}}$ (tzn. z rozkładu $F_\theta(x)$ z parametrem $\theta = \hat{\theta}^{(b)}$).

4: Oblicz statystykę testu K-S: $T_{KS}^{(b)} = \sup_{x\in\mathbb{R}}\{|\widehat{F}_n^{(c)}(x) - F_{\hat{\theta}^{(b)}}(x)|\}$, gdzie $\widehat{F}_n^{(c)}(x)$ jest dystrybuantą empiryczną $X^{(c)}$.

5: Oszacuj p - wartość ze wzoru: $\frac{\#\{b:T_{KS}^{(b)} > T_{KS}\}+1}{B+1}$, gdzie $T_{KS} = \sup_{x\in\mathbb{R}}\{|\widehat{F}_n(x) - F_{\hat{\theta}}(x)|\}$ jest statystyką testową testu K-S opartą na danych oryginalnych z oszacowaniem parametru θ metodą MLE.

Warto wspomnieć, że w 5. kroku do licznika i mianownika dodawana jest liczba 1, aby uniknąć zerowej p - wartości. W języku R powyższy algorytm realizuje metoda dis.kstest() z biblioteki iZID.

Test Andersona - Darlinga

Test Andersona - Darlinga (patrz Grace, Wood [27]) stosuje się do testowania hipotezy $H_0 : F = F_0$ przy hipotezie alternatywnej $H_1 : F \neq F_0$, gdzie F to ciągły rozkład zmiennych z próby *i.i.d.*, a F_0 to ciągły rozkład teoretyczny. Statystyka testowa tego testu jest równa:

$$A_n = n\int_{-\infty}^{+\infty} \frac{[F_n(x) - F_0(x)]^2}{F_0(x)(1 - F_0(x))} dF_0(x),$$

gdzie F_n jest dystrybuantą empiryczną z danych, a n to liczność próby. Przypomnijmy, że rozkład A_n jest zależny od F_0, (w przeciwieństwie do testu Kołmogorowa – Smirnowa).

Analogicznie do testu K-S, gdy pewne parametry rozkładu muszą być oszacowane na podstawie prób, to tutaj też trzeba stosować odpowiednią poprawkę. W języku R jest on realizowany przez funkcję ad.test() z biblioteki goftest (zarówno w wersji podstawowej, jak i z poprawką). W teście, w celu skorygowania wpływu estymacji parametrów, została zastosowana metoda Brauna [70]. Metoda ta powinna być wykorzystywana na zbiorach danych, które łatwo można podzielić na kilka mniejszych podzbiorów. Zatem liczba obserwacji w analizowanym wektorze danych powinna być odpowiednio duża. Procedura polega na wykorzystaniu całej próbki do oszacowania parametrów zakłócających (z ang. *nuisance parameters*[3]), a następnie przeprowadzeniu odpowiedniego przekształcenia danych z wykorzystaniem oszacowanej dystrybuanty. Potem oblicza się statystykę zgodności dopasowania, oddzielnie dla każdego podzbioru przekształconych

[3]Parametr zakłócający to każdy parametr, który nie jest obiektem bezpośredniego zainteresowania, ale musi być uwzględniony w analizie parametrów będących przedmiotem zainteresowania. Przykładem takiego parametru jest wariancja rozkładu normalnego, gdy głównym obiektem zainteresowania jest średnia.

obserwacji. Hipoteza zerowa jest odrzucana, jeśli którakolwiek ze statystyk przekracza swoje wartości krytyczne. Szczegóły implementacyjne są zawarte na stronie dokumentacji testu lub w publikacji H. Braun [70].

Test Shapiro - Wilka

Test Shapiro - Wilka (patrz Koronacki, Mielniczuk [26] i Shapiro, Wilk [41]) służy do testowania normalności rozkładu badanej próby na podstawie charakterystyk rozkładu bez konieczności specyfikacji parametrów μ oraz σ. Jest on testem uniwersalnym[4].
Bardziej formalnie testuje on hipotezę zerową, mówiącą, że próba $x_1, ..., x_n$ pochodzi z rozkładu normalnego, przeciw hipotezie alternatywnej, mówiącej, że z niego nie pochodzi. Statystyka testowa dla testu Shapiro - Wilka ma następującą postać:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

gdzie $\overline{x} = \frac{x_1 + ... + x_n}{n}$ to średnia próbkowa, natomiast a_i to współczynniki zależne od rozmiaru próby n jak i od i.
W języku R test ten realizuje funkcja shapiro.test() z biblioteki stats.

Test Chi - kwadrat

Test Chi - kwadrat (patrz Shier [39]) używany jest do testowania hipotezy, która mówi, że obserwowane dane mają określony rozkład. Procedura testowa polega tu na uporządkowaniu n obserwacji z próby w tabelę częstości z k klasami. Statystyka testu chi - kwadrat ma postać:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

gdzie O_i to wartość mierzona, a E_i to wartość teoretyczna (oczekiwana) wynikająca z hipotezy, odpowiadająca wartości mierzonej. Liczba stopni swobody wynosi $k-p-1$, gdzie p jest liczbą parametrów oszacowanych na podstawie danych (z próbki) użytych do wygenerowania zakładanego rozkładu.
Test Chi - kwadrat może być stosowany dla każdego rozkładu, tylko nieco inaczej dla ciągłego.
W języku R test ten realizuje np. funkcja chisq() z biblioteki stats.

[4] Test Shapiro - Wilka charakteryzuje się dużą mocą, gdy prawdziwy rozkład jest wyraźnie skośny lub spłaszczony i w przybliżeniu symetryczny.

3.2.2. Metody oceny i selekcji modelu

Przejdę teraz do omówienia podstawowych testów pozwalających na weryfikację jakości dopasowania w przypadku modelowania procesów stochastycznych. Do weryfikacji założeń wybranych modeli wykorzystałem testy Kołmogorowa - Smirnowa, Andersona - Darlinga, oraz opisane poniżej.

Test Craméra-von - Misesa

Hipotezy testu Craméra-von – Misesa mają analogiczną postać jak w testach Andersona - Darlinga oraz Kołmogorowa - Smirnowa. Niech $X_1, ..., X_n$ będą niezależnymi zmiennymi losowymi o takim samym rozkładzie i dystrybuancie empirycznej $F_n(x),\ x \in \mathbb{R}$. Statystyka testowa Craméra - von Misesa (patrz Csorgo, Faraway [38]) dana jest wzorem:

$$\omega_n^2 = n \int_{-\infty}^{\infty} \Big(F_n(x) - F(x)\Big)^2 dF(x) = n \int_0^1 \Big(G_n(s) - s\Big)^2 ds, \tag{3.1}$$

gdzie $G_n(s)$ jest dystrybuantą empiryczną niezależnych zmiennych losowych $U_1 = F(X_1), ..., U_n = F(X_n)$ o rozkładach jednostajnych na odcinku $[0, 1]$, $F(x)$ jest dystrybuantą rozkładu wzorcowego, a n to liczność próby.

Jeśli ponadto przez $U_{1,n} \leqslant ... \leqslant U_{n,n}$ oznaczymy statystyki pozycyjne o rozkładach jednostajnych na odcinku $[0, 1]$, to prawą stronę równania (3.1) możemy zapisać w postaci:

$$\omega_n^2 = \frac{1}{12n} + \sum_{k=1}^{n} \Big(U_{k,n} - \frac{2k-1}{2n}\Big)^2, \quad n = 1, 2, ...,$$

Wartość statystyki testowej ω_n^2 w praktyce oblicza się z podanego powyżej wzoru.
W języku R test ten realizuje funkcja cvm.test() z biblioteki goftest.

Test Ljunga - Boxa

Test Ljunga - Boxa (patrz Widz, Bar [37]) służy do wykrywania autokorelacji szeregów czasowych. W praktyce jest to najczęściej stosowany test, który pozwala na weryfikację hipotezy o braku korelacji dowolnego rzędu. Stosuje się go do testowania hipotezy H_0 : autokorelacje wszystkich rzędów równają się zero, przy alternatywie H_1 : istnieje przynajmniej jedna autokorelacja różna od zera.

Statystyka Ljunga - Boxa ma postać:

$$Q_{LB} = n(n+2)\Big(\sum_{j=1}^{h} \frac{\rho^2(j)}{n-j}\Big),$$

gdzie n - liczba obserwacji, $\rho(j)$ - współczynniki autokorelacji rzędu j, $j = 1, 2, ..., h$ - rząd autokorelacji.

Statystyka Ljunga - Boxa Q_{LB} posiada rozkład zbieżny do rozkładu χ^2 o p stopniach swobody. W języku R test ten realizuje funkcja Box.test() z ustalonym parametrem type = "Ljung-Box", z biblioteki stats.

Test McLeoda - Li

Test McLeoda - Li (patrz [71]) opiera się na tej samej statystyce, którą stosuje się w teście Ljung-Boxa, z tym że próbkowe współczynniki autokorelacji z danych zastępuje się próbkowymi współczynnikami autokorelacji z danych podniesionych do kwadratu $\rho_{WW}(k)$. Statystyka McLeoda - Li przyjmuje postać:

$$Q_{ML} = n(n+2)\Big(\sum_{k=1}^{h} \frac{\rho^2_{WW}(k)}{n-k}\Big),$$

Koncepcja tego testu polega na tym (patrz [26]), że czasami nieliniowa zależność między zmiennymi losowymi ujawnia się, gdy zastosuje się na nich pewne (nieliniowe) przekształcenie. W języku R test ten realizuje funkcja McLeod.Li.test() z biblioteki TSA.

Selekcja modelu (kryteria informacyjne)

Przejdę teraz do omówienia metod wyboru rzędu procesu ARMA(p, q), które oparte są na podejściu kryterialnym do wyboru odpowiedniego modelu (patrz Mielniczuk [40]).
Załóżmy, że chcemy dopasować model ARMA(p, q) do danych i musimy zdecydować, jakie wybrać p i q. Jeśli chcielibyśmy postępować naiwnie patrząc na $(p, q) = \arg\min\{-\log L(\hat{\theta})\}$, to zawsze będzie to prowadziło do wyboru maksymalnych rozpatrywanych p i q. Wiąże się to z przeszacowaniem, ponieważ dane zostały już raz użyte do oszacowania parametrów, a chcielibyśmy użyć ich ponownie do oszacowania wymiaru modelu. Sposobem uniknięcia tego efektu jest włączenie do funkcji kryterialnej kary za złożoność tj. kosztu dopasowania modelu zawierającego wiele parametrów. Wśród najczęściej stosowanych kar wyróżniamy:

1. Kryterium Akaike (AIC):

$$AIC = -2\ln L(\hat{\theta}) + 2(p+q+1),$$

   gdzie $L(\hat{\theta})$ to funkcja wiarygodności dla danego modelu, która jest funkcją liczby parametrów modelu i liczby obserwacji n oraz $(p+q+1)$ to liczba parametrów modelu.

   W języku R tę funkcję realizuje metoda AIC() z biblioteki stats.

2. Kryterium Bayesowskie (BIC):

$$BIC = -2\ln L(\hat{\theta}) + \ln n(p+q+1),$$

gdzie $L(\hat{\theta})$ i $(p+q+1)$ - jak dla AIC, i n - liczba obserwacji.

Dla n takiego, że $\ln n > 2$ kryterium BIC karze mocniej niż AIC i prowadzi do oszczędniejszych modeli.

W języku R tę funkcję realizuje metoda BIC() z biblioteki stats.

3. Kryterium Hannana-Quinna (HQIC):

$$HQIC = -2\ln L(\hat{\theta}) + 2(p+q+1)\ln(\ln n),$$

gdzie $L(\hat{\theta})$ i $(p+q+1)$ - jak dla AIC, i n - liczba obserwacji.

W języku R tę funkcję realizuje metoda HQIC() z biblioteki qpcR.

Stosując kryteria informacyjne do wyboru modelu wybieramy ten model, któremu odpowiada minimalna wartość danego kryterium informacyjnego. AIC i BIC są najczęściej stosowanymi kryteriami przy wyborze modelu w statystyce.

Funkcja Autokorelacji (ACF) i częściowej autokorelacji (PACF)

Zanim podam formalne definicje funkcji autokorelacji (w skrócie ACF) oraz funkcji częściowej autokorelacji (w skrócie PACF) (patrz J. Mielniczuk [40]), najpierw przedstawię definicje pomocnicze.

Definicja 3.1 (proces stacjonarny w szerszym sensie). Szereg czasowy $(X_t)_{t\in\mathbb{Z}}$ nazywamy procesem stacjonarnym w szerszym sensie jeśli:

(a) $\mathbb{E}(X_t) = m$, $t \in \mathbb{Z}$,

(b) $\mathrm{Var}(X_t) < \infty$, $t \in \mathbb{Z}$,

(c) $\gamma_X(s,t) = \gamma_X(s+r, t+r)$, $r,s,t \in \mathbb{Z}$, gdzie

$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t)$ to funkcja autokowariancji procesu X

Definicja 3.2 (biały szum). Mówimy, że ciąg $(\varepsilon_t)_{t\in\mathbb{Z}}$ jest (słabym) białym szumem jeśli (ε_t) jest nieskorelowanym ciągiem zmiennych losowych takich, że $\mathbb{E}(\varepsilon_t) = 1$ oraz $\mathrm{Var}(\varepsilon_t) = \sigma^2$ $\forall_t$. Oznaczamy go przez $\mathrm{WN}(0, \sigma^2)$.

Definicja 3.3 (ACF). Niech $(X_t)_{t\in\mathbb{Z}}$ będzie procesem stacjonarnym w szerszym sensie. Wtedy funkcja autokorelacji (ACF) procesu X dana jest wzorem:

$$\rho_X(h) = \rho(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\gamma_X(0)}, \quad \rho_X(0) = 1,$$

gdzie $\gamma_X(h) = \gamma_X(s,t)$, dla $s - t = h$.

Autokorelacja pozwala na ocenę stopnia zależności wyrazu szeregu od wyrazów poprzednich. Dokładniej, mierzy korelację pomiędzy obserwacjami we wcześniejszych okresach i obserwacjami w późniejszych okresach bez pomijania obserwacji pomiędzy nimi.
W języku R tę funkcję realizuje metoda acf() z biblioteki stats.

Definicja 3.4 (PACF). Funkcja częściowej autokorelacji pokazuje zależność między X_t, a wartościami X_{t+h} z usuniętą zależnością liniową między X_{t+1}, X_{t+2}, $\ldots$, X_{t+h-1} (pomijamy wpływ korelacji zmiennych pośrednich).
W języku R tę funkcję realizuje metoda pacf() z biblioteki stats.

Dzięki funkcjom ACF i PACF można określić rząd procesu ARMA(p, q). Dokładniej, urywająca się funkcja ACF wskazuje na rząd procesu MA(q), a funkcja PACF na rząd procesu AR(p). W praktyce wykresy tych funkcji są skomplikowane do jednoznacznej interpretacji. Ogólnie, celem jest wybór modelu o jak najmniejszej liczbie parametrów.

Współczynnik korelacji τ Kendalla

Korelacja τ Kendalla (patrz [43]) mierzy zależność jako tendencję ruchu dwóch zmiennych losowych X i Y w tym samym (lub przeciwnym) kierunku. Załóżmy że (X_i, Y_i) i (X_j, Y_j) to dwie pary obserwacji z rozkładu łącznego (X, Y). Jeśli $(X_j - X_i)$ i $(Y_j - Y_i)$ mają ten sam znak, to para jest zgodna, a jeśli mają przeciwne znaki, to niezgodna. Wtedy korelacją τ Kendalla nazywamy:

$$\tau(X, Y) = \mathbb{P}[(X_j - X_i)(Y_j - Y_i) > 0] - \mathbb{P}[(X_j - X_i)(Y_j - Y_i) < 0].$$

Wyrazić go można także przy pomocy kopuł, które wykorzystuję w dalszej części pracy:

$$\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

Nieparametryczny test oparty na τ Kendalla

Współczynnik korelacji Kendalla jest często stosowany jako statystyka testowa w celu ustalenia, czy dwie zmienne można uznać za statystycznie zależne. Mówimy, że test ten jest nieparametryczny, gdyż nie opiera się na żadnych założeniach dotyczących rozkładów zmiennych losowych X lub Y, lub rozkładu pary (X, Y).
Formalnie, jest to test hipotezy $H_0 : \rho = 0$, przeciwko $H_1 : \rho \neq 0$. Natomiast jego statystyka testowa ma następującą postać (patrz [76]):

$$T = \sqrt{\frac{9n(n-1)}{2(2n+5)}} \cdot |\hat{\tau}|,$$

gdzie n oznacza liczbę obserwacji, a $\hat{\tau}$ oznacza empiryczny wskaźnik korelacji τ Kendalla.
Wtedy p - wartość hipotezy zerowej o niezależności dwuwymiarowej jest równa:

$$p = 2 \cdot (1 - \Phi(T)),$$

gdzie $\Phi(\cdot)$ jest dystrybuantą standardowego rozkładu normalnego.
W języku R test ten realizuje funkcja cor.test() z ustalonym parametrem method = "kendall" w bibliotece stats.

Współczynnik korelacji ρ Spearmana

Korelacją ρ Spearmana (patrz [43]) nazywamy:

$$\rho^S(X, Y) = \rho(F(X), F(Y)),$$

gdzie $\rho(X, Y) = \mathrm{Cov}(X, Y)/\sqrt{\mathrm{Var}\, X \,\mathrm{Var}\, Y}$, a $F(\cdot)$ oznacza dystrybuantę wybranej zmiennej losowej. Wzór ten można także wyrazić za pomocą kopuł następującym wzorem:

$$\rho^S(X, Y) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2,$$

gdzie $C(\cdot, \cdot)$ oznacza wybraną kopułę, używaną do modelowania zależności.

Nieparametryczny test oparty na ρ Spearmana

Podobnie jak nieparametryczny test oparty na współczynniku τ Kendalla, test oparty na ρ Spearmana stosuje się jako statystykę testową w celu ustalenia, czy dwie zmienne można uznać za statystycznie zależne. Analogicznie testuje on hipotezę $H_0 : \rho = 0$, przeciwko $H_1 : \rho \neq 0$.
Jego statystyka testowa przyjmuje postać (patrz [76]):

$$T = \sqrt{n - 1} \cdot |\widehat{\rho^S}|,$$

gdzie n to liczba obserwacji, a $\widehat{\rho^S}$ to empiryczny wskaźnik korelacji ρ Spearmana.
W języku R test ten realizuje funkcja cor.test() z ustalonym parametrem method = "spearman" w bibliotece stats.

3.3. Wyniki

3.3.1. Analiza zgodności rozkładów

Przejdę teraz do wyników dotyczących badania zgodności rozkładów zmiennych *breach size* oraz *interarrival times*, por. publikację Eling i Loperfido [23] opisaną w skrócie w rozdz. 2.1.

Autorzy korzystali z danych pobranych ze wspomnianej strony PRC [25] z okresu od 10 stycznia 2005r. do 15 grudnia 2015r., co dało im dokładnie 2266 obserwacji.

Przypomnę, że przeprowadzona przez autorów analiza MDS pokazywała, iż różne typy naruszeń danych muszą być modelowane jako odrębne kategorie ryzyka. W celu testowania zgodności rozkładów zdecydowali się wybrać tylko typ naruszenia PORT. Ich wybór padł właśnie na ten typ dlatego, że ma on najwięcej, tj. 629 obserwacji. Eling i Loperfido wspominają, że przeprowadzili również analizę na innych typach naruszeń, jednakże ze względu na ograniczoną długość publikacji wybrali tylko wspomniany powyżej typ.

Zacznę od przedstawienia wyników analizy przeprowadzonej na zmiennej *breach sizes*.

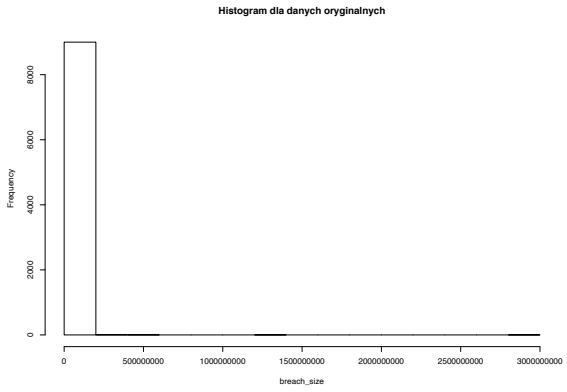
Zmienna *breach sizes*

Na początku przedstawię w skrócie podstawowe statystyki i wykresy wprowadzające, które wykorzystam do zaprezentowania interesujących wniosków w następującej części tego rozdziału. Przyjrzyjmy się tablicy 3.2.

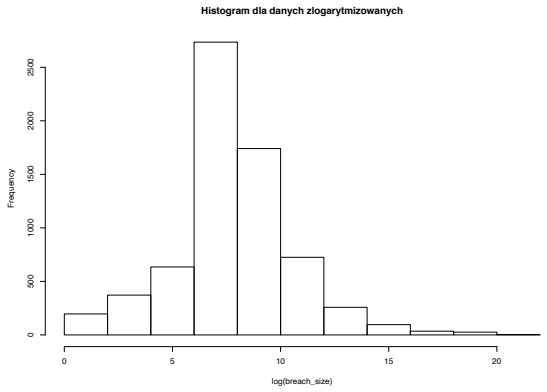
0%	**10**%	**20**%	**30**%
1	91	500	754.6
40%	**50**%	**60**%	**70**%
1200	2000	3385.6	6923
80%	**90**%	**100**%	
15500	63000	3000000000	

Tablica 3.2: Tabela przedstawiająca wartości kolejnych decyli dla danych oryginalnych.

Tablica 3.2 przedstawia wartości kolejnych decyli (tj. kwantyli rzędu $k/10$ dla $k = 1, ..., 9$) dla całego zbioru danych oryginalnych. Z tabelki możemy odczytać, że 50% (mediana) obserwacji jest mniejszych od 2000, 70% obserwacji jest mniejszych niż 6923 oraz 90% obserwacji jest mniejszych od 63000.



Rysunek 3.1: Histogram dla danych oryginalnych.



Rysunek 3.2: Histogram dla danych zlogarytmowanych.

Na rysunkach 3.1 oraz 3.2 widzimy histogramy dla całego zbioru danych: oryginalnych oraz w skali logarytmicznej. Patrząc na pierwszy histogram (rysunek 3.1) widzimy, że znacząca większość obserwacji ma małe wielkości w porównaniu do maksimum równego $3000000000 = 3e^{-9}$. Na rysunku 3.2 przejrzystość przedstawionych danych pokazanych w skali logarytmicznej jest znacznie większa. Nie widać tu już tak ogromnego rozrzutu. Z tego więc powodu zasadnym wydaje się próba wzięcia logarytmu z danych. Dodatkowo widzimy, że rozkład może być zbliżony do rozkładu normalnego lub T-studenta. Mimo wszystko obserwujemy delikatną skośność prawostronną, co sugerowałoby próby dopasowania rozkładu skośnego - normalnego lub skośnego - T-studenta.

Zanim zaprezentuję otrzymane przeze mnie wyniki empiryczne, w celu porównania przytoczę i omówię rezultaty analizy przedstawionej w publikacji [23].

Model	**Kolmogorov - Smirnov Test**		**Anderson - Darling Test**	
Panel A: Original Data	test stat.	p - value	test stat.	p - value
Exponential	0.66	< 0.01	1,083.10	< 0.01
Gamma	no covergence			
GPD	0.06	< 0.05	4.12	< 0.01
Log-logistic	1.00	< 0.01	0.55	
Log-normal	**0.03**		**0.27**	
Normal	0.47	< 0.01	218.74	< 0.01
Weibull	0.08	< 0.01	9.51	< 0.01
Skew-normal	0.86	< 0.01	1,532.70	< 0.01
Skew-student	0.94	< 0.01	1,141.60	< 0.01
Panel B: Log-transformed Data (LN)	test stat.	p - value	test stat.	p - value
Exponential	0.30	< 0.01	103.75	< 0.01
Gamma	0.14	< 0.01	22.55	< 0.01
GPD	0.24	< 0.01	63.81	< 0.01
Log-logistic	0.99	< 0.01	11.57	< 0.01
Log-normal	0.28	< 0.01	110.02	< 0.01
Normal	0.03		0.28	
Weibull	0.08	< 0.01	4.90	< 0.01
Skew-normal	**0.02**		**0.26**	
Skew-student	0.02		0.25	

Rysunek 3.3: Analiza zgodności rozkładów dla rozmiarów naruszeń, typ PORT, 629 obserwacji (źródło: [23]).

Na rys. 3.3 przedstawiona jest tabela z wynikami dla badania zgodności rozkładów w przypadku naruszenia PORT. W pierwszej kolumnie widzimy analizowane rozkłady. W drugiej kolumnie - kolejne liczby oznaczają wartości statystyki testowej. Następnie widzimy p - wartości dla testu Kołmogorowa - Smirnowa. Autorzy zdecydowali się ustalić poziom ufności równy 0.01 (oczywiście p - wartości mniejsze od 0.01 oznaczają odrzucenie hipotezy zerowej mówiącej, że próbka pochodzi z zadanego rozkładu). Puste komórki w tablicy oznaczają zaś brak podstaw do odrzucenia hipotezy zerowej tj. p-wartość słabo większą od 0.01. Ostatnie dwie kolumny powyższej tablicy mają analogiczne znaczenie, ale tym razem dla testu Andersona - Darlinga.

Zobaczmy najpierw *Panel A* zawierający analizę dla danych oryginalnych. Widzimy, że dane mogą być dobrze dopasowane przez rozkład log - normalny. Hipoteza zerowa nie została odrzucona przez obydwa testy K-S i A-D jedynie dla tego właśnie rozkładu. Dla wszystkich innych rozkładów dopasowanie jest gorsze. Jeśli chodzi o *Panel B*, czyli analizę na danych zlogarytmowanych, widzimy, że można by dopasować więcej rozkładów, a mianowicie: normalny, skośny – normalny oraz skośny – T-studenta. Dla nich hipoteza zerowa nie jest odrzucana.

	Original data	LN data
Best fit	**Log-normal not rejected in:** - 6 out of 7 cases for the different entities - 7 out of 8 cases for the different types of data breaches	**Skew-normal not rejected in:** - 7 out of 7 cases for the different entities - 8 out of 8 cases for the different types of data breaches
2nd best fit	**GPD not rejected:** - 3 out of 7 cases for the different entities - 6 out of 8 cases for the different types of data breaches	**Normal not rejected in:** - 6 out of 7 cases for the different entities - 7 out of 8 cases for the different types of data breaches

Rysunek 3.4: Podsumowanie zgodności rozkładów dla rozmiarów naruszeń, 2266 obserwacji (źródło: [23]).

W kolejnej tabelce (rys. 3.4) widzimy krótkie podsumowanie dla wszystkich 15 typów, tzn. 8 typów naruszeń (włącznie z *Unknown*) oraz 7 typów organizacji (*entities*). Według autorów [23], najlepszym wyborem rozkładu, który można dopasować do danych oryginalnych, jest log - normalny. Drugim najlepszym jest uogólniony rozkład Pareto, tzn. GPD. Natomiast dla danych zlogarytmowanych dwa najlepsze rozkłady to skośny – normalny oraz normalny. Patrząc tylko i wyłącznie na powyższe testy statystyczne, możemy zauważyć, że najlepiej wypada rozkład skośny – normalny na danych zlogarytmowanych, dla którego hipoteza zerowa nie została odrzucona ani razu. Autorzy wnioskują, że można by się zastanawiać, który z tych dwóch rozkładów wybrać. Skośny – normalny ma jeden parametr więcej do estymacji niż log – normalny (tzn. skośny ma 3 parametry, a normalny 2), więc ciężko im jednoznacznie stwierdzić, który rozkład byłby lepszy.

Zmienna *interarrival times*

Z kolei dla zmiennej *interarrival times*, przeprowadzona analiza była dużo krótsza. Uzyskane wyniki przedstawione są w poniższej tablicy (rys. 3.5).

Model	**Chi - square Test**		**Kolmogorov - Smirnov Test**	
Original Data	test stat.	p - value	test stat.	p - value
Poisson	> 10000	< 0.01	0.80	< 0.01
Negative Binomial	51.57	< 0.01	0.22	

Rysunek 3.5: Analiza zgodności rozkładów dla *interarrival times*, typ PORT, 629 obserwacji (źródło: [23]).

Tabela wynikowa skonstruowana jest analogicznie jak dla zmiennej *breach size*. W pierwszej kolumnie widzimy analizowane rozkłady (tj. rozkład Poissona oraz rozkład ujemny dwumianowy), w drugiej - kolejne liczby oznaczają wartości statystyki testowej, a w trzeciej - widzimy

p - wartości dla testu Chi - kwadrat. Ostatnie dwie kolumny mają analogiczne znaczenie, ale tym razem dla testu Kołmogorowa - Smirnowa. W tym miejscu należy zaznaczyć, że z publikacji nie wynika wprost, z jakiej wersji tego testu korzystano. Ponieważ podstawową wersję stosować można tylko na rozkładach ciągłych, to założyłem, że autorzy korzystali z wersji dyskretnej tego testu (opisanej w skrócie w rozdz. 3.2.1) oraz przedstawiam, jakie otrzymali wyniki.

Łatwo widać, że rozkład ujemny dwumianowy najlepiej pasuje do danych. Analiza pokazuje, że hipotezy zerowej nie trzeba odrzucać dla K-S testu w przypadku tego rozkładu. Autorzy dodają, że podobne wnioski wyciągnąć można dla innych typów naruszeń. Wspominają też, że próbowali dopasować do danych także dwa dodatkowe rozkłady tj. obcięty rozkład Poissona oraz obcięty rozkład ujemny dwumianowy. Jednakże wyniki były gorsze, niż dla standardowych wersji tych rozkładów, więc zdecydowali się je pominąć.

Wyniki na danych (PRC)

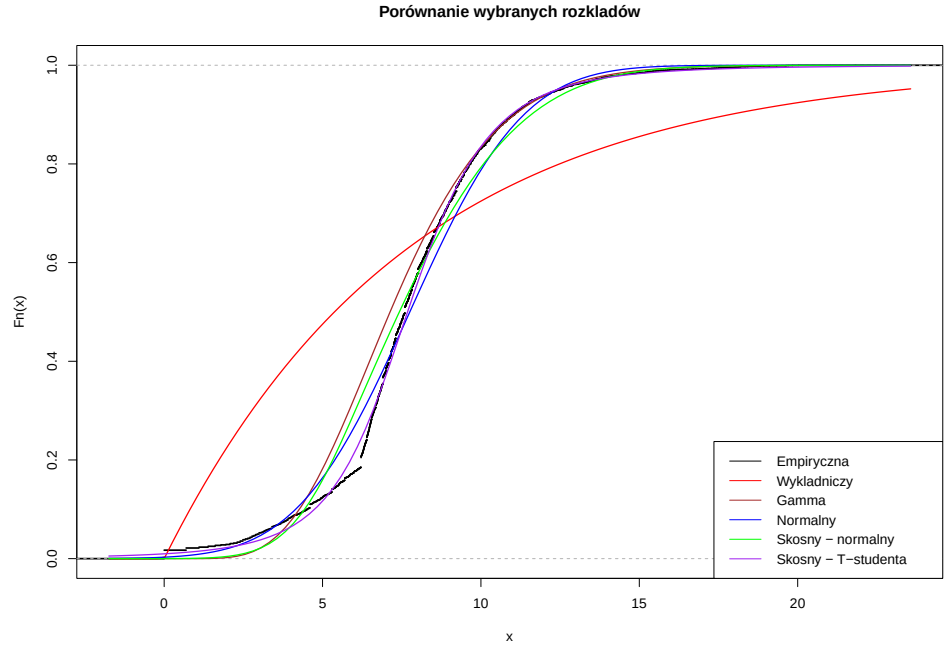
Moim celem była weryfikacja empiryczna, czy modelowanie rozkładami (jak również jakimi) jest zasadne w przypadku zmiennych *breach size* i *interarrival times*.

Przeprowadziłem podobne, ale znacznie rozszerzone analizy na danych (PRC). Mianowicie dane podzieliłem na 7 typów organizacji oraz 8 typów naruszeń. Dodatkowo rozpatrywałem też cały zbiór danych. Do odpowiednio oczyszczonych i przetworzonych danych (m.in. usunąłem rekordy zawierające wartości 0 lub NA dla zmiennej *Total.Records* - rozmiar naruszenia oraz podzieliłem zbiór danych na dane oryginalne i zlogarytmowane) próbowałem dopasować następujące rozkłady: Wykładniczy, Gamma, Log - Normalny, Normalny, Weibulla, Skośny - Normalny, Log - Skośny - Normalny, Skośny - T-Studenta, Uogólniony Rozkład Pareto (GPD) oraz Log - Logistyczny (patrz tablice 2.1 i 2.2).

Tak jak Eling i Loperfido [23] zająłem się głównie modelowaniem dotkliwości szkód tzn. rozmiarów naruszeń (z ang. *breach sizes*). Pod koniec tego rozdziału wspomnę także o proponowanych przez autorów 2 rozkładach, które próbowałem dopasować do zmiennej *interarrival times*. Analogicznie, jak powyżej, zacznę od ciekawszej części, czyli modelowania dotkliwości szkód.

Zmienna *breach sizes*

Aby dopasować wybrany rozkład do danych, musiałem najpierw oszacować jego parametry (patrz kolumna *Estymacja* w tablicach 2.1 i 2.2). Dla rozkładów o mniejszej liczbie parametrów (głównie 2 parametrowych) szacowałem na podstawie znanych zależności rozkładów, czyli wartości oczekiwanej, wariancji lub mediany (dla rozkładu log - normalnego), a dla rozkładów o większej ilości parametrów (czyli 3 i więcej) wykorzystałem metodę największej wiarygodności. Korzystałem z istniejących bibliotek tj. fGarch, sn, SpatialExtremes oraz actuar, dla języka R.



Rysunek 3.6: Dystrybuanta empiryczna oraz dystrybuanty wybranych rozkładów.

Na rysunku 3.6 widzimy wykres dystrybuanty empirycznej (czarna łamana linia) oraz dystrybuanty rozkładów: wykładniczego – linia czerwona, gamma – linia brązowa, normalnego – niebieska, skośnego normalnego – zielona oraz skośnego T-studenta – fioletowa. Od razu widać, że rozkład wykładniczy w ogóle nie pasuje do danych. Jest to jednak rozkład o jednym parametrze, więc warto przynajmniej spróbować dopasować prosty rozkład. Dalej widzimy, że rozkłady: normalny, skośny - normalny oraz skośny T-studenta wyglądają na zbliżone do dystrybuanty empirycznej. Widać tu zgodność z wnioskami z artykułu Elinga, Loperfido [23]. Rozkład Gamma dopasowuje się gorzej niż wspomniane powyżej, choć nadal znacznie lepiej, niż rozkład wykładniczy.

W celu weryfikacji powyższych hipotez przeprowadziłem testy statystyczne Kołmogorowa – Smirnowa oraz Andersona – Darlinga.

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	< 0.01	0.00060
Normal	< 0.01	**0.16184**
Weibull	< 0.01	0.00060
Skew-Normal	< 0.01	**0.38374**
Log-Logistic	< 0.01	0.00060
Skew-Student	< 0.01	**0.34460**
GPD	< 0.01	0.00060

Tablica 3.3: Wyniki testów dla całego zbioru danych zlogarytmowanych (6822 obserwacji).

Tablica 3.3 przedstawia wyniki dla całego zbioru danych zlogarytmowanych (tzn. 6822 obserwacji). W pierwszej kolumnie widzimy rozkłady, które próbowałem dopasować do danych. Kolejne dwie kolumny to p - wartości odpowiadające wspomnianym testom K-S oraz A-D. Obserwujemy, że p - wartości odpowiadające rozkładom: normalny, skośny - normalny i skośny - T-studenta są większe od 0.05, co oznacza brak podstaw do odrzucenia hipotezy zerowej, mówiącej, że próbka pochodzi z danego rozkładu. Jest to zgodne z naszymi przypuszczeniami, wypływającymi z analizy wykresu dystrybuant (rysunek 3.6).

Przechodząc do testu K-S (z gwiazdką) w wersji podstawowej zauważmy, że do jego przeprowadzenia musimy znać lub oszacować parametry rozkładów. Ponieważ parametry testowanych rozkładów były szacowane, to znalezione oszacowania p - wartości są obciążone. Z tego powodu wyniki testu K-S skonfrontowałem z wynikami testu A-D. W drugim teście odpowiednia poprawka na szacowanie istnieje i została przeze mnie uwzględniona. Wykorzystałem funkcję ad.test() z biblioteki goftest dostępnej w języku R, w której, w celu skorygowania efektu estymacji parametrów, wystarczyło ustawić parametr estimated na TRUE. Bez względu na powyższe, standardowy test Kołmogorowa - Smirnowa stosowany jest w wielu publikacjach (patrz np. [22] lub [23]). Postanowiłem więc umieścić go także w swojej pracy.

W ogólności, dla testu K-S taka poprawka też jest możliwa (wrócę do tego w dalszej części pracy). Patrząc jedynie na test K-S (w tym kontekście), do danych nie pasuje żaden z proponowanych

wanych rozkładów.

Przyjrzyjmy się teraz tabelkom, które przedstawiają wyniki testów dla każdego kolejnego typu naruszenia.

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00072
Gamma	< 0.01	0.00060
Log-Normal	0.503795	0.00072
Normal	0.321004	**0.23798**
Weibull	0.576420	0.00072
Skew-Normal	0.478986	**0.16427**
Log-Logistic	0.907220	0.00072
Skew-Student	0.926493	**0.53810**
GPD	< 0.01	0.00060

Tablica 3.4: Wyniki testów dla typu naruszenia: CARD (32 obserwacji).

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.00058
Normal	< 0.01	**0.51637**
Weibull	< 0.01	0.00058
Skew-Normal	< 0.01	**0.48853**
Log-Logistic	< 0.01	0.00058
Skew-Student	< 0.01	0.02603
GPD	< 0.01	0.00058

Tablica 3.5: Wyniki testów dla typu naruszenia: DISC (1553 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00059
Log-Normal	< 0.01	0.00059
Normal	< 0.01	**0.76131**
Weibull	< 0.01	0.00059
Skew-Normal	< 0.01	**0.42059**
Log-Logistic	< 0.01	0.00060
Skew-Student	0.026033	**0.57370**
GPD	< 0.01	0.00059

Tablica 3.6: Wyniki testów dla typu naruszenia: HACK (1603 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00057
Gamma	< 0.01	0.00057
Log-Normal	< 0.01	0.00057
Normal	0.235794	**0.31404**
Weibull	0.939118	0.00057
Skew-Normal	0.804825	**0.16432**
Log-Logistic	0.167850	0.00057
Skew-Student	0.802525	**0.15780**
GPD	< 0.01	0.00057

Tablica 3.7: Wyniki testów dla typu naruszenia: INSD (376 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.00058
Normal	< 0.01	**0.28315**
Weibull	< 0.01	0.00058
Skew-Normal	< 0.01	**0.33631**
Log-Logistic	< 0.01	0.00058
Skew-Student	< 0.01	**0.72665**
GPD	< 0.01	0.00058

Tablica 3.8: Wyniki testów dla typu naruszenia: PHYS (1474 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00225
Gamma	< 0.01	0.00060
Log-Normal	< 0.01	0.00207
Normal	0.028998	**0.35129**
Weibull	< 0.01	**0.98587**
Skew-Normal	0.048748	**0.49386**
Log-Logistic	0.578447	**0.70014**
Skew-Student	0.178852	**0.32423**
GPD	< 0.01	0.00060

Tablica 3.9: Wyniki testów dla typu naruszenia: PORT (874 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.08373
Gamma	< 0.01	0.00060
Log-Normal	0.583496	**0.60865**
Normal	0.681660	**0.47383**
Weibull	0.396781	**0.54568**
Skew-Normal	0.774812	**0.45951**
Log-Logistic	0.933342	**0.94034**
Skew-Student	0.853719	**0.27313**
GPD	< 0.01	0.00077

Tablica 3.10: Wyniki testów dla typu naruszenia: STAT (184 obserwacji)

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.00058
Normal	< 0.01	**0.87225**
Weibull	< 0.01	0.00058
Skew-Normal	< 0.01	**0.65201**
Log-Logistic	< 0.01	0.00058
Skew-Student	< 0.01	**0.78149**
GPD	< 0.01	0.00058

Tablica 3.11: Wyniki testów dla typu naruszenia: UNKN (637 obserwacji)

Patrząc na kolejne typy widzimy, że:

- Dla typu naruszenia CARD, patrząc na test A-D widzimy, że ponownie nie zostały odrzucone te same rozkłady, co dla całego zbioru danych. Z kolei wg. testu K-S (z gwiazdką) zostały odrzucone tylko trzy rozkłady. Zauważmy, że dla tego typu naruszenia dostępne były tylko 32 obserwacje;
- Dla typu DISC dostępnych było już stosunkowo dużo obserwacji tj. 1553. Nie odrzucone rozkłady przez test A-D to normalny i skośny normalny. Natomiast test K-S odrzuca wszystkie;
- Dla typu HACK, wg. testu A-D nie zostały odrzucone dokładnie te same rozkłady, tymczasem test K-S ponownie odrzuca wszystko;
- Dla typu INSD: to samo wg. testu A-D, lecz więcej możliwości przyjęcia wg. testu K-S. Ponownie dostępnych było mało obserwacji;
- Dla typu PHYS: analogicznie te same wnioski co powyżej;
- Dla typu PORT moglibyśmy przyjąć już znacznie więcej rozkładów. Dodatkowo zaskakująca może być wyznaczona p - wartość dla rozkładu Weibulla;

- Dla typu STAT ponownie widzimy, że mało było dostępnych obserwacji. Kolejny raz rozkładów możliwie przyjętych przez obydwa testy jest więcej. Wspomniane trzy rozkłady powtarzają się cały czas;

- Dla typu UNKN - wyniki analogiczne do typów INSD i PHYS.

Typ	**Wykładniczy**	**Log-Normalny**	**Normalny**	**Weibulla**
CARD	0.0002	**0.3254**	0.0106	0.0534
DISC	0.0002	0.0002	0.0002	0.0002
HACK	0.0002	0.0002	0.0002	0.0002
INSD	0.0002	0.0002	0.0218	**0.4728**
PHYS	0.0002	0.0002	0.0002	0.0002
PORT	0.0002	0.0002	0.0002	0.0002
STAT	0.0002	0.0166	**0.2408**	0.0400
UNKN	0.0002	0.0002	0.0002	0.0002
Wszystkie	0.0002	0.0002	**0.0902**	0.0002

Tablica 3.12: Wyniki (p-wartości) testu Kołmogorowa - Smirnowa z poprawką.

Dalej widzimy tablicę 3.12, która zawiera p - wartości dla kolejnych typów naruszeń oraz dla całych danych (*Wszystkie*). Tym razem wykorzystałem wspominany w rozdziale 3.2.1 test K-S z poprawką Lillieforsa [56, 57]. Jest on odpowiedni, gdy parametry są nieznane i muszą być szacowane. Implementacja tego testu pochodzi z biblioteki KScorrect (funkcja LcKS()), dla języka R (rozkłady obsługiwane przez tę funkcję, ze zbioru rozważanych przeze mnie, to: rozkład normalny, wykładniczy, Weillbula i log-normalny).

Przechodząc już do wyników widzimy, że do typu naruszenia CARD mógłby pasować rozkład log - normalny, dla typu INSD rozkład Weibulla lub ewentualnie normalny, a dla STAT i dla całych danych można by dopasować rozkład normalny.

Breach type	**p-wartość**	**Org type**	**p-wartość**
CARD	0.02148	BSF	< 0.001
DISC	< 0.001	BSO	< 0.001
HACK	< 0.001	BSR	< 0.001
INSD	< 0.001	EDU	**0.07897**
PHYS	< 0.001	GOV	0.00588
PORT	< 0.001	MED	< 0.001
STAT	**0.19844**	NGO	**0.66797**
UNKN	< 0.001	UNKN	< 0.001
Wszystkie	< 0.001		

Tablica 3.13: Wyniki (p-wartości) testu Shapiro - Wilka normalności.

W kolejnej tablicy 3.13 widzimy p - wartości dla testu Shapiro – Wilka normalności zmiennej *breach size* z podziałem na typy naruszeń (*breach type*), typy organizacji (*org type*) oraz (na samym dole) mamy też wartość dla całych danych (*Wszystkie*). Ponownie wyniki nie są zbyt optymistyczne. Na poziomie ufności równym 0.05 moglibyśmy nie odrzucić hipotezy o normalności rozkładu jedynie dla typu naruszenia STAT oraz typów organizacji EDU i NGO.

Model	Kolmogorov.Smirnov.Test*	Anderson.Darling.Test
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	< 0.01	**0.09629**
Normal	< 0.01	0.00060
Weibull	< 0.01	0.00060
Skew-Normal	< 0.01	0.00060
Log-Logistic	< 0.01	**0.48023**
Skew-Student	< 0.01	0.00060
GPD	< 0.01	**0.07241**

Tablica 3.14: Wyniki analizy na danych oryginalnych: cały zbiór danych (6822 obserwacji).

Na koniec przyjrzyjmy się jeszcze tablicy 3.14 (analogicznej do tablicy 3.3) dla danych oryginalnych (czyli danych niezlogarytmowanych). Widzimy, że można by dopasować rozkład log - normalny według A-D testu, co jest oczywiście równoważne z dopasowaniem rozkładu normalnego na danych zlogarytmowanych. Jest to jedyny rozkład, który pokrywa się z wcześniejszymi wynikami. Bardzo dużą p - wartość widzimy dla rozkładu log - logistycznego. Analizując wyniki testu A-D można się zatem zastanowić, czy rozkład log - logistyczny nie byłby dobrym wyborem. Analogiczne badania przeprowadziłem dla różnych typów naruszeń i organizacji. Otrzymałem wyniki prowadzące do podobnych wniosków. Jednak ze względu na przejrzystość pracy zdecydowałem się nie umieszczać tabel z wynikami w tym miejscu. Znajdują się one w dodatku A. na samym końcu mojej pracy.
Według testu K-S moglibyśmy przyjąć jedynie rozkłady: log - normalny, log - logistyczny i GPD dla poszczególnych typów naruszeń oraz organizacji. Wyniki testu A-D są dość podobne, choć ewentualnie przyjętych rozkładów mogłoby być więcej.

Zmienna *interarrival times*

Aby analogicznie jak powyżej dopasować 2 wybrane rozkłady do danych, musiałem najpierw oszacować ich parametry. Dla rozkładu Poissona wystarczyło szacować na podstawie wartości oczekiwanej, z kolei dla rozkładu ujemnego dwumianowego wykorzystałem metodę największej

wiarygodności. Do estymacji wykorzystałem funkcję fitdist() z biblioteki fitdistrplus w języku R.

W celu porównania wyników mojego badania z tymi otrzymanymi przez autorów [23], przeprowadziłem testy statystyczne Chi - kwadrat oraz Kołmogorowa – Smirnowa. Jak już wspominałem w poprzednim podrozdziale, podstawowa wersja testu K-S działa tylko dla rozkładów ciągłych. Z tego względu wykorzystałem wersję dyskretną testu Kołmogorowa - Smirnowa, w skrócie opisaną w rozdziale 3.2.1. Została ona zaimplementowana w bibliotece iZID, z której wykorzystałem funkcję dis.kstest() z odpowiednimi parametrami.

Wyniki obydwu testów dla poszczególnych typów naruszeń umieściłem w dwóch poniższych tablicach. Zauważmy od razu, że liczba obserwacji w przypadku analizy zmiennej *interarrival times* (tj. 9014) jest znacznie wyższa niż przy *breach sizes* (tj. 6822). Wynika to np. z faktu, że przy liczeniu czasów *interarrival* i oczyszczaniu danych nie pozbywałem się wartości równych 0.

Typ	Model	Test Kołmogorowa Smirnowa**	Test Chi kwadrat
Wszystkie (9014)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
CARD (67)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	$\geqslant$ **0.05**
DISC (1860)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
HACK (2532)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
INSD (605)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
PHYS (1732)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
PORT (1171)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
STAT (248)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	$\geqslant$ **0.05**
UNKN (703)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05

Tablica 3.15: Wyniki testów dla całych danych oraz z podziałem na typy naruszeń. Liczby w nawiasach oznaczają ilość obserwacji.

Typ	Model	Test Kołmogorowa Smirnowa**	Test Chi kwadrat
BSF (81)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
BSO (180)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
BSR (107)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
EDU (212)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
GOV (83)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
MED (553)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05
NGO (21)	Poisson	< 0.01	< 0.05
	Negative - Binomial	< 0.01	< 0.05

Tablica 3.16: Wyniki testów dla poszczególnych typów organizacji. Liczby w nawiasach oznaczają ilość obserwacji.

Tablica 3.15 przedstawia wyniki testów dla całego zbioru danych (Wszystkie), a także z podziałem na typy naruszeń (analizowałem rozkłady Poissona oraz ujemny dwumianowy). Z kolei tablica 3.16 prezentuje wyniki testów dla kolejnych typów organizacji. W obydwu tablicach wartości w trzeciej kolumnie to przybliżone p - wartości wspomnianego testu K-S w wersji dyskretnej, a w czwartej - dla testu Chi kwadrat. Analiza pokazuje, że hipotezy zerowej nie trzeba odrzucać dla testu Chi - kwadrat w przypadku rozkładu ujemnego dwumianowego jedynie dla typów naruszeń CARD i STAT. Dla nich p - wartości tego testu są większe lub równe 0.05 co mówi nam, że dane mogą być stosunkowo dobrze dopasowane przez ten rozkład. Z kolei, w przypadku testu K-S, hipotezy zerowe dla wszystkich rozkładów i typów zostały odrzucone.

Wnioski

W przypadku zmiennej *breach size* możemy sformułować następujące wnioski:

1. Na podstawie wyników dla testu Andersona - Darlinga, możemy powiedzieć, że:

- Dla danych przekształconych logarytmicznie moglibyśmy się zastanowić nad dopasowaniem jednego z następujących rozkładów: normalnego, skośnego - normalnego lub skośnego - T-studenta;
- Dla danych oryginalnych można by dopasować rozkłady: log - normalny, log – logistyczny lub Weibulla dla odpowiednich typów.

2. Najbardziej obiecujący jest tu rozkład log – normalny. Wniosek ten pokrywa się z wynikami publikacji [23], w której autorzy zdecydowali się wybrać właśnie ten rozkład obok rozkładu skośnego - normalnego.

3. Test Kołmogorowa - Smirnowa (czy to w wersji podstawowej, czy też z poprawką) jest tutaj trudny w użyciu. Nie daje wyników, na podstawie których można wysnuć sensowne wnioski.

4. Test Shapiro - Wilka nie znajduje powodów do odrzucenia hipotezy zerowej o normalności danych zlogarytmowanych dla typu naruszenia STAT oraz typów organizacji EDU i NGO. Jest to zgodne z wynikami testu A-D dotyczącego log - normalności danych.

Podsumowując, w przypadku zmiennej *breach sizes* modelowanie przy użyciu rozkładów jest trudne. Problemem jest rozbieżność w wynikach różnych testów statystycznych, co widzimy po rezultatach powyższej analizy. Ponadto, istnieje konieczność korekt wynikających z nieznanych, szacowanych na podstawie danych parametrów. Co więcej, wyniki są trudne w interpretacji. Na podstawie przeprowadzonych analiz jedynie rozkład log - normalny wydaje się być sensownym wyborem.

Z kolei dla zmiennej *interarrival times* możemy sformułować następujące wnioski:

1. Modelowanie rozkładami prawdopodobieństwa tej zmiennej jest trudne;

2. Podobnie jak dla zmiennej *breach size*, otrzymane wyniki nie są łatwe w interpretacji;

3. Test Kołmogorowa - Smirnowa w wersji dyskretnej nie daje żadnych sensownych wyników, na podstawie których moglibyśmy wybrać jeden z proponowanych rozkładów;

4. Według testu Chi - kwadrat jedyny rozkład, który moglibyśmy dopasować do danych, to ujemny dwumianowy dla typów naruszeń CARD oraz STAT.

Widzimy zatem, że dla obydwu zmiennych modelowanie rozkładami prawdopodobieństwa może nie być perfekcyjnym podejściem. Nie jest to jednak jedyne znane podejście w literaturze (patrz rozdz. 2.2). Przejdę teraz do omówienia wyników z podejścia stochastycznego.

3.3.2. Procesy stochastyczne

W poniższym podrozdziale przedstawię wyniki dotyczące modelowania procesami stochastycznymi. Zacznę od modelowania znanej już zmiennej *interarrival times*, aby następnie przejść do modelowania zmiennej *breach sizes*. Przypomnę, że kolejne podrozdziały odnoszą się do kroków ogólnego schematu przedstawionego w sekcji 2.2.

Autorzy analizowanej przeze mnie publikacji [24] korzystali z danych pobranych ze wspomnianej już strony PRC [25]. Zostały one przygotowane przez nich w następujący sposób:

- wybranie okresu od 01.01.2005 r. do 07.04.2017 r.

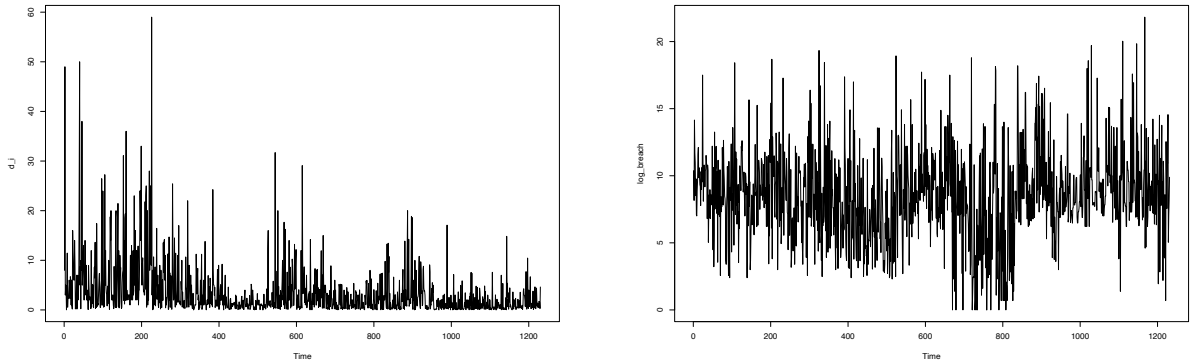
- określenie wybranego typu naruszeń (analizy rozpocząłem od typu HACK wraz ze wszystkimi typami organizacji);

- przekształcenie zmiennych zgodnie z opisem: losowe porządkowanie incydentów odpowiadających temu samemu dniu, a następnie wstawienie małego, losowego przedziału czasowego pomiędzy dwa kolejne incydenty, zapewniając jednocześnie, że incydenty te odpowiadają nadal temu samemu dniu.

W ten sposób pozostało im dokładnie 600 rekordów. Przypomnę, że nie miałem dostępu do tych samych danych co autorzy analizowanego artykułu. Źródłem moich danych była ta sama strona internetowa, ponieważ jednak nie udostępnia ona danych archiwalnych, nie mogłem użyć dokładnie tych samych danych co oni. Miałem dostęp do znacznie większej ich ilości. W moim wypadku po analogicznym przygotowaniu danych pozostało dokładnie 1244 obserwacji.

W celu głębszego zbadania tego problemu wykonałem analogiczne analizy na typie naruszenia HACK, a także na innych typach naruszeń (badanych też przez innych autorów artykułów, patrz np. [2, 59, 74]).

Pod koniec tego rozdziału przedstawię skrócone wyniki analizy, którą przeprowadziłem na wszystkich typach naruszeń. Zacznę jednak od przedstawienia wyników porównawczych dla typu HACK.



Rysunek 3.7: Wykresy szeregów czasowych *interarrival times* oraz zlogarytmowanych *breach sizes* odpowiednio (dla naruszenia typu HACK).

Rysunek 3.7 przedstawia wykresy szeregów czasowych interesujących nas zmiennych dla typu naruszenia hakerskiego. Na pierwszym z nich znajduje się szereg czasowy *interarrival times* (jednostka: dzień). Zauważamy, że większość czasów *interarrival* jest niewielka (mniej niż 20 dni), a ostatnie czasy między zdarzeniami są jeszcze mniejsze, co sugeruje, że częstotliwość naruszeń hakerskich wzrasta. Widzimy zatem, że wiele zdarzeń występuje w krótkim okresie czasu. Wykres pokazuje również, że incydenty są rozmieszczone nieregularnie (tzn. wykazują zarówno duże, jak i małe czasy *interarrival*).

Drugi wykres przedstawia szereg czasowy *breach sizes* przekształconych logarytmicznie, również dla typu naruszenia hakerskiego (jednostka: jeden rekord). Został on przedstawiony w skali logarytmicznej, gdyż rozmiary naruszeń wykazują bardzo dużą zmienność oraz skośność, co utrudnia ich modelowanie. Nadal obserwujemy dużą zmienność, jednak widzimy też, że niektóre rozmiary naruszeń są szczególnie wielkie (oznaczają poważne incydenty naruszenia hakerskiego). Dodatkowo można zauważyć duże (odp. małe) zmiany, po których następują duże (odp. małe) zmiany.

Modelowanie *interarrival times*

Swoją analizę rozpocząłem od modelowania czasów *interarrival*. W tablicy 3.17 widzimy podstawowe statystyki czasów między zdarzeniami dla całego zbioru danych (co odpowiada rys. 3.7) oraz z podziałem na kategorie biznesu. Ponownie, tak jak autorzy, rozpatrywałem tylko typ naruszenia hakerskiego[5]. Możemy tu zaobserwować, że odchylenie standardowe w każdej kategorii jest widocznie większe od średniej, co sugeruje, że procesy opisujące te naruszenia nie są procesami Poissona (gdyż wartość oczekiwana i wariancja rozkładu Poissona są sobie równe, patrz [24]).

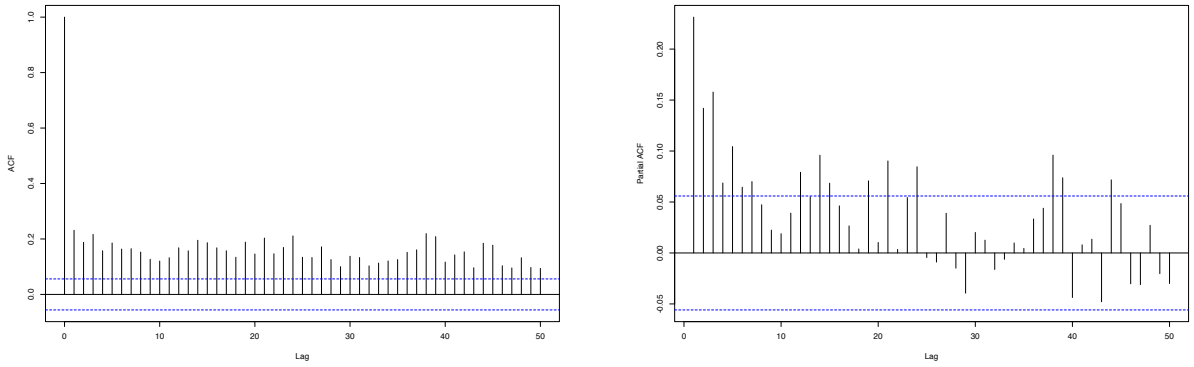
[5]Tablicę przedstawiającą analogiczne statystyki dla czasów między zdarzeniami, łącznie dla wszystkich typów naruszeń przedstawiłem w dodatku A. znajdującego się na końcu mojej pracy.

Zauważmy też np., że maksymalny czas *interarrival* incydentów naruszenia NGO wynosi aż 1178 dni, a maksymalny czas *interarrival* dla wszystkich wynosi tylko 59 dni.

Typ naruszenia	**Min**	**Median**	**Mean**	**SD**	**Max**	**Total**
BSF	0.01621	22.90002	53.14204	60.37435	260.0000	81
BSO	0.01416	10.00000	22.86189	34.71244	220.0000	180
BSR	0.03301	15.00000	40.71872	74.09488	444.9033	107
EDU	0.00380	10.00000	20.42721	28.69755	217.6519	212
GOV	0.13379	24.00000	50.08434	69.70996	324.6988	83
MED	0.00165	1.92241	7.19990	28.52360	497.0000	553
NGO	0.18270	52.68858	156.64528	298.38103	1177.1355	21
Wszystkie	0.00075	1.76772	3.59416	5.47395	59.0000	1243

Tablica 3.17: Statystyki opisowe dla *interarrival times*, dla typu naruszenia hakerskiego (jednostka: dzień), (kolejno: minimum, mediana, średnia, odchylenie standardowe, maksimum i liczba obserwacji).

Następnie popatrzmy na rysunek 3.8. Przedstawia on odpowiednio wykresy próbkowych funkcji ACF i PACF. Na obu wykresach widać, że wartości korelacji przekraczają przerywane niebieskie linie (tj. przedziały ufności dla białego szumu). Na pierwszym rysunku słupki nigdy nie kończą się poniżej tej linii, pomimo wysokiego lagu równego 50. Z kolei na drugim rysunku siódmy słupek znajduje się już w przedziale ufności, jednakże 11-ty, a tym bardziej 13-ty wystaje już znacznie bardziej. Dodatkowo widać też minimalny trend malejący (por. rys. 3.7). Z powyższego wnioskujemy, że pomiędzy czasami *interarrival* występują wyraźne korelacje. Wyniki moich badań potwierdzają to, co zasugerowali autorzy [24], że czasy między zdarzeniami dla typu naruszenia hakerskiego powinny być modelowane przez procesy stochastyczne, a nie przez rozkłady.



Rysunek 3.8: Wykresy próbkowych funkcji odpowiednio ACF i PACF, dla *interarrival times*.

Przypomnę, że zbiór danych reprezentowany jest przez ciąg $\{(t_i, y_{t_i})\}_{0 \leqslant i \leqslant n}$, gdzie n to liczba obserwacji, a t_i to czas, w którym wystąpił incydent o rozmiarze y_{t_i}, dla $i \geqslant 1$. Ponadto $d_i = t_i - t_{i-1}$ to czasy *interarrival*, dla $i = 1, 2, ..., n$.

Do modelowania czasów *interarrival* wykorzystałem model ACD (z ang. *Autoregressive Conditional Mean*) lub jedną z jego logarytmicznych wersji[6]. Modele te wyselekcjonowałem zgodnie z literaturą (patrz np. [24, 75, 62]). Ich podstawową ideą jest standaryzacja czasów *interarrival* d_i poprzez wykorzystanie informacji historycznych, dla $i = 1, 2, ..., n$. Ogólnie model ACD może służyć jako narzędzie do modelowania czasów *interarrival* pomiędzy wybranymi zdarzeniami procesu transakcyjnego, jak np. zawarcie transakcji lub zmiana ceny. Wyniki takich badań służą do weryfikacji wybranych hipotez dla modeli rynku finansowego.

Przejdę teraz do definicji formalnej. W modelu ACD (oraz w jego logarytmicznych wersjach) zmienną d_i wyraża się jako następujący iloczyn:

$$d_i = \Psi_i \cdot \varepsilon_i, \tag{3.2}$$

gdzie Ψ_i to funkcje historycznych czasów *interarrival* tj.:

$$\Psi_i = \mathbb{E}(d_i | \mathcal{F}_{i-1}) = \mathbb{E}(d_i | d_{i-1}, ..., d_1), \tag{3.3}$$

$\mathcal{F}_{i-1}$ reprezentuje wiedzę zgromadzoną do momentu t_{i-1}, natomiast ε_i to innowacje - niezależne zmienne losowe o tym samym rozkładzie (*i.i.d.*), o wartości oczekiwanej równej 1 oraz gęstości $f_\varepsilon(\varepsilon_i)$. Najbardziej popularnym rozkładem zmiennej losowej ε_i jest rozkład wykładniczy. Ponadto zakładamy, że $\mathcal{F}_{i-1}$ jest niezależne od ε_i.

Wtedy specyfikacje modeli, uwzględniając równanie średniej warunkowej (3.3) i (3.2), dane są następująco:

[6]Model ACD początkowo zaproponowany został przez Engle i Russell w 1997 r. do modelowania dynamiki finansowych czasów między zdarzeniami.

1. Standardowy model ACD(p, q):

$$\Psi_i = \omega + \sum_{j=1}^{p} a_j d_{i-j} + \sum_{j=1}^{q} b_j \Psi_{i-j}, \tag{3.4}$$

gdzie indeks i oznacza i-ty incydent (naruszenie); $\omega > 0$; $a_j \geqslant 0$, dla $j = 1, ..., p$; $b_j \geqslant 0$, dla $j = 1, ..., q$, natomiast p i q to dodatnie liczby całkowite oznaczające rząd wyrażenia autoregresyjnego;

2. Model log-ACD typu I (w skrócie LACD$_1(p, q)$):

$$\log(\Psi_i) = \omega + \sum_{j=1}^{p} a_j \log(\varepsilon_{i-j}) + \sum_{j=1}^{q} b_j \log(\Psi_{i-j});$$

3. Model log-ACD typu II (w skrócie LACD$_2(p, q)$):

$$\begin{aligned}\log(\Psi_i) &= \omega + \sum_{j=1}^{p} a_j \log(d_{i-j}) + \sum_{j=1}^{q} b_j \log(\Psi_{i-j}) = \\ &= \omega + \sum_{j=1}^{p} a_j \log(\Psi_{i-j}\varepsilon_{i-j}) + \sum_{j=1}^{q} b_j \log(\Psi_{i-j});\end{aligned}$$

Różne warianty modeli ACD i LACD mogą być wyprowadzone np. poprzez określenie różnych funkcji gęstości dla zmiennej ε_i.
Różnice pomiędzy modelami ACD i log-ACD są dość małe. Widać tu, że różnica pomiędzy modelem LACD$_1$, a LACD$_2$ to w pierwszej sumie przemnożenie członu ε_{i-j} przez Ψ_{i-j} w logarytmie. Ponadto, modele log-ACD są bardziej wszechstronne, ponieważ nie mają one ograniczeń dotyczących znaku ich współczynników.

W pracy [24, str. 6] wybór powyższych modeli został zarekomendowany po przeprowadzeniu analizy wstępnej. Zwróćmy uwagę, że modele te są wystarczająco uniwersalne w tym sensie, że są dostatecznie proste, ale nie zbyt proste (por. np. proste modele wykorzystywane w analizie regresji liniowej). Mogą więc być skutecznie dopasowywane w praktyce.

W dalszej części pracy będę rozpatrywał powyższe modele ACD z ustalonymi parametrami $p = q = 1$. Jest to zgodne z podejściem przedstawionym w pracach [24, 29, 62]. Przedstawione tam wnioskowanie opiera się na założeniu, że wyższe rzędy modelu niekoniecznie poprawiają dokładność predykcji. W tej sytuacji uproszczenie modelu może okazać się znacznie korzystniejsze.

W przypadku innowacji ε_i zakładamy, że mają one uogólniony rozkład Gamma, którego funkcja gęstości prawdopodobieństwa ma postać:

$$f(x|\lambda, \gamma, k) = \frac{\gamma x^{k\gamma-1}}{\lambda^{k\gamma}\Gamma(k)} \exp\Big[-\Big(\frac{x}{\lambda}\Big)^{\gamma}\Big], \tag{3.5}$$

gdzie $\lambda > 0$ to parametr skali; $\gamma, k > 0$ to parametry kształtu; a $\Gamma(k) = \int_0^\infty t^{k-1}\, e^{-t}\, dt$, $k > 0$ to funkcja Gamma. Rozkład ten rekomendowany jest np. przez autorów prac [24, 29], gdyż

pasuje on dobrze do nieregularnych danych. Dalej, zauważmy, że n-ty moment tego rozkładu ma następującą formułę (patrz [30]):

$$\mathbb{E}(X^n) = \frac{\lambda^n \Gamma(n/\gamma + k)}{\Gamma(k)}.$$

Podstawiając $n = 1$ dostajemy:

$$\mathbb{E}(X) = \frac{\lambda \Gamma(1/\gamma + k)}{\Gamma(k)},$$

skąd już widać, że aby spełniony był wspomniany warunek: $\mathbb{E}(\varepsilon_i) = 1$, parametr λ musimy ustalić następująco:

$$\lambda = \frac{\Gamma(k)}{\Gamma(k + 1/\gamma)}. \tag{3.6}$$

Znaczna część literatury dotyczącej estymacji parametrów w modelu ACD(p, q) (patrz np. [75] lub [62]) skupia się na metodzie największej wiarygodności (MLE). Biorąc pod uwagę dane $\{d_i\}_{i=1}^n$ pochodzące z modelu ACD(p, q) o nieznanych parametrach $\theta = (\omega, a_1, ..., a_p, b_1, ..., b_q)$, wiarygodność θ równa jest gęstości łącznej z danych i ma postać:

$$f(d_n|\theta) = f(d_g|\theta) \cdot \prod_{i=g+1}^{n} f(d_i|d_{i-1}, \theta),$$

gdzie $g = \max\{p, q\}$. Wtedy warunkowa funkcja wiarygodności wygląda następująco:

$$\mathcal{L}(\theta|d_n) = \prod_{i=g+1}^{n} f(d_i|d_{i-1}, \theta).$$

Ponadto, warunkowa funkcja log - wiarygodności przyjmuje postać:

$$\ell(\theta|d_n) = \sum_{i=t_0+1}^{n} \log(f(d_i|d_{i-1}, \theta)),$$

gdzie $t_0 = g = \max\{p, q\}$.

Jeśli znana jest poprawna specyfikacja ε_i (jak również $f(d_i|d_{i-1}, \theta)$), to oszacowania warunkowej funkcji największej wiarygodności (ML) można otrzymać przez maksymalizację warunkowej funkcji log - wiarygodności.

W celu dopasowania parametrów dla powyższej klasy modeli ACD wykorzystałem właśnie metodę największej wiarygodności (MLE), która zaimplementowana została w bibliotece ACDm w języku R. Z tej biblioteki wykorzystałem funkcję acdFit() z odpowiednimi parametrami (tj. w celu dopasowania wybrałem odpowiedni model ACD, ustaliłem rozkład innowacji i ustaliłem parametr λ zgodnie z (3.6)). W poniższej tabelce zobaczyć możemy moje wyniki dopasowania.

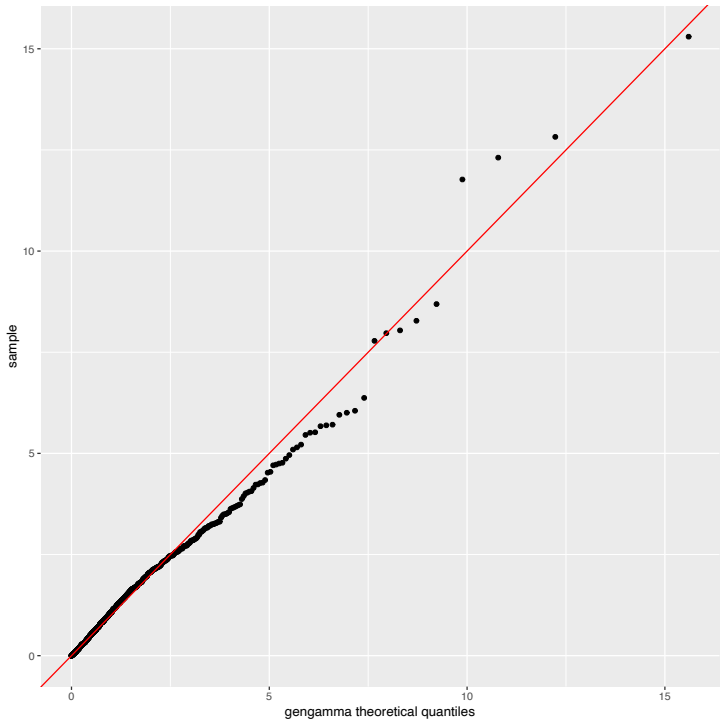
Model	ω	a_1	b_1	k	γ	**AIC**	**BIC**
ACD	0.015869	0.052606	0.942631	1.808068	0.584506	5171.133	5196.759
LACD_1	0.042547	0.041589	0.990978	1.939075	0.559538	5180.519	5206.146
LACD_2	-0.042365	0.045819	0.996123	1.821092	0.581990	5172.415	5198.041

Tablica 3.18: Wyniki dopasowań modelów ACD i log-ACD dla *interarrival times*

Patrząc od lewej strony tablicy 3.18 widzimy najpierw kolejne modele, czyli ACD, LACD_1 oraz LACD_2, a następnie dopasowane parametry modeli tj. ω, a_1 i b_1. Dalej mamy wyestymowane parametry rozkładu zmiennej losowej ε_i, a mianowicie k oraz γ. Przypomnę, że parametr λ rozkładu można otrzymać podstawiając obydwie wyżej wymienione oszacowane wartości do wzoru (3.6). W ostatnich dwóch kolumnach tabelki widzimy wartości kryteriów AIC oraz BIC. Intuicyjnie mówiąc kryteria te mierzą, jak dobrze dopasowywany model pasuje do obserwacji (tzn. czym mniejszą wartość ma kryterium, tym lepsze jest dopasowanie). Formalne definicje obydwu kryteriów przedstawiłem w rozdziale 3.2.2.

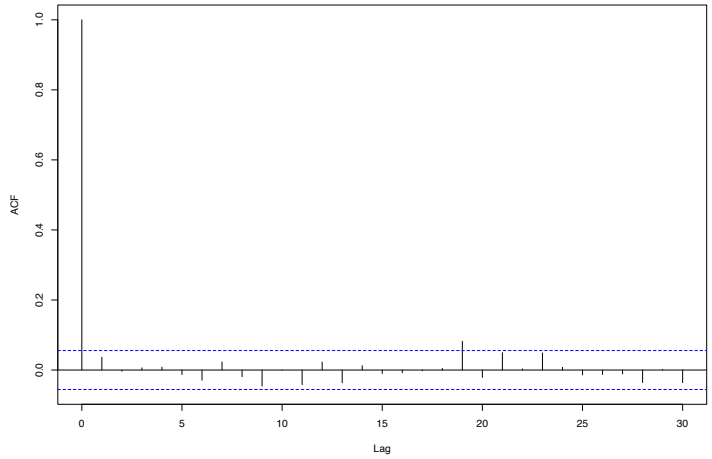
Najmniejszą wartość dla obydwu kryteriów możemy zaobserwować dla modelu ACD. Ta obserwacja nie pokrywa się z artykułem, gdyż patrząc na analogiczne wartości AIC i BIC autorzy zdecydowali się wybrać model LACD_1 (uzyskali oni, że właśnie dla tego modelu obydwie wartości AIC i BIC były najmniejsze).

Aby dodatkowo ocenić dokładność dopasowania modelu ACD, przyjrzyjmy się wykresowi na rysunku 3.9. Widzimy tu wykres kwantylowy (z ang. *qq-plot*) dla residuów modelu $\text{ACD}(1,1)$ z innowacjami o rozkładzie uogólnionym gamma (patrz (3.5)) dla dopasowanych residuów. Jest on wykresem rozrzutu utworzonym przez naniesienie na siebie dwóch zbiorów kwantyli. Jeśli wartości leżą wzdłuż czerwonej linii, to rozkład ma taki sam kształt, jak rozkład teoretyczny, który założyliśmy powyżej. Widzimy, że większość obserwacji, z wyjątkiem kilku po przeciwnych stronach prostej nachylonej pod kątem 45°, znajduje się wokół niej, co oznacza, że dopasowanie jest stosunkowo dokładne. Ta metoda jest tylko wizualnym sprawdzeniem, a nie niepodważalnym dowodem, jednak pozwala zobaczyć, czy założenie jest wiarygodne.



Rysunek 3.9: Wykres kwantylowy dla residuów modelu ACD, dla *interarrival times*.

W celu zbadania, czy proponowany model ACD jest wystarczający do uchwycenia zależności między czasami *interarrival*, przyjrzyjmy się kolejnemu wykresowi tj. rysunkowi 3.10 przedstawiającemu wykres próbkowej funkcji ACF dla residuów. Intuicyjnie mówiąc, ACF mierzy korelację pomiędzy obserwacjami we wcześniejszych okresach i obserwacjami w późniejszych okresach, bez pomijania obserwacji pomiędzy nimi. Funkcja ACF jest szeroko stosowana do wykrywania korelacji w szeregach czasowych. Z wykresu widać, że lagi nie mają istotnego wpływu (w granicach oznaczających przedziały ufności dla białego szumu - nie można ich odróżnić od zera), tzn. korelacje zostały usunięte. Ponieważ nie ma zależności autokorelacyjnej, to można wnioskować, że residua są białym szumem.



Rysunek 3.10: Wykres funkcji ACF dla residuów dla *interarrival times*, modelu ACD.

Przejdźmy do kolejnych wykorzystanych przeze mnie narzędzi. Poniżej przedstawiłem tablicę 3.19 z wynikami testów, które opisywałem szerzej w rozdz. 3.2. Z prawej strony tablicy widzimy testy statystyczne McLeoda - Li oraz Ljunga - Boxa. Intuicyjnie mówiąc testy te mierzą, czy istnieją jeszcze jakieś korelacje, które pozostały w residuach. Zauważmy, że przyjmując poziom ufności równy nawet $\alpha = 0.1$, to obydwie p - wartości dla wspomnianych testów są większe od α. Oznacza to, że nie ma pozostałych korelacji w residuach, oraz że proponowany model ACD może odpowiednio opisać szereg *interarrval times*.

Test	**test K-S**	**test A-D**	**test C-M**	**McLeod-Li**	**Ljung-Box**
p - wartość	0.435382	0.4465614	0.6975355	0.46029	0.31025

Tablica 3.19: P - wartości testów statystycznych przeprowadzonych na residuach dla modelu ACD$(1, 1)$.

W celu potwierdzenia założenia, że innowacje mają uogólniony rozkład gamma o gęstości danej wzorem (3.5), wykorzystałem testy Kołmogorowa - Smirnova, Andersona - Darlinga oraz Craméra-von Misesa formalnie zdefiniowane w rozdziale 3.2. Intuicyjnie mówiąc testy te sprawdzają, jak dobrze rozkład empiryczny danej próbki pasuje do zadanego rozkładu teoretycznego. Zauważmy, że test K-S skupia się na największym odchyleniu próbek od rozkładu teoretycznego, natomiast testy A-D oraz C-M biorą pod uwagę całkowite odchylenie.

W celu przeprowadzenia testów musiałem estymować parametry uogólnionego rozkładu gamma. Jest to rozkład 3 parametrowy, dlatego też zdecydowałem się wykorzystać metodę największej wiarygodności. Została ona zaimplementowana w bibliotece gamlss, z której wykorzy-

stałem funkcję fitDist() z parametrem extra ustawionym na GG2, co oznacza wybór rozkładu uogólnionego gamma.

Popatrzmy ponownie na tablicę 3.19. Od lewej strony widzimy kolejne p - wartości dla wspomnianych powyżej testów. Wynikowe p - wartości są tutaj duże (tj. znacznie większe od 0.1). W związku z tym nie mamy podstaw do odrzucenia hipotezy zerowej. Stąd wnioskujemy, że założenia modelu są spełnione.

Modelowanie *breach sizes*

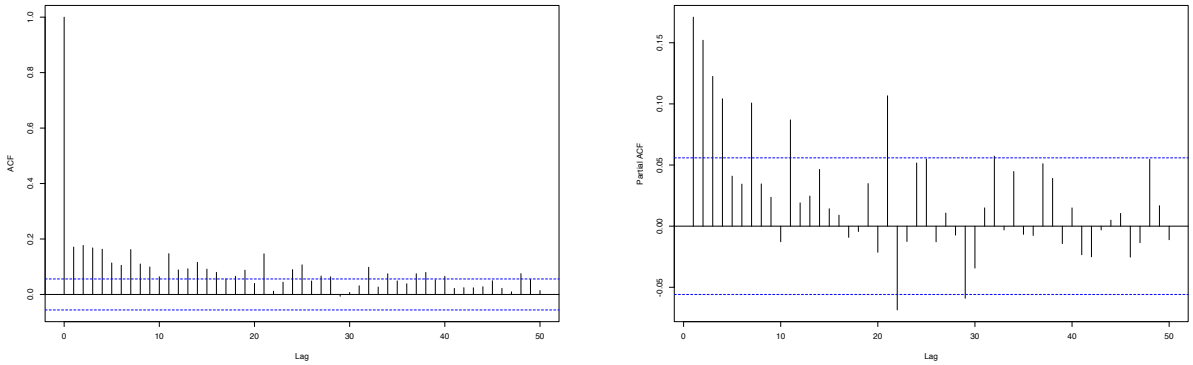
Kolejną wielkością, którą modelowałem, była zmienna ozn. rozmiar naruszeń (z ang. *breach sizes*). Popatrzmy na tablicę 3.20 zawierającą zestawienie podstawowych statystyk dotyczących rozmiarów naruszeń, dla typu naruszenia HACK[7]. Możemy w niej zauważyć zauważyć, że trzy kategorie biznesu (tj. BSF, BSO i BSR) mają znacznie większą średnią wielkość naruszenia niż pozostałe. Ponadto, analogicznie jak dla *interarrival times*, odchylenia standardowe dla każdej kategorii są dużo większe niż odpowiadające im średnie. Zauważmy jeszcze, że rozmiary naruszeń wykazują dużą zmienność i skośność (zob. też rysunek 3.7), na co wskazuje znaczna różnica między medianą, a wartością średnią. To sprawia, że trudno je modelować bez dokonywania przekształceń (por. rozdz. 3.3.1, rysunki 3.1 oraz 3.2), więc podobnie jak przy modelowaniu rozkładami, zastosowałem przekształcenie logarytmiczne na tej zmiennej.

Typ biznesu	Min	Median	Mean	SD	Max	Total
BSF	7	4113.5	4727952.24	19050450.8	130000000	82
BSO	2	8000.0	29307113.00	230120803.3	3000000000	181
BSR	12	2369.0	3820812.72	15696896.0	101600000	108
EDU	12	9000.0	44742.69	104594.2	800000	213
GOV	11	5700.0	435944.60	2440232.1	21500000	84
MED	1	3364.0	273006.54	3435882.7	78800000	554
NGO	13	4028.0	154452.64	636219.7	3000000	22
Wszystkie	1	4500.0	5068907.10	88437613.6	3000000000	1244

Tablica 3.20: Statystyki opisowe dla *breach sizes*, dla typu naruszenia HACK (kolejno: minimum, mediana, średnia, odchylenie standardowe, maksimum i liczba obserwacji).

[7]Podobnie jak dla zmiennej *interarrival times*, statystyki opisowe łącznie dla wszystkich typów naruszeń umieściłem w analogicznej tablicy, która znajduje się w dodatku A. na końcu mojej pracy.

Aby odpowiedzieć na pytanie, czy rozmiary naruszeń powinny być modelowane przez rozkład, czy przez proces stochastyczny, autorzy [24] ponownie proponują narysowanie wykresu próbkowych funkcji ACF i PACF (por. rys. 3.8) dla zlogarytmowanych wielkości *breach sizes*. Można tu zaobserwować, że istnieją widoczne korelacje pomiędzy rozmiarami naruszeń (por. z opisem rys. 3.8). W związku z tym autorzy [24] sugerują, że do modelowania powinniśmy użyć procesu stochastycznego, a nie rozkładu.



Rysunek 3.11: Wykresy próbkowych funkcji ACF i PACF odpowiednio, dla zlogarytmowanych wielkości *breach sizes* (dla typu naruszenia HACK).

Do modelowania zmiennej *breach sizes* wykorzystałem proces *AutoRegressive Moving Average*, w skrócie ARMA(p, q), gdzie p to rząd procesu AR(p), natomiast q to rząd procesu MA(q). Został on przeze mnie wyselekcjonowany w myśl publikacji [24, 45, 61]. Proces ARMA(p, q) to model szeregów czasowych, który stosuje się do modelowania szeregów stacjonarnych lub niestacjonarnych. Najprościej mówiąc, szeregi stacjonarne to takie, w których występują jedynie wahania losowe wokół średniej. Dla szeregów niestacjonarnych trzeba wykorzystać odpowiednią metodę, by sprowadzić je do stacjonarnych - dopiero wtedy można wykorzystać ten proces. Budowa procesu ARMA, którego definicję formalną podam za chwilę, opiera się na zjawisku autokorelacji, czyli na korelacji zmiennej prognozowanej z nią samą, ale opóźnioną w czasie.
Modele ARMA są użytecznym narzędziem prognostycznym. Przed estymacją parametrów należy ustalić rzędy p i q procesów. Jeśli procesy są prawidłowo ustalone, to residua będą białym szumem.

Dodatkowo, do modelowania zmienności w *breach sizes*, autorzy np. [24] zdecydowali się wybrać model GARCH(p, q) (ang. *Generalized AutoRegressive Conditional Heteroskedasticity*). Jest to model nieliniowy.
GARCH jest jednym z najpopularniejszych modeli do opisu i prognozowania zmienności. Wynika to z faktu, że pozwala on opisać większość empirycznych własności np. finansowych szeregów

czasowych. Ważna jest dla niego również łatwość rozszerzania modeli, a także łatwość estymacji jego parametrów.

Analiza w [24] przeprowadzona przez autorów na residuach sugeruje, że proces GARCH$(1,1)$ jest wystarczający do opisu zmienności residuów. Opierają się oni głównie na publikacji Hansen, Lunde [31] z której wynika, że modele GARCH wyższego rzędu niekoniecznie są lepsze niż zwykły GARCH$(1,1)$.

Powyższe prowadzi nas do następującego modelu ARMA(p,q)-GARCH(p,q):

$$Y_t = \mathbb{E}(Y_t|\mathcal{F}_{t-1}) + \varepsilon_t = \mu_t + \varepsilon_t, \tag{3.7}$$

gdzie $\mathcal{F}_{t-1}$ reprezentuje wiedzę zgromadzoną do momentu $t-1$, natomiast ε_t to innowacje - niezależne zmienne losowe o tym samym rozkładzie (*i.i.d.*).

Ogólna postać procesu ARMA(p,q) ze średnią μ:

$$Y_t = \mu + \sum_{k=1}^{p} \phi_k Y_{t-k} + \sum_{l=1}^{q} \theta_l \varepsilon_{t-l} + \varepsilon_t, \tag{3.8}$$

gdzie $\varepsilon_t = \sigma_t \cdot Z_t$; Z_t to innowacje *i.i.d.*; a odpowiednio ϕ_k i $\theta_l \in \mathbb{R}$ to współczynniki procesów AR(p) i MA(q).

Dwa szczególne przypadki, czyli procesy AR(p) i MA(q) mają następujące postaci:

1. Proces ARMA$(p,0)$, czyli AR(p):

$$Y_t = \mu + \sum_{k=1}^{p} \phi_k Y_{t-k} + \varepsilon_t,$$

2. Proces ARMA$(0,q)$, czyli MA(q):

$$Y_t = \mu + \sum_{l=1}^{q} \theta_l \varepsilon_{t-l} + \varepsilon_t.$$

Formalnie $(\varepsilon_t)_{t\in\mathbb{Z}}$ nazywamy procesem GARCH(p,q) gdy:

$$\sigma_t^2 = \mathrm{Var}(\varepsilon_t|\varepsilon_{t-s}, \sigma_{t-s}^2, s>0) = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

gdzie σ_t^2 jest wariancją warunkową; $\varepsilon_t = \sigma_t \cdot Z_t$ – innowacje; $p, q \in \mathbb{N}$ to rzędy procesu; $\alpha_i, \beta_j \geqslant 0$; a ω to wyraz wolny.

Zatem szczególna postać modelu GARCH, to znaczy GARCH$(1,1)$ wygląda następująco:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \tag{3.9}$$

Wtedy równ. 3.7 możemy zapisać w następującej postaci:

$$Y_t = \mathbb{E}(Y_t|\mathcal{F}_{t-1}) + \varepsilon_t = \mu_t + \sigma_t \cdot Z_t, \tag{3.10}$$

gdzie μ_t i σ_t^2 to zdefiniowane powyżej odpowiednio średnia warunkowa i wariancja warunkowa, a Z_t to zmienna *i.i.d.*

Do wyboru odpowiedniego modelu autorzy ponownie wykorzystują kryterium AIC i BIC. Wspominają, że kilka modeli ARMA(p,q)-GARCH$(1,1)$ miało podobne wartości dla $p,q \in [0,5]$ naturalnych. Zdecydowali się więc wybrać prostszy model tj. ARMA$(1,1)$-GARCH$(1,1)$, z innowacjami o rozkładzie normalnym.

Powyższe wyniki pokrywają się z przeprowadzoną przeze mnie analizą.
Do wyboru odpowiedniego modelu ARMA-GARCH wykorzystałem obszerną bibliotekę rugarch, zawierającą implementacje niezbędnych funkcji. Dokładniej, funkcje, które w tym celu wykorzystałem to:

1. ugarchspec() - służy ona do ustalenia specyfikacji modelu. Najważniejsze parametry tej funkcji, to mean.model, który odpowiada współczynnikom (p,q) procesu ARMA oraz variance.model, który odpowiada współczynnikom modelu GARCH(p,q). Dodatkowo ustalić w niej można rozkład innowacji. Domyślnie ustawionym rozkładem jest rozkład normalny;

2. ugarchfit() - służy ona do dopasowywania różnych modeli GARCH. Najważniejszym parametrem tej funkcji jest spec, czyli specyfikacja modelu zwracana przez powyższą funkcję ugarchspec().

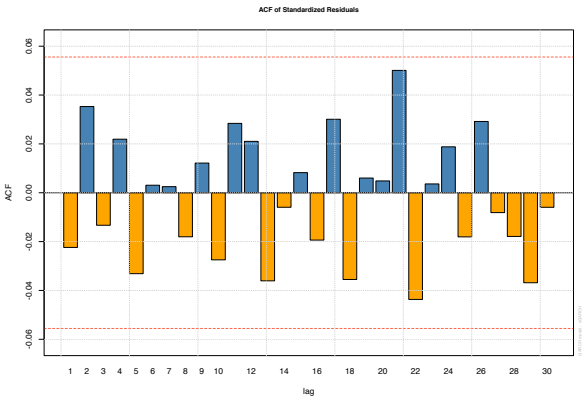
Następnie, patrząc na wartości kryteriów Akaike, Bayesa, oraz Hannana-Quinna (w skrócie HQIC), które zdefiniowałem w rozdziale 3.2.2, próbowałem dopasować odpowiedni model ARMA(p,q)-GARCH$(1,1)$ dla $p,q \in \{0,1,2,3,4,5\}$. W tablicy 3.21 podałem znalezione wartości wspomnianych kryteriów.

Model	**AIC**	**BIC**	**HQIC**
ARMA$(1,1)$-GARCH$(1,1)$	5.1748	**5.1995**	5.1841
ARMA$(1,2)$-GARCH$(1,1)$	5.1760	5.2049	5.1869
ARMA$(2,1)$-GARCH$(1,1)$	5.1761	5.2050	5.1870
ARMA$(2,2)$-GARCH$(1,1)$	5.1763	5.2093	5.1887
ARMA$(0,0)$-GARCH$(1,1)$	5.2451	5.2616	5.2513
ARMA$(4,3)$-GARCH$(1,1)$	**5.1595**	5.2048	**5.1766**

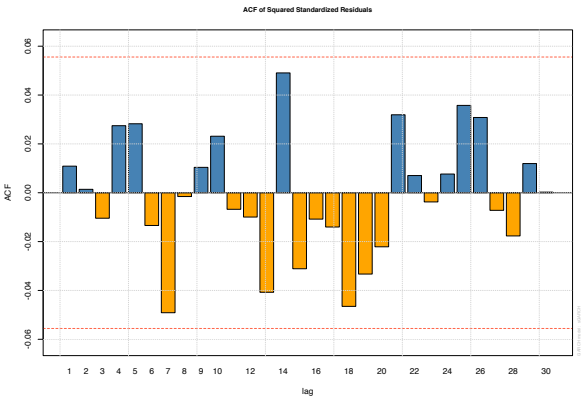
Tablica 3.21: Wartości kryteriów AIC, BIC i HQIC dla różnych modeli ARMA(p,q)-GARCH$(1,1)$, dla $p,q \in \{0,1,2,3,4,5\}$

W pierwszej kolumnie widzimy model ARMA(p, q)-GARCH dopasowany z wybranymi zestawami parametrów. Kolorem czerwonym oznaczone są najmniejsze wartości kryteriów. Ostatni wiersz tabeli nie jest przypadkowy. To właśnie dla niego wartości kryteriów AIC i HQIC były najniższe dla $p, q \in \{0, 1, 2, 3, 4, 5\}$. Wartość kryterium BIC okazała się najniższa dla podstawowego modelu ARMA$(1, 1)$-GARCH$(1, 1)$. Okazuje się, że różnice pomiędzy nimi są bardzo małe. Zatem aby nie brać modelu ze zbyt dużą liczbą parametrów oraz patrząc na powyższe wartości kryteriów informacyjnych, zdecydowałem się wybrać ten sam model co autorzy artykułu [24] tj. ARMA$(1, 1)$-GARCH$(1, 1)$, z innowacjami o rozkładzie normalnym.

W celu dalszej oceny dopasowania tego modelu, na rysunku 3.12 przedstawiłem wykres próbkowej funkcji ACF dla standaryzowanych residuów, a na rysunku 3.13 wykres kwadratowych standaryzowanych residuów (wykonane one zostały przy użyciu biblioteki rugarch). Patrząc na obydwa rysunki łatwo można zobaczyć, że korelacje zostały usunięte.



Rysunek 3.12: Wykres próbkowej funkcji ACF dla standaryzowanych residuów, dla *breach sizes*.



Rysunek 3.13: Wykres próbkowej funkcji ACF dla kwadratowych standaryzowanych residuów, dla *breach sizes*.

Następnie wykonałem testy Ljunga-Boxa na standaryzowanych residuach oraz kwadratowych standaryzowanych residuach. Otrzymane p - wartości znajdują się w tablicy 3.22. Widzimy, że obydwie wartości znacznie przekraczają poziom ufności równy nawet $\alpha = 0.1$. Oznacza to, że nie ma podstaw, by odrzucić hipotezę zerową, która mówi, że w residuach nie występują korelacje. Uzyskane wyniki zgadzają się z wynikami publikacji [24].

Test	**Ljung - Box na std. res.**	**Ljung - Box na kwadr. std. res.**
p - wartość	0.4296	0.7006

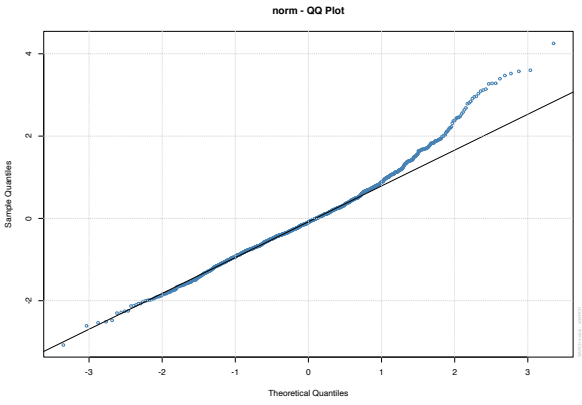
Tablica 3.22: P - wartości testów Ljunga - Boxa na standaryzowanych residuach i kwadratowych standaryzowanych residuach.

Dalej, w tablicy 3.23 widzimy wyniki dopasowania modelu ARMA$(1,1)$-GARCH$(1,1)$. Patrząc na oszacowane odchylenia standardowe obserwujemy, że wszystkie oszacowane współczynniki dla części ARMA i GARCH $\pm$ odchylenie standardowe są różne od zera.

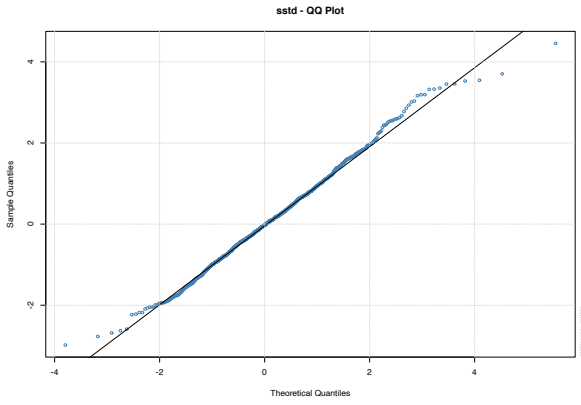
Parametry	μ	ϕ_1	θ_1	ω	α_1	β_1
Est.	8.806370	0.949907	-0.857678	0.334865	0.038426	0.930318
Odch. std.	0.251761	0.018719	0.031325	0.194207	0.013240	0.028656

Tablica 3.23: Wyniki dopasowania modelu ARMA(1,1)-GARCH(1,1) dla *breach sizes*, gdzie Odch. std. oznacza szacowane odchylenie standardowe.

Z powyższej analizy wiemy już, że model ARMA$(1,1)$-GARCH$(1,1)$ może generalnie pasować do modelowania zmiennej *breach sizes*. Zobaczmy teraz, jak ten model dopasowuje się w ogonach. Na kolejnym rysunku przedstawiłem wykres kwantylowy dla residuów modelu ARMA$(1,1)$-GARCH$(1,1)$ z innowacjami o rozkładzie normalnym (rysunek 3.14). Z kolei na rysunku 3.15 z innowacjami o rozkładzie skośnym T-studenta.



Rysunek 3.14: Wykres kwantylowy dla residuów modelu ARMA$(1,1)$-GARCH$(1,1)$ z innowacjami o rozkładzie normalnym dla zlogarytmowanych *breach sizes*.



Rysunek 3.15: Wykres kwantylowy dla residuów modelu ARMA$(1,1)$-GARCH$(1,1)$ z innowacjami o rozkładzie skośnym T-studenta dla zlogarytmowanych *breach sizes*.

Widać niestety, że ogony są zbyt ciężkie dla innowacji normalnych. Można jednak stwierdzić, że rozkład skośny T-studenta zapewnia już stosunkowo dokładne dopasowanie w ogonach. Zatem w dalszej części pracy będę korzystał z tego modelu z innowacjami o rozkładzie skośnym T-studenta. Ta obserwacja nie pokrywa się z wynikami artykułu [24], w którym autorzy zdecydowali się wybrać inny rozkład, ponieważ dla niego właśnie dopasowanie było najlepsze.

Zależność pomiędzy zmiennymi *interarrival times*, a *breach sizes*

Zgodnie z podejściem przedstawionym w Xu i inni [24], przeprowadziłem analizę na residuach otrzymanych po dopasowaniu szeregów czasowych do zmiennych *interarrival times* oraz *breach sizes*. W celu zbadania zależności obliczyłem współczynniki korelacji ρ Spearmana oraz τ Kendala dla reszt z modeli ACD i ARMA-GARCH. Wynoszą one odpowiednio -0.01869585 oraz -0.01219713. Widzimy, że obie wartości są bardzo bliskie 0, co może sugerować brak zależności. Intuicję tę potwierdza test hipotezy istotności współczynnika korelacji. Dokładnej, dla hipotez postaci $H_0 : \rho = 0$, przeciwko $H_1 : \rho \neq 0$ zarówno dla Spearmana, jak i Kendalla (których formalne definicje przedstawiłem w podrozdziale 3.2.2) uzyskane p - wartości nie dały podstaw do odrzucenia hipotezy zerowej. Były one odpowiednio równe 0.5101 i 0.5195. Okazało się więc, że, w przeciwieństwie do wyników z pracy Xu i inni [24], uzyskane przeze mnie wyniki nie wskazują na istnienie zależności pomiędzy *interarrival times*, a *breach sizes*.

W celu modelowania dwuwymiarowej zależności pomiędzy *interarrival times* incydentów, a *breach sizes*, autorzy publikacji [24, 60] proponują zastosowanie teorii kopuł. Mimo przesłanek świadczących o braku zależności, przeprowadziłem odpowiednią analizę. Zbadałem zarówno podejście parametryczne, jak i nieparametryczne do dopasowania kopuły, estymacji jej parametrów i parametrów rozkładów brzegowych. Najprostszym podejściem do szacowania jest przedstawione poniżej podejście dwuetapowe (patrz [24, 60]).

- Na początku szacujemy parametry rozkładów brzegowych, oddzielnie dla każdej zmiennej, a następnie wyciągamy standaryzowane residua r_{it}.
  - W podejściu parametrycznym zakłada się, że rozkład standaryzowanych residuów jest znany dla każdego rozkładu brzegowego. Niech $F_i(\cdot|\hat{\theta}_m)$ oznacza ten rozkład, gdzie $\hat{\theta}_i^m$ są oszacowanymi parametrami brzegowymi dla rozkładu brzegowego i. Dalej, definiujemy przekształcenie $u_{it} = F^{-1}(r_{it}|\hat{\theta}_i^m)$.
  - W podejściu nieparametrycznym zamiast $F_i(\cdot|\hat{\theta}_m)$ stosuje się dystrybuantę empiryczną $\{r_{it} : t = 1, ..., T\}$.
- W następnym kroku dane $u_t = (u_{1t}, ..., u_{dt})$, $t = 1, ..., T$ służą do oszacowania parametrów

kopuł przy użyciu jednej z metod estymacji np. przedstawionych w [60].

Podejście parametryczne

W podejściu parametrycznym rozpocząłem od dobrania do danych odpowiedniego modelu. W podrozdziale 3.3.2 stwierdziłem, że najlepszym wyborem dla zmiennej *breach sizes* okazał się model ARMA$(1,1)$-GARCH$(1,1)$, zaś dla zmiennej *interarrival times* model ACD$(1,1)$. Następnie, standaryzowane residua, które otrzymałem na podstawie tych modeli, stanowią próbkę *i.i.d.* Dalej, wzorując się na [24], dla obydwu modeli zastosowałem następujące przekształcenia:

- Dla residuów modelu ACD$(1,1)$, które oznaczyłem przez $e_1, ..., e_n$, zastosowałem uogólniony rozkład gamma $\Gamma(\cdot|\gamma, k)$ (patrz równ. (3.5)) z dopasowanymi parametrami:

$$e_i \to \Gamma(e_i|\gamma, k), \quad i = 1, ..., n.$$

- Dla residuów modelu ARMA$(1,1)$-GARCH$(1,1)$ wykonałem przekształcenie analogiczne do powyższego, jednakże tym razem dla rozkładu skośnego T-studenta z dopasowanymi parametrami. Było to oczywiście podyktowane wyborem rozkładu skośnego T-studenta jako rozkładu innowacji w modelu ARMA$(1,1)$-GARCH$(1,1)$.

Następnie do tak przygotowanych danych spróbowałem dopasować odpowiednią kopułę, tzn. dopasować ją do zmiennych dwuwymiarowych (por. [24, 60]). Wspomniana powyżej dwuetapowa metoda została zaimplementowana dla wielu znanych typów kopuł w bibliotece VineCopula, języka R. Wykorzystałem ją i otrzymałem wyniki, które umieściłem w tablicy 3.24. Widać w niej moje próby dopasowania popularnych w literaturze kopuł (patrz [24, 60, 61, 65, 75]).

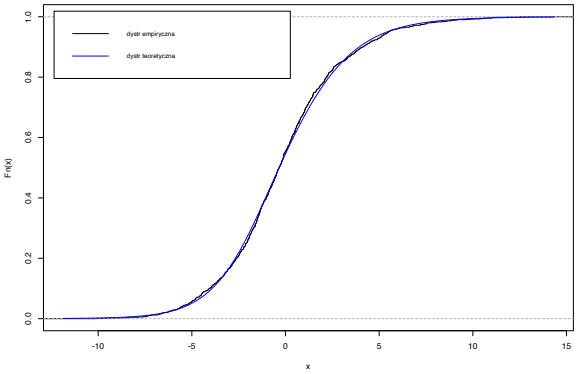
Model	**log-lik.**	**AIC**	**BIC**	**τ Kendalla**
Independence	0.00000	**0.00000**	**0.00000**	0.00000
Gaussian	0.00884	1.98231	7.10759	-0.00243
Gumbel	-0.00214	2.00428	7.12957	0.00015
Clayton	0.00914	1.98172	7.10701	0.00196
Joe	-0.00085	2.00170	7.12698	0.00010
BB6	-0.02499	4.04998	14.30054	0.00158
BB8	0.00010	3.99980	14.25036	0.00001
Tawn	1.35228	1.29544	11.54601	0.01982

Tablica 3.24: Wyniki dopasowania wybranych kopuł (podejście parametryczne).

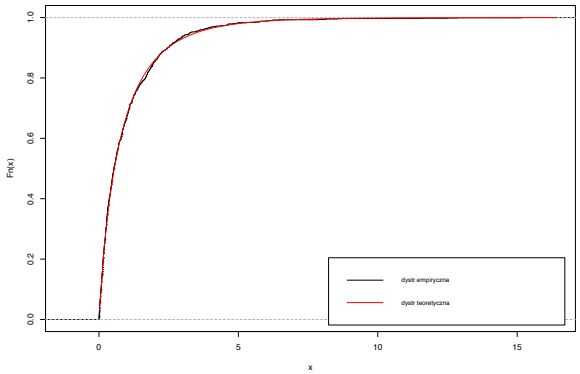
W tablicy 3.24 kolejne kolumny odpowiadają: typowi dopasowywanej kopuły, wartości logarytmicznej funkcji wiarygodności, kryteriom AIC i BIC oraz wartości współczynnika korelacji τ Kendalla. Próbowałem tu dopasować następujące typy kopuł: Gaussa, Gumbela, Claytona, Joe'a, BB6, BB8, Tawn (opisane przeze mnie w podrozdziale 2.3.1), a także rozpatrywałem niezależność analizowanych dwóch zmiennych. Zauważmy, że najmniejsze wartości kryteriów AIC i BIC (oznaczone w tablicy 3.24 kolorem czerwonym) ma model *Independence*. Potwierdza to poprzednio uzyskane przeze mnie wnioski o braku zależności obydwu zmiennych.

Podejście nieparametryczne

W podejściu nieparametrycznym przekształciłem residua wykorzystując w tym celu dystrybuantę empiryczną. Rysunek 3.16 przedstawia porównanie dystrybuanty empirycznej oraz założonej dystrybuanty teoretycznej (z wyestymowanymi parametrami) dla obydwu zmiennych. Obserwujemy tu bardzo dużą zgodność.



Rysunek 3.16: Wykres dystrybuanty empirycznej i założonej dystrybuanty teoretycznej wyznaczonej dla reszt z modelu ARMA$(1,1)$ - GARCH$(1,1)$ dla zmiennej *breach sizes*.



Rysunek 3.17: Wykres dystrybuanty empirycznej i założonej dystrybuanty teoretycznej wyznaczonej dla reszt z modelu ACD$(1,1)$ dla zmiennej *interarrival times*.

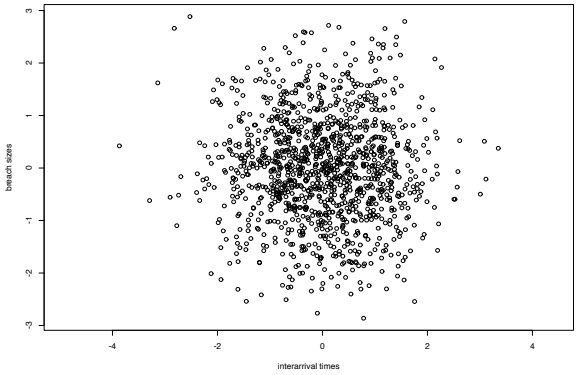
Następnie tak przekształcone residua wykorzystałem do dopasowania kopuł (por. tablica 3.24). Wyniki przedstawia tablica 3.25. Zauważmy, że ponownie najmniejsze wartości kryteriów AIC i BIC (oznaczone w tej tablicy kolorem czerwonym) ma model *Independence*. Zatem kolejny raz otrzymujemy wnioski o braku zależności obydwu zmiennych.

Model	**log-lik.**	**AIC**	**BIC**	**τ Kendalla**
Independence	0.00000	**0.00000**	**0.00000**	0.00000
Gaussian	0.07987	1.840254	6.965538	0.00728
Gumbel	-0.00323	2.006477	7.131760	0.00015
Clayton	0.01052	1.978948	7.104231	0.00211
Joe	-0.00382	2.007649	7.132932	0.00009
BB6	-0.04988	4.099779	14.350345	0.00157
BB8	0.00008	3.999838	14.250404	0.00001
Tawn	-0.00002	4.000019	14.250586	0.00001

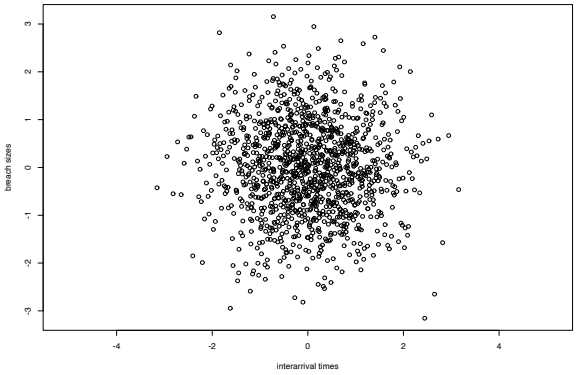
Tablica 3.25: Wyniki dopasowania wybranych kopuł (podejście nieparametryczne).

W celu określenia zależności można stosować także przekształcenie normalne (z ang. *normal score transformation*, patrz [24]). Dokładniej, zmienne przygotowane zgodnie z opisem na str. 90

należy przekształcić przez odwrotność dystrybuanty rozkładu standardowego normalnego. Rysunki 3.18 oraz 3.19 przedstawiają wykresy *normal scores* (czyli wykresy rozproszenia innowacji po wspomnianych przekształceniach) zarówno w podejściu parametrycznym jak i nieparametrycznym. Porównując je z wykresami konturowymi dla wybranych kopuł, przedstawionych na rys. 2.4, widzimy bardzo duże podobieństwo do kopuły *Independence*, co wskazuje na niezależność zmiennych.



Rysunek 3.18: Wykres *normal scores* w podejściu parametrycznym.



Rysunek 3.19: Wykres *normal scores* w podejściu nieparametrycznym.

Wnioski z analizy na wszystkich typach

Podobne analizy przeprowadziłem na posiadanych danych, dla wszystkich typów naruszeń. Poniżej przedstawiam skrócone wnioski z moich badań:

1. Dla zmiennej *interarrival times* modelem, który najczęściej pasował do danych, był ACD$(1,1)$. Właśnie przy tym modelu wartości kryteriów AIC i BIC były najmniejsze dla typów HACK, INSD, PORT i UNKN. Z kolei dla typów DISC, STAT oraz PHYS wartości tych kryteriów były najmniejsze dla modelu LACD_2. Jednakże ponieważ różnica między modelem ACD, a LACD_2 była bardzo mała, to zdecydowałem się wybrać prostszy model. Ponadto dla typu CARD najlepszym modelem okazał się LACD_1. Należy jednak pamiętać, że dla tego właśnie typu dostępne są tylko 32 obserwacje.

2. W celu sprawdzenia zależności pomiędzy zmiennymi *interarrival times*, a *breach sizes* (jak w analizie na typie HACK powyżej), obliczyłem empiryczne współczynniki korelacji ρ Spearmana oraz τ Kendalla pomiędzy obydwiema zmiennymi. Dla każdego typu naruszenia, ich wartości prowadziły do tych samych wniosków. Podobnie, nieparametryczne testy rangowe, zarówno dla Spearmana, jak i Kendalla, nie dały podstaw do odrzucenia hipotezy zerowej. Okazało się więc, że wyniki dla każdego z typów naruszeń nie wskazują na istnie-

nie zależności pomiędzy analizowanymi zmiennymi. Wyniki mojej analizy przedstawiłem w tablicy 3.26.

3. Dla każdego typu naruszenia wykresy *normal scores* były podobne do wykresów przedstawionych na rysunkach 3.18 oraz 3.19. Zatem, podobnie jak w powyższym podpunkcie, wyniki dla wszystkich typów naruszeń wyszły identycznie.

Typ breach	**ρ Spearman**	**τ Kendall**	**p-wart. testu Sp**	**p-wart. testu Ke**
CARD	0.10788177	0.04926108	0.57608577	0.72392020
HACK	-0.00706274	-0.00472209	0.80352048	0.80307176
INSD	0.00843042	0.00681137	0.87259068	0.84617901
PHYS	-0.02383878	-0.01594890	0.38287907	0.38143762
PORT	0.03902938	0.02610624	0.24955054	0.24848150
STAT	-0.10832557	-0.07079805	0.14427565	0.15473904
DISC	-0.01731521	-0.01191413	0.53863823	0.52592405
UNKN	-0.10689291	-0.07564822	0.32376552	0.29950377

Tablica 3.26: Wyniki dopasowania dla poszczególnych typów naruszeń, dla punktu 2 powyższych wniosków (kolejno: współczynnik korelacji ρ Spearmana i τ Kednala oraz p - wartości nieparametrycznych testów rangowych dla Spearmana i Kendalla).

Porównanie z artykułem

Na koniec, celem porównania moich wyników z wynikami pracy Xu i inni [24], wrócę do krótkiego streszczenia ich analizy. Jako rozkład innowacji, zamiast rozkładu skośnego T - Studenta, zdecydowali się wybrać tzw. rozkład *mixed extreme value distribution* (czyli mieszankę rozkładu normalnego, z dwoma rozkładami GPD w ogonach). Dalej, do modelowania zależności pomiędzy zmiennymi *interarrival times*, a *breach sizes* wybrali kopułę Gumbela. Swój wybór potwierdzili wykorzystaniem kryteriów informacyjnych AIC i BIC – dla tej właśnie kopuły ich wartości były najmniejsze. Dodatkowo, przedstawili porównanie wykresu *normal scores* (por. rys. 3.18 i 3.19 z rys. 2.4) z wykresem konturowym dla kopuły Gumbela, skąd wyciągnęli wnioski, że dopasowanie jest dokładne. Na końcu przeprowadzili test zgodności Cramera-von Misesa, którego wyniki wskazują, że zależność można modelować za pomocą kopuły Gumbela. Według autorów [24], zależność ta oznacza, że jeśli istnieje długi okres czasu, w którym nie pojawiają się żadne incydenty hakerskie, to bardziej prawdopodobne jest, że w momencie wystąpienia incydentu dojdzie

do dużego naruszenia hakerskiego. Potwierdzenia takich wniosków nie zaobserwowałem jednak w swojej analizie.

3.3.3. Wartość narażona na ryzyko i predykcja

Przejdę teraz do kolejnej części mojej pracy, która będzie odnosiła się do podpunktów 3 oraz 4. schematu 2.2.
We wcześniejszych rozdziałach próbowałem wybrane rozkłady dopasować do danych lub modelować przy pomocy wyselekcjonowanych procesów stochastycznych. W przypadku modelowania procesami stochastycznymi, do zmiennej *breach sizes* zdecydowałem się dobrać odpowiedni model ARMA$(1,1)$-GARCH$(1,1)$. Z kolei do zmiennej *interarrival times* dopasowałem adekwatny model ACD$(1,1)$. Po dopasowaniu do danych wybranych modeli, zbadałem jak prognozować przyszłe trajektorie szeregu, a także jak obliczać i prognozować wartości narażone na ryzyko (które opisałem w poniższym podrozdziale). Następnie przedstawiłem uzyskane przeze mnie wyniki w odpowiednich tabelkach i na wykresach. Dodatkowo skupiłem się na ocenie dokładności otrzymanej predykcji szeregu, a także na predykcji wartości narażonej na ryzyko. Rozdział ten zakończyłem wnioskami z przeprowadzonej analizy.

Wartość narażona na ryzyko

Zanim jednak przejdę do prognozy dot. wspomnianych zmiennych, rozpocznę od wprowadzenia potrzebnych mi definicji oraz odnoszących się do nich prostych przykładów (zob. [24, 45, 61, 77]). Następnie wymienię najbardziej popularne metody liczenia wartości narażonej na ryzyko, a także opiszę tę, którą wykorzystałem w swojej analizie.
Poniższa definicja mówi, czym jest wartość narażona na ryzyko (patrz Jakubowski [77]):

Definicja 3.5 (wartość narażona na ryzyko). Dla każdego $\alpha \in (0,1)$ wartość narażona na ryzyko (z ang. *Value at Risk*, w skrócie VaR), ozn. $\mathrm{VaR}_\alpha(X)$, na poziomie tolerancji α dla zmiennej losowej X definiowana jest wzorem:

$$\mathrm{VaR}_\alpha(X) = -\sup\{x \in \mathbb{R} : \mathbb{P}(X < x) \leqslant \alpha\}. \tag{3.11}$$

Wartość narażona na ryzyko wyrażona jest w języku kwantyli, więc $\mathrm{VaR}_\alpha(X)$ można zapisać przy pomocy kwantyli:

$$\mathrm{VaR}_\alpha(X) = -q_\alpha^+(X) = q_{1-\alpha}^-(-X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X + x < 0) \leqslant \alpha\}, \tag{3.12}$$

gdzie $q_\alpha^+(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leqslant x) > \alpha\}$ to górny α - kwantyl, natomiast $q_\alpha^-(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leqslant x) \geqslant \alpha\}$ to dolny α - kwantyl.

Ponadto, gdyby na VaR spojrzeć z pozycji strat Y, tzn. $Y = -X$, to ze wzoru (3.12) dostaniemy, że $\mathrm{VaR}_\alpha(Y) = q^-_{1-\alpha}(X)$, więc VaR_α dla strat Y to po prostu dolny $(1-\alpha)$ - kwantyl zmiennej losowej X.
Poniżej przedstawiłem przykład, który mówi jak interpretować wartość narażoną na ryzyko VaR.

Przykład 3.6 (Interpretacja słowna). Przykładowe interpretacje wartości narażonej na ryzyko to:

- wielkość $\mathrm{VaR}_{0.95}$ oznacza, że istnieje tylko 5 procentowe prawdopodobieństwo, że obserwowana wartość jest większa niż przewidywana wartość $\mathrm{VaR}_{0.95}$;
- wielkość $\mathrm{VaR}_{0.99}$ w horyzoncie rocznym można opisać: "tylko raz na 100 lat strata będzie większa lub równa niż kwota X";
- mówiąc językiem finansowym, wielkość VaR_α to najmniejsza dodatkowa ilość kapitału, którą trzeba zainwestować w instrument (bez ryzyka), aby prawdopodobieństwo straty z pozycji X było poniżej poziomu tolerancji α (patrz [77]).

Value at Risk jest wygodną i praktyczną miarą ryzyka, ale należy pamiętać, że jest ona tylko statystyką. Zatem nadal możliwe jest, że rzeczywiste przyszłe straty będą dużo wyższe niż poziom określony przez VaR nawet przy poziomie tolerancji równym 0.99.

Podstawowe metody wyznaczania wartości VaR to:

- metoda nieparametryczna (np. symulacja historyczna);
- metoda parametryczna (np. tzw. podejście wariancji – kowariancji);
- metoda numeryczna (np. symulacja metodą Monte – Carlo).

Każda z tych metod ma swoje wady i zalety. Szacowanie VaR jest problemem, który nie ma swojego uniwersalnego rozwiązania.

Najprostsza z tych metod to symulacja historyczna. Jest ona łatwa również w implementacji. Wystarczy tylko znaleźć rozkład empiryczny i na jego podstawie wyznaczyć VaR z definicji. Zaletą tej metody jest fakt, że to podejście nieparametryczne. Nie trzeba więc estymować parametrów rozkładów na podstawie danych historycznych. Poniżej przedstawiłem prosty przykład zastosowania tej metody.

Przykład 3.7 (Symulacja historyczna). Załóżmy, że chcemy obliczyć jednodniowy VaR na poziomie tolerancji $\alpha = 95\%$, przy użyciu 100 – dniowych danych. Dane wartości należy posor-

tować od najmniejszej do największej. Wtedy wartość narażona na ryzyko to po prostu piąta najmniejsza wartość odpowiadająca piątemu dniu, ponieważ wykorzystujemy 100 dni danych.

Z uwagi na to, że w poprzednich podrozdziałach dopasowałem już modele do danych, wyestymowałem ich parametry oraz ustaliłem rozkład innowacji, to zdecydowałem się wykorzystać trzecią metodę, tj. symulację metodą Monte – Carlo (wybraną także przez [24, 48, 63]).
Ogólnie, przy symulacjach Monte – Carlo, przyjmuje się pewien model, który najlepiej opisuje zachowanie kształtowania się np. cen pewnych instrumentów finansowych. Dalej, należy wygenerować dużą ilość (np. 10000, patrz [24] lub [77]) kolejnych obserwacji, otrzymując w ten sposób rozkład empiryczny. Następnie, wyznaczenie kwantyla tego rozkładu umożliwia wyliczenie VaR wprost z definicji (patrz równ. (3.12)).

Kroki algorytmu szacowania VaR metodą Monte – Carlo mogą być trochę inne w zależności od np. wybranego modelu lub analizowanych danych. Jak już wspominałem wcześniej, do danych zdecydowałem się dopasować modele ACD$(1,1)$ oraz ARMA$(1,1)$-GARCH$(1,1)$. Wybrana przeze mnie metoda symulacji dla modelu ACD$(1,1)$, którą wykonałem w oparciu o publikacje [24, 48, 63], opiera się na następujących krokach:

1. Dopasowanie wybranego modelu ACD$(1,1)$.

2. Wyznaczenie standaryzowanych residuów z modelu i estymacja na ich podstawie parametrów odpowiedniego rozkładu (w moich badaniach był to uogólniony rozkład gamma; do estymacji wykorzystałem funkcję fitGGammaDist z biblioteki MethylIT; w ten sposób otrzymałem estymowane parametry k, γ oraz λ tego rozkładu;)

3. Wygenerowanie $N = 10000$ losowych wartości z rozkładu uogólnionego gamma z ustalonymi parametrami, które zostały wyestymowane w 1. kroku. W tym celu można użyć np. funkcji rggamma() z tej samej biblioteki;

4. Obliczenie 10000 przewidywanych czasów *interarrival*, korzystając z równ. (3.2), gdzie Ψ_i to wektor średnich warunkowych, a ε_i to wektor otrzymany w 2. kroku[8];

5. Obliczenie kwantyla wybranego rzędu dla otrzymanej próbki. Znaleziona w ten sposób wartość to wyliczone wprost z definicji $\mathrm{VaR}_\alpha(X)$, przy danym poziomie tolerancji;

6. Powtórzenie wszystkich powyższych kroków tyle razy, ile kroków predykcji chcemy otrzymać.

[8] O tych wektorach wspomnę więcej w kolejnym podrozdziale podczas analizy metod prognostycznych oraz podczas przedstawiania wyników.

Dla modelu ARMA-GARCH, odpowiadającemu zmiennej *breach sizes*, algorytm jest analogiczny. W powyższym algorytmie wystarczy tylko podmienić rozkład innowacji na rozkład skośny T-studenta. Dodatkowo zamiast równ. (3.2), korzystamy z równ. (3.10), gdzie μ_t to wektor średnich warunkowych, σ_t^2 to wektor wariancji warunkowych, a ε_t to wektor otrzymany w 3 kroku. Wyniki symulacji ukazane na odpowiednich wykresach przedstawiłem w kolejnych podrozdziałach.

Zaletą symulacji metodą Monte – Carlo jest możliwość stosowania jej przy niepełnych danych. Wadą natomiast jest duża zależność wyników od przyjętego modelu i dokładności estymacji parametrów.

Ocena adekwatności prognozy VaR

Jedną z technik oceniania otrzymanej wartości narażonej na ryzyko jest *backtesting*. Metoda ta jest sposobem na odróżnienie modelu dokładnego od niedokładnego. W istocie polega na porównywaniu codziennych (lub dla okresów o innej długości) zysków lub strat (w moim przypadku po prostu naruszeń) z miarami VaR, generowanymi np. przez symulacje Monte – Carlo, aby odróżnić modele dokładne od niedokładnych. Zatem *backtesting* polega na ocenie, jak model sprawowałby się, gdyby był wykorzystywany w przeszłości. Jakość prognoz *Value at Risk* ewaluujemy ex-post, czyli po fakcie.
Prowadzi to do następującej definicji (patrz [44]):

Definicja 3.8. Liczbę przekroczeń (lub naruszeń, z ang. *violations*) *backtestingu* nazywa się liczbę dni (lub wybranych okresów o innej długości), w których wartość narażona na ryzyko $\mathrm{VaR}_\alpha(L)$ jest przekroczona przez poziom zysków (lub odpowiednio strat) na analizowanym portfelu.
W szczególności, obserwowana wartość większa od przewidywanej $\mathrm{VaR}_\alpha(L)$ nazywana jest naruszeniem (z ang. *violation*), wskazującym na niedokładność przewidywań.

Analiza liczby naruszeń *backtestingu* pozwala stwierdzić, czy badany model wartości narażonej na ryzyko powinien być zaakceptowany lub odrzucony.
Ogólnie, jeśli stosunek liczby przekroczeń do wielkości próbki jest większy niż $(1-\alpha)$, gdzie α to poziom tolerancji, to mówimy, że model niedoszacowuje ryzyka, a gdy mniejszy, to mówimy, że model przeszacowuje ryzyko. W rzeczywistości stosunek ten bardzo rzadko równy jest $1-$ ustalonemu poziomowi tolerancji.

Aby ocenić dokładność predykcji zarówno pod kątem niedoszacowania jak i przeszacowania wartości VaR, wykorzystałem bezwarunkowy oraz warunkowy test pokrycia (patrz [45] oraz [46]),

które zdefiniowałem poniżej:

1. Test Kupca (bezwarunkowy test pokrycia):

   Test Kupca (1995) (patrz [45, 47]), czyli bezwarunkowy test pokrycia, pozwala sprawdzić, czy spodziewany i obserwowany udział przekroczeń wartości narażonej na ryzyko (VaR) są sobie równe, biorąc pod uwagę wybrany kwantyl i poziom ufności α. Dokładniej, hipotezy mają następującą postać: H_0: spodziewany i obserwowany udział przekroczeń są sobie równe, kontra H_1: są od siebie różne.

   Dla tego testu istnieje także alternatywna hipoteza H_0: stosunek liczby przekroczeń do wielkości próbki, czyli spodziewany udział przekroczeń, jest równy $(1-\alpha)$, gdzie α to poziom ufności.

   Prawdopodobieństwo poniżej wybranego poziomu ufności prowadzi do odrzucenia hipotezy zerowej.

   Statystyka testowa ma postać testu ilorazu wiarygodności i dana jest następującym wzorem:

   $$LR_{uc} = 2[\ln(\hat{p}^X(1-\hat{p})^{N-X}) - \ln(p^X(1-p)^{N-X})] \sim \chi^2(1), \tag{3.13}$$

   gdzie p to spodziewany udział przekroczeń; $\hat{p} = \frac{X}{N}$ to obserwowany udział przekroczeń; X to liczba zaobserwowanych przekroczeń, a N to wielkość próbki. Przy hipotezie zerowej statystyka testowa pochodzi z asymptotycznego rozkładu χ^2 z 1 stopniem swobody.

   Test bezwarunkowego pokrycia wskazuje zarówno modele, które niedoszacowują, jak i te, które przeszacowują VaR. Na podstawie wyniku tego testu nie można jednak powiedzieć, czy model ma skłonność do przeszacowywania czy niedoszacowywania prognozowanej *Value at Risk*.

2. Test Christoffersena (warunkowy test pokrycia):

   Test Christoffersena (2001) (patrz [47, 48]), czyli warunkowy test pokrycia to rozszerzenie bezwarunkowego testu Kupca. Jest on połączeniem testu Kupca oraz tzw. testu niezależności przekroczeń (który także wprowadzony został przez Christoffersena (1998)).

   Hipoteza zerowa testu niezależności przekroczeń mówi, że przekroczenie wartości VaR w okresie t nie zależy od przekroczenia wartości VaR w okresie $t-1$.

   Zatem hipotezą zerową warunkowego testu pokrycia jest to, że przekroczenie wartości VaR w okresie t nie zależy od przekroczenia wartości VaR w okresie $t-1$ i jest równy $(1-\alpha)$, gdzie α to poziom ufności. Statystyka testu Christoffersena dana jest następującym wzorem (zob. [47]):

   $$LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2(2),$$

gdzie:

$$LR_{ind} = 2\ln\left(\frac{(1-\pi_{01})^{N_{00}}(\pi_{01})^{N_{01}}(1-\pi_{11})^{N_{10}}(\pi_{11})^{N_{11}}}{(1-\hat{p})^{N_{00}+N_{10}}(\hat{p})^{N_{01}+N_{11}}}\right),$$

natomiast N_{ij} to liczba obserwacji, dla których zaobserwowano stan j (przekroczenie), pod warunkiem zaobserwowania stanu i w poprzednim okresie (przekroczenie). Dalej, $\pi_{01} = N_{01}/(N_{01} + N_{00})$ to prawdopodobieństwo wystąpienia przekroczenia pod warunkiem braku przekroczenia w poprzednim okresie, π_{11} to prawdopodobieństwo wystąpienia przekroczenia pod warunkiem wystąpienia przekroczenia w poprzednim okresie, $\pi_{11} = N_{11}/(N_{11} + N_{10})$, a $\hat{p} = X/N$ to udział przekroczeń.

Zatem jest to test, który uwzględnia zarówno liczbę przekroczeń jak i ich niezależność w czasie. Ponadto, przy hipotezie zerowej, statystyka testowa pochodzi z asymptotycznego rozkładu χ^2 z 2 stopniami swobody.

Obydwa powyższe testy pozwalają ocenić modele prognozowania wartości VaR ze względu na odpowiedniość jej prognoz. Zostały one zaimplementowane w języku R, w bibliotece GAS pod nazwą BacktestVaR().

Po zaprezentowaniu definicji oraz testów, jakie wykorzystywałem w swojej analizie, przejdę do kolejnej części, w której ukażę metody predykcji dla modelu ACD, a następnie dla modelu ARMA-GARCH. Dalej przedstawię wyniki dot. wartości narażonej na ryzyko (VaR). Pokażę też jej predykcję w oparciu o dwa wspomniane procesy. Zacznę od modelu ACD.

Predykcja w modelu ACD

Jak już wspomniałem na początku tego rozdziału, dla zmiennej *interarrival times* zdecydowałem się wybrać model ACD$(1,1)$, z innowacjami o rozkładzie uogólnionym gamma (patrz równ. (3.5)).

Zacznę więc od predykcji dot. zmiennej *interarrival times* przy użyciu odpowiedniego modelu ACD. W rozdziale 3.3.2 wprowadziłem odpowiednie oznaczenia oraz definicję tego procesu (patrz równania (3.2) – (3.4)). Warto jednak ponownie przypomnieć najważniejsze z nich.

Przez $d_i = t_i - t_{i-1}$ oznaczyłem czasy *interarrival*, gdzie t_i to czas, w którym wystąpił incydent. Ponadto standardowy model ACD$(1,1)$ ma następującą postać:

$$\begin{cases} d_i = \Psi_i \cdot \varepsilon_i \\ \Psi_i = \mathbb{E}(d_i|\mathcal{F}_{i-1}) = \omega + a_1 d_{i-1} + b_1 \Psi_{i-1} \end{cases} \tag{3.14}$$

gdzie $\omega > 0$, $a_1, b_1 \geqslant 0$ to parametry modelu; $\mathcal{F}_{i-1}$ reprezentuje wiedzę zgromadzoną do momentu t_{i-1}; ε_i to innowacje – niezależne zmienne losowe o tym samym rozkładzie (*i.i.d.*), $\mathbb{E}(\varepsilon_i) = 1$, $\forall_{i=1,\ldots,n}$. Zakładamy też, że $\mathcal{F}_{i-1}$ jest niezależne od ε_i.

Przedstawię teraz trzy wykorzystane przeze mnie metody predykcji przyszłych trajektorii zmiennej *interarrival times*. Rozpocznę od metody predykcji, która opisana została przez L. Cheung w publikacji [62]. Następnie przedstawię zmodyfikowaną metodę pierwszą, która zaproponowana została w dodatku do [62], a także w [78], której wyniki okazały się lepsze. Trzecia wykorzystana przeze mnie metoda to modyfikacja metody ukazanej w Xu i inni w publikacji [24]. Następnie, w postaci obserwacji, przedstawiłem własną interpretację.

Pierwsze kroki każdej z tych metod są takie same, więc przedstawiłem je poniżej.

W celach prognostycznych dane (zmienna *interarrival times*) $\{d_1, d_2, ..., d_n\}$ podzieliłem na próbę uczącą, oraz na próbę testową. Próbę uczącą oznaczyłem przez $\{d_1, d_2, ..., d_h\}$, dla $h < n$, a próbę testową, czyli resztę posiadanych danych, oznaczyłem przez $\{d_{h+1}, ..., d_n\}$, gdzie $n = h + L$. Dostępny mi zbiór danych składa się z 1243 obserwacji, więc w moim wypadku $n = h + L = 1243$. Takie h w literaturze (patrz. [24, 62]) nazywane jest początkiem prognozy (z ang. *forecast origin*).

W kolejnych paragrafach opisałem następne kroki wybranych metod.

Pierwsza metoda (na podst. [62]):

Do danych aż do h, czyli na próbie uczącej, dopasowałem proces ACD$(1, 1)$, z innowacjami o rozkładzie uogólnionym gamma, przy użyciu funkcji acdFit() z biblioteki ACDm. Jak już wspomniałem wcześniej, funkcja ta estymuje różne modele ACD z różnymi założonymi rozkładami innowacji przy użyciu metody *MLE*. W metodzie tej dopasowanie modelu do danych odbywa się tylko raz – na samym początku. Przy użyciu zaimplementowanej metody coef() na dopasowanym przez funkcję acdFit() obiekcie, znalazłem estymowane wartości parametrów ω, a_1 i b_1, których wartości widoczne są w tab. 3.18. Dodatkowo, korzystając z metody muHats() z modelu wyciągnąłem wektor dopasowanej średniej warunkowej Ψ_i. Ponadto wiemy (patrz [62]), że otrzymane z modelu ε_i to standaryzowane residua dopasowanego modelu. Wtedy dopasowany do danych model ACD, z wyestymowanymi parametrami, można wykorzystać do przewidywania przyszłych czasów *interarrival* z l-krokowym wyprzedzeniem, dla $l = 1, ..., L$. W ten sposób uzyska się $h + l = \{h + 1, ..., h + L\}$ prognoz.

Niech $d_h(l)$ oznacza l-krokową prognozę, z początkiem prognozy h oraz $l = 1, ..., L$. Wówczas prognoza 1-krokowa ma następującą postać (patrz [62]):

$$d_h(1) = \mathbb{E}(d_{h+1}|\mathcal{F}_h) = \Psi_{h+1}. \tag{3.15}$$

Ponadto, z definicji wiemy, że $\Psi_i = \mathbb{E}(d_i|\mathcal{F}_{i-1}) = \mathbb{E}(\Psi_i\varepsilon_i|\mathcal{F}_{i-1})$ oraz $\mathbb{E}(\varepsilon_i) = 1$. Dodatkowo d_i, dla $i = 1, ..., h$ jest znaną stałą, zatem $\Psi_i = \mathbb{E}(d_i|\mathcal{F}_{i-1}) = d_i$, $i = 1, ..., h$. Z tego wynika, że Ψ_{h+1}

jest znane dla $(h+1)$-ej obserwacji:

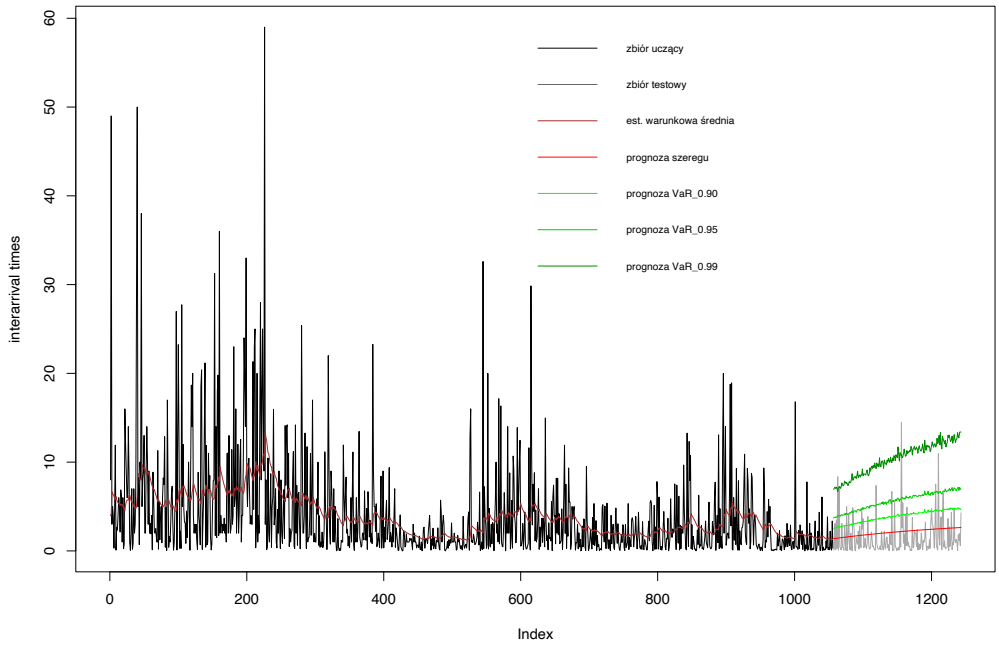
$$\Psi_{h+1} = d_h(1) = \omega + a_1 d_h + b_1 \Psi_h = \omega + (a_1 + b_1) d_h. \tag{3.16}$$

Wykorzystując powyższe równ. (3.16), kolejne predykcje z wielo-krokowym wyprzedzeniem można wyprowadzić podstawiając do modelu $d_h(1)$ i Ψ_{h+1} następująco:

$$\begin{aligned}
\Psi_{h+2} &= d_h(2) = \omega + a_1 d_h(1) + b_1 \Psi_{h+1} = \omega + (a_1 + b_1)\Psi_{h+1}, \\
\Psi_{h+3} &= d_h(3) = \omega + a_1 d_h(2) + b_1 \Psi_{h+2} = \omega + (a_1 + b_1)\Psi_{h+2}, \\
&\vdots \\
\Psi_{h+L} &= d_h(L) = \omega + a_1 d_h(L-1) + b_1 \Psi_{h+L-1} = \omega + (a_1 + b_1)\Psi_{h+L-1},
\end{aligned} \tag{3.17}$$

Zatem średnią warunkową modelu autorzy utożsamiają z d_i, czyli czasami *interarrival*. Wtedy otrzymane $\{\Psi_{h+1}, ..., \Psi_{h+L}\} = \{d_h(1), ..., d_h(L)\}$ to L kolejnych wartości prognozy.

Następnie korzystając z kroków na str. 98 algorytmu symulacji metodą Monte – Carlo, obliczyłem predykcję wartości narażonej na ryzyko. Do jej obliczenia skorzystałem ze wzoru $d_i = \Psi_i \cdot \varepsilon_i$ (patrz. [62]), gdzie $\{\Psi_{h+1}, ..., \Psi_{h+L}\}$ otrzymałem już z powyższej analizy, a $\{\varepsilon_{h+1}, ..., \varepsilon_{h+L}\}$ to wygenerowane próbki z rozkładu uogólnionego gamma zgodnie ze wspomnianymi krokami. Wtedy licząc kwantyl wybranego rzędu (patrz rys. 3.20) z otrzymanych przeskalowanych próbek d_i, dla $i = h+1, ..., h+L$, otrzymałem prognozowane wartości VaR_α.
Wyniki predykcji, do której wykorzystałem opisaną powyżej metodę, ukazałem na rysunku 3.20



Rysunek 3.20: Predykcja zmiennej *interarrival times* z wykorzystaniem modelu ACD$(1,1)$ (pierwsza metoda, patrz [62]).

Na rysunku 3.20 widzimy wykres szeregu czasowego odp. danym z próbki uczącej (1056 obserwacji) oraz próbki testowej (187 obserwacji), dla zmiennej *interarrival times*. Zaznaczone są one odpowiednio krzywą czarną oraz krzywą szarą. Z kolei brązowa krzywa ukazuje oszacowaną średnią warunkową odp. modelowi ACD$(1,1)$. Otrzymałem ją korzystając z metody muHats(), o której wspomniałem powyżej. Dodatkowo czerwona krzywa przedstawia prognozę szeregu. Ponadto, trzema odcieniami zielonego koloru wyrysowane są predykcje wartości VaR_α, dla poziomów tolerancji kolejno $\alpha = 0.9$, 0.95 i 0.99 (zob. legenda). Widzimy, że czerwona krzywa predykcyjna kształtem przypomina prostą, lecz nią nie jest. Jej nieskomplikowany kształt jest uzasadniony, gdyż we wzorach (3.16) i (3.17) nie ma żadnej losowości. Z kolei słabo rosnące wartości prognozy szeregu nie wydają się zbyt prawdopodobne, gdy popatrzymy na ogólny kształt trajektorii całego szeregu czasowego. Aby jednak formalnie, numerycznie sprawdzić dokładność otrzymanej prognozy, wykorzystałem pewne miary.

Jednym ze sposobów porównania przewidywanych przyszłych *interarrival times* $\{d_h(1), ..., d_h(L)\}$ z obserwowanymi przyszłymi *interarrival times* $\{d_{h+1}, ..., d_{h+L}\}$ jest użycie miary MAPE (ang. *Mean Absolute Percentage Error*, patrz [62]). Definiuje się ją następująco:

$$\mathsf{MAPE} = \frac{1}{L}\sum_{l=1}^{L}\Big|\frac{d_{h+l} - d_h(l)}{d_{h+l}}\Big|. \tag{3.18}$$

Wskaźnik MAPE stosuje się jako miarę dokładności prognozy dopasowanego modelu. Jest on łatwy w interpretacji: np. MAPE $= 0.14$ oznacza, że średnia różnica między wartością prognozowaną, a rzeczywistą wynosi 14%. W ogólności, model o najniższym MAPE wykazuje najmniejszą liczbę błędów w prognozach.

Druga miara, analogiczna do powyższej, to MedianAPE (ang. *Median Absolute Percentage Error Loss*). Jej definicja jest bardzo podobna do definicji MAPE, ale w tym wypadku liczy się medianę zamiast średniej. Jej zaletą jest odporność na obserwacje odstające. Przykładowo MedianAPE $=$ 0.14 oznacza, że w połowie przypadków błędy prognozy przekraczały 14%, a w drugiej połowie były mniejsze niż 14%.

W moim przypadku wskaźnik MAPE wyniósł 0.80767, co wskazuje na to, że średnia różnica wynosi ok. 80.76%. Wartość ta oczywiście jest dość duża. Ponadto wskaźnik MedianAPE wyniósł $0.720848 \approx 72\%$. Można więc przypuszczać, że otrzymana prognoza nie jest najlepsza.

Wróćmy ponownie do prognozy wartości VaR i testowania jej przekroczeń. Przeprowadziłem porównanie spodziewanej liczby przekroczeń z zaobserwowaną liczbą, a także wykonałem testy Kupca i Christoffersena. Wyniki tej analizy widoczne są w tablicy 3.27.

α	Zaobs.	Spodziew.	p-val LR_{uc}	p-val LR_{cc}	Decyzja$_{uc}$	Decyzja$_{cc}$
0.90	23	19	0.3098153	0.2252992	Fail to Reject H0	Fail to Reject H0
0.95	9	10	0.90595586	0.6282587	Fail to Reject H0	Fail to Reject H0
0.99	3	2	0.4451566	0.7113009	Fail to Reject H0	Fail to Reject H0

Tablica 3.27: Wyniki testów Kupca i Christoffersena dla liczby przekroczeń przy ustalonym poziomie istotności równym 0.05. Model ACD(1,1) z innowacjami o rozkładzie uogólnionym gamma (pierwsza metoda).

Próbka testowa liczy 187 obserwacji zatem oczekiwana liczba przekroczeń VaR przy poziomie tolerancji $\alpha = 0.9$ to $(1-0.9)\cdot 187 \approx 19$, przy poziomie $\alpha = 0.95$ to $(1-0.95)\cdot 187 \approx 10$, a przy $\alpha = 0.99$ to $(1-0.99)\cdot 187 \approx 2$. W tab. 3.27 wartości te widoczne są w kolumnie *spodziew.* Dodatkowo, patrząc też na rys. 3.20, przy poziomie tolerancji $\alpha = 0.9$ można zaobserwować 23 naruszenia, na poziomie $\alpha = 0.95$: 9 naruszeń, a przy poziomie $\alpha = 0.99$ już tylko 3 naruszenia. Widzimy więc, że dla poziomów $\alpha = 0.9$ i $\alpha = 0.99$ model niedoszacowuje wartości VaR, natomiast dla $\alpha = 0.95$ model przeszacowuje VaR. Dodatkowo, zarówno test Kupca, jak i Christoffersena, których p – wartości widoczne są w kolumnach odpowiednio *p-val* LR_{uc} i *p-val* LR_{cc} wskazują na brak podstaw na odrzucenie hipotez zerowych dla każdego poziomu tolerancji VaR. W przypadku testu Kupca oznacza to, że spodziewany i obserwowany udział przekroczeń są sobie w przybliżeniu równe. Natomiast dla testu Christoffersena oznacza to, że przekroczenie VaR w okresie t nie zależy od przekroczenia VaR w $t-1$ i jest w przybliżeniu równe $1-\alpha$ (patrz hipoteza alternatywna dla testu Kupca na str. 100). Można więc zaobserwować, że wybrany model predykcyjny przechodzi wszystkie testy.

Reasumując, wskaźniki takie jak MAPE i MedianAPE pokazują, że predykcja szeregu czasowego nie wydawała się zbyt dobra. Jednakże patrząc dalej na wyniki testów Kupca i Christoffersena, a także analizując zaobserwowane naruszenia, można stwierdzić, że wybrany model jest stosunkowo dobry.

Przejdę teraz do kolejnej metody predykcji.

Druga metoda (modyfikacja metody pierwszej, [78]):

Kolejną analizowaną przeze mnie metodą, jest zmodyfikowana pierwsza metoda. Tym razem wzorowałem się na publikacji [78] oraz dodatku do [62]. Analogicznie jak powyżej, zacząłem od jednorazowego dopasowania modelu do danych z próbki uczącej. Następnie z otrzymanego obiektu wyciągnąłem wyestymowane parametry i dopasowaną średnią warunkową. Pierwszy krok prognozy ma taką samą postać jak w metodzie pierwszej (patrz równ. (3.16)). Kolejne kroki

zostały trochę zmodyfikowane. Jak już wspominałem, w równaniach (3.17) $\{\Psi_{h+1}, ..., \Psi_{h+L}\}$ to kolejne średnie warunkowe. Różnicą natomiast jest założenie, że $\{d_h(1), ..., d_h(L)\}$ są po prostu równe kolejnym wartościom $\{d_{h+1}, ..., d_{n-1}\}$ z próbki testowej.

Taka metoda predykcji może budzić wątpliwości co do swojej poprawności. Z reguły zakłada się, że dane z próbki testowej nie są jeszcze dostępne. Jednakże w świecie rzeczywistym (np. banki lub inne instytucje finansowe) model dopasowuje się przeważnie raz, na początku, a ponowne dopasowywania nie są wykonywane codziennie, tylko co jakiś określony dłuższy okres czasu. W tym wypadku przeprowadzam jedynie predykcję 1-krokową, tzn. np. dzisiaj wykonuję prognozę tylko na dzień jutrzejszy. Następnego dnia kolejna wartość jest już znana, więc tę nowo zaobserwowaną wartość mogę dodać do analizowanej próbki uczącej. Wtedy ponownie można przeprowadzić prognozę na kolejny dzień, korzystając z poszerzonego zbioru danych, a kolejnego dnia zaobserwowaną wartość znowu dodać do tego zbioru. Czynność ta może być powtarzana wybraną liczbę razy. Taka metoda prognozowania, w której zbiór danych uczących jest poszerzany po każdym kroku predykcji nazywa się prognozą kroczącą. Jej definicję zaprezentowałem poniżej.

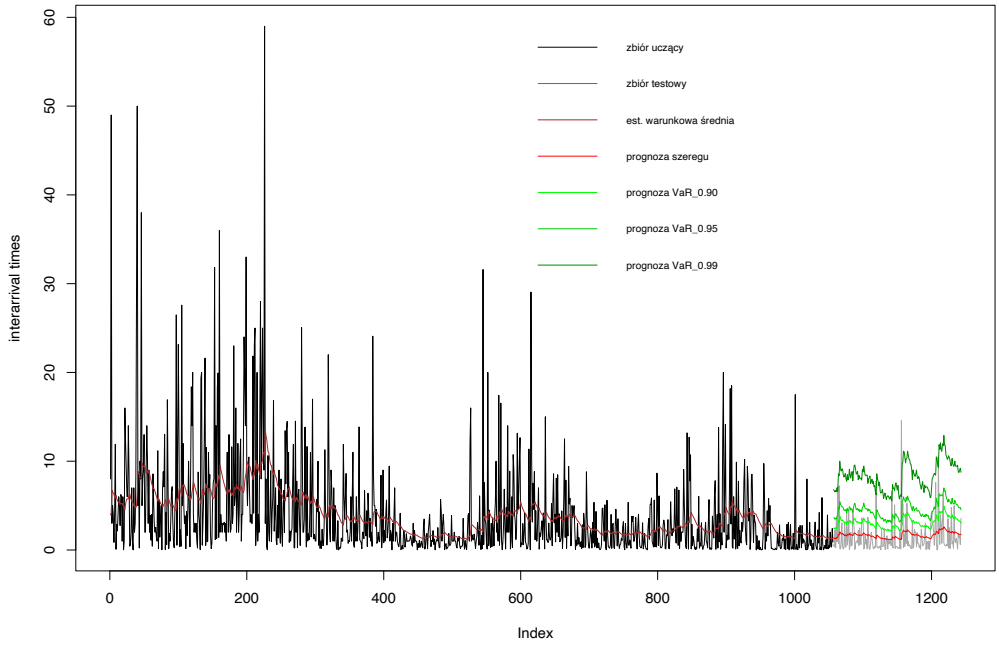
Definicja 3.9 (Prognoza krocząca). Prognoza krocząca (z ang. *rolling forecast*) to metoda predykcyjna, która wykorzystuje dane historyczne do przewidywania przyszłych wartości szeregu w sposób ciągły, przez pewien okres czasu[9]. Dokładniej, liczba danych z próbki uczącej rośnie wraz z kolejnymi krokami procesu predykcji. Wybrany model może być na nowo dopasowywany do powiększającego się zbioru danych albo po każdej prognozie 1-krokowej, albo co kilka kroków.

Jedyna różnica jest taka, że w tym wypadku modelu nie dopasowywałem po każdym kroku, ani nawet co kilka kroków, tylko na samym początku.

Taka metoda, dla której dostajemy tylko prognozę na przyszły dzień, może nie dawać zbyt dużo informacji, ale np. w przypadku gdy chcielibyśmy przewidzieć, czy następnego dnia nastąpi wzrost czy spadek kursu akcji, to może okazać się ona użyteczna.

Wracając ponownie do metody predykcji – wszystkie kolejne kroki (wraz ze sposobem liczenia VaR) są takie same jak w oryginalnej metodzie z publikacji [62], które opisałem powyżej. Wyniki tej analizy przedstawiłem na wykresie, na rysunku 3.21.

[9] W języku ekonomiczno - biznesowym prognoza krocząca to sprawozdanie, które wykorzystuje dane historyczne do przewidywania przyszłych liczb i pozwala organizacjom przewidywać przyszłe wyniki dot. budżetu, wydatków i innych danych finansowych, w oparciu o ich wyniki z przeszłości.



Rysunek 3.21: Predykcja zmiennej *interarrival times* z wykorzystaniem modelu ACD(1, 1) (druga metoda, patrz [62] i dodatek do [62]).

Na rysunku 3.21 widzimy wykres szeregu czasowego wraz z otrzymanymi predykcjami. Wszystkie oznaczenia są tutaj takie same jak wspomniane na rys. 3.20. Kształt trajektorii znalezionej prognozy nie przypomina już linii prostej. Małe skoki widoczne przy predykcji szeregu jak i predykcji VaR-ów wydają się bardziej sensowne, niż analogiczne w metodzie pierwszej. Przejdę od razu do oceny dokładności otrzymanej prognozy. W metodzie pierwszej wskaźnik MAPE wyniósł 0.80767, a wskaźnik MedianAPE wyniósł 0.720848. W tej metodzie natomiast wyniosły one odpowiednio 0.707619 oraz 0.5671815. Wskazuje to na lepsze odwzorowanie przyszłych trajektorii przez tę prognozę, a także na mniejszą liczbę błędów w prognozach.

Dalej, liczby naruszeń VaR w pierwszej metodzie wynosiły odpowiednio 23, 9 oraz 3, dla VaR_α na poziomie tolerancji $\alpha = 0.9$, 0.95 i 0.99 (patrz tab. 3.27). Natomiast analogiczne wyniki dla drugiej metody przedstawiłem w następującej tablicy 3.28.

α	Zaobs.	Spodziew.	p-val LR_{uc}	p-val LR_{cc}	$Decyzja_{uc}$	$Decyzja_{cc}$
0.90	30	19	0.01056706	0.03543554	Reject H0	Reject H0
0.95	13	10	0.1436913	0.3415856	Fail to Reject H0	Fail to Reject H0
0.99	5	2	0.0000919465	0.0004156057	Reject H0	Reject H0

Tablica 3.28: Wyniki testów Kupca i Christoffersena dla liczby przekroczeń przy ustalonym poziomie istotności równym 0.05. Model ACD(1,1) z innowacjami o rozkładzie uogólnionym gamma (druga metoda).

W tablicy 3.28 w drugiej kolumnie liczby przekroczeń były odpowiednio równe 30, 13 i 5 przy analogicznych poziomach tolerancji. Widzimy tu niestety pogorszenie wyników. Testy Kupca i Christoffersena przy poziomie istotności równym 0.05 odrzucają hipotezę zerową dla poziomów tolerancji $\alpha = 0.9$ oraz $\alpha = 0.99$. Hipoteza zerowa mogłaby być przyjęta jedynie dla poziomu tolerancji $\alpha = 0.95$. Dodatkowo widzimy, że metoda ta niedoszacowuje ryzyka dla każdego badanego poziomu ufności.

Reasumując, z jednej strony widać, że wyniki predykcji przyszłych *interarrival times* w tej metodzie są lepsze, niż w pierwszej metodzie, na co wskazują miary MAPE oraz MedianApe. Z drugiej jednak strony testy przekroczeń sygnalizują, że predykcja wartości narażonej na ryzyko okazała się lepsza w metodzie pierwszej niż w drugiej. Dodatkowo, bardziej skokowy kształt lepiej odzwierciedla realne zachowanie czasów *interarrival*. Jej wadą jest możliwość przewidywania tylko jednego kroku wprzód, podczas gdy metoda pierwsza pozwala przewidywać wybraną lczbę kroków wprzód w dowolnej chwili. Należy jednak pamiętać, że do obliczenia predykcji VaR wykorzystywałem metodę symulacji Monte – Carlo, zatem po każdym jej wykonaniu wyniki będą się od siebie trochę różnić. Może to więc wpływać na wyniki obydwu zastosowanych przeze mnie testów.

Trzecia metoda (modyfikacja metody [24]):

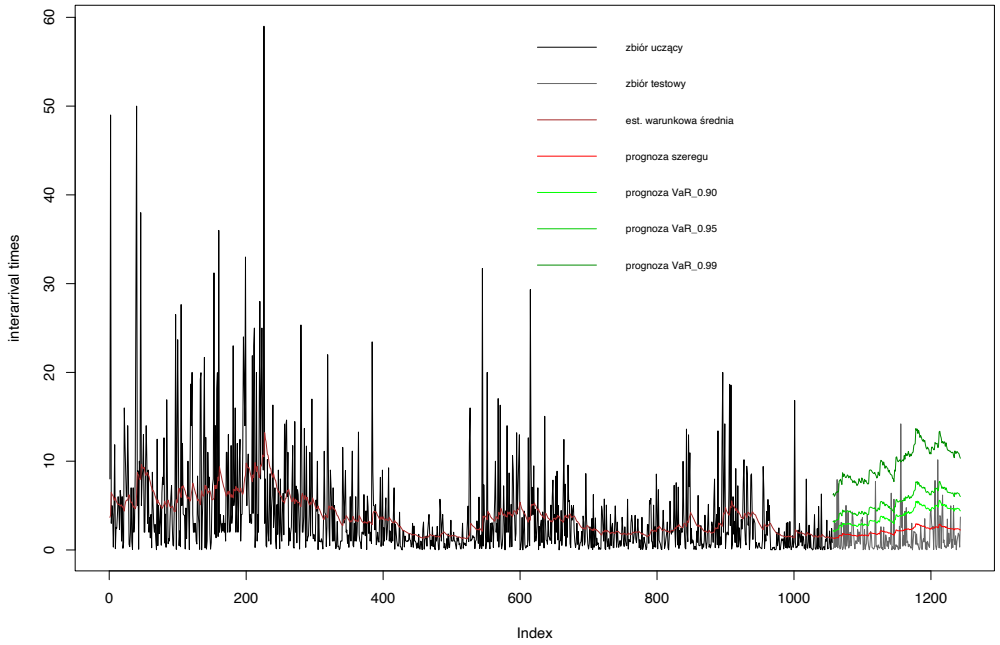
Ostatnia – trzecia metoda, którą zdecydowałem się opisać w swojej pracy, przedstawiona została przez Xu i inni w publikacji [24]. Jest ona dużo bardziej skomplikowana obliczeniowo, oraz bardziej losowa. Aby lepiej odzwierciedlić rzeczywisty kształt krzywej predykcyjnej otrzymanej w pierwszej lub drugiej metodzie, w tym wypadku, tak jak powyżej, wykorzystałem metodę prognozy kroczącej. Metoda ta różni się jednak liczbą dopasowań modelu. Zaproponowana przeze mnie modyfikacja tej metody dotyczy chwili dopasowywania modelu do danych. W publikacji autorzy proponują dopasowanie modelu tylko na początku oraz poszerzanie zbioru danych o kolejne obliczone wartości, ja natomiast zdecydowałem się dopasowywać model do danych po każdej

prognozie 1-krokowej. Kiedy przetestowałem obydwie metody na dostępnych mi danych, okazało się, że zaproponowana przeze mnie modyfikacja pozwoliła uzyskać lepsze wyniki. Dlatego też postanowiłem z niej skorzystać.

Przejdę teraz do opisu metody predykcji. Podobnie jak przy pierwszej metodzie, zacząłem od dopasowania do danych z próbki uczącej odpowiedniego modelu ACD$(1,1)$ przy pomocy funkcji acdFit(). Dalej, analogicznie wyciągnąłem dopasowaną średnią warunkową (przy pomocy metody muHats()) oraz wszystkie wyestymowane parametry modelu. Następnie, wprost ze wzoru $\Psi_{h+1} = \omega + a_1 d_h + b_1 \Psi_h$, obliczyłem 1 krok prognozowanej średniej warunkowej, gdzie Ψ_h to ostatnia wartość wektora dopasowanej średniej warunkowej, a d_h to ostatnia wartość zbioru uczącego. Dalsze kroki prognozy są bardziej skomplikowane, ponieważ aby otrzymać następne wartości predykcji oraz predykcję VaR, trzeba znać przyszłe d_i, dla $i = h+1, ..., h+L-1$.

W tej metodzie, podobnie jak w drugiej, zakładamy, że $\{d_h(1), ..., d_h(L)\}$ to przyszłe wartości *interarrival times*. Aby znaleźć kolejne $d_h(i)$ przypomnę pierwsze równanie definicji modelu ACD, czyli: $d_i = \Psi_i \cdot \varepsilon_i$. By uzyskać kolejną predykcję *interarrival times*, trzeba znać wektory Ψ_i i ε_i. Korzystając z pierwszego kroku powyżej dostałem pierwszy z tych wektorów. Natomiast ε_i otrzymałem korzystając z metody przedstawionej w algorytmie na str. 98. W 3 punkcie algorytmu jest mowa o wygenerowaniu 10000 losowych wartości z rozkładu uogólnionego gamma z ustalonymi parametrami. W tym wypadku, ponieważ jest to metoda 1-krokowa, losowałem po prostu 1 wartość zamiast 10000 (patrz [24]). Zatem skalując prognozę Ψ_i przez wylosowaną próbkę, otrzymałem 1-krokową predykcję d_{h+1}. Następnie, wykonując wspomniany algorytm i licząc kwantyl wybranego rzędu z otrzymanej przeskalowanej próbki d_{h+1}, otrzymałem prognozowaną wartość VaR_α. Dalej, wektor wyestymowanych średnich warunkowych oraz wektor czasów *interarrival* (z próbki uczącej) poszerzałem o wartość otrzymanej predykcji (prognoza krocząca). Wszystkie powyższe kroki powtarzałem w pętli, aby otrzymać 187 kroków predykcji.

Wyniki predykcji, do której wykorzystałem tę metodę, pokazałem na rysunku 3.22.



Rysunek 3.22: Predykcja zmiennej *interarrival times* z wykorzystaniem modelu ACD(1, 1) (trzecia metoda, patrz [24]).

Podobnie jak na rysunku 3.20 i 3.21 widzimy tu wykres szeregu czasowego odp. danym z próbki uczącej (1056 obserwacji) oraz próbki testowej (187 obserwacji), dla zmiennej *interarrival times*. Wszystkie oznaczenia są tu takie same jak na wspomnianych powyżej rysunkach. Ponownie widzimy, że kształt przyszłych trajektorii różni się od tych przedstawionych na rysunku 3.20. Nie przypominają już one prostej. Dodatkowo, widać pewną losowość, która wprowadzona została dzięki skalowaniu średniej warunkowej. Trajektorie najpierw łagodnie rosną, a następnie dość łagodnie maleją. Porównując prognozę szeregu z zaobserwowanymi przyszłymi czasami *interarrival* widzimy, że prognoza nadal nie była w stanie odzwierciedlić kilku dużych skoków, tak jak na rys. 3.20 oraz 3.21.

Aby sprawdzić poprawność otrzymanej prognozy, po raz kolejny wykorzystałem wskaźniki MAPE i MedianAPE. W tym wypadku obydwa okazały się dużo wyższe, niż wartości otrzymane w metodzie pierwszej lub drugiej. Dokładniej, MAPE był równy 6.48362, a MedianAPE równy 1.979727. Oznacza to, że średnia różnica między wartością prognozowaną, a rzeczywistą wynosi ok. 648%, z kolei odpowiadająca mediana wynosi 197%. Wartości te oczywiście są ogromne. Wskazuje to na dużo większą liczbę błędów w prognozach. Biorąc jednak pod uwagę wspomniany czynnik losowy, który sprawia, że trajektorie procesu są bardziej skokowe, to wartość tego współczynnika rzeczywiście może być dużo większa.

Z kolei wyniki analizy naruszeń okazały się bardziej optymistyczne. Dla przypomnienia, w tablicach 3.27 i 3.28 podałem wyniki odpowiednio dla metody pierwszej i drugiej. Natomiast w tablicy 3.29 przedstawiłem wyniki tej analizy dot. metody trzeciej.

α	Zaobs.	Spodziew.	p-val LR_{uc}	p-val LR_{cc}	Decyzja$_{uc}$	Decyzja$_{cc}$
0.90	26	19	0.09085755	0.2214065	Fail to Reject H0	Fail to Reject H0
0.95	11	10	0.5897152	0.4328867	Fail to Reject H0	Fail to Reject H0
0.99	3	2	0.4451566	0.7113009	Fail to Reject H0	Fail to Reject H0

Tablica 3.29: Wyniki testów Kupca i Christoffersena dla liczby przekroczeń przy ustalonym poziomie istotności równym 0.05. Model ACD(1,1) z innowacjami o rozkładzie uogólnionym gamma (trzecia metoda).

Zaobserwowane przekroczenia wartości narażonej na ryzyko wynosiły odpowiednio: 26 przy poziomie tolerancji $\alpha = 0.9$; 11 przy poziomie tolerancji $\alpha = 0.95$ oraz 3 przy poziomie tolerancji $\alpha = 0.99$. Zatem, ponownie jak przy metodzie drugiej, model niedoszacowuje ryzyka. Patrząc na znalezione p – wartości testów Kupca i Christoffersena widzimy, że ten model predykcyjny przechodzi wszystkie testy na poziomie istotności 0.05. Gdyby jednak poziom istotności testu ustalić na równy 0.1, to należałoby odrzucić hipotezę zerową testu Kupca, dla poziomu istotności VaR równego $\alpha = 0.9$.

Obserwacja 3.10. Można się zastanawiać nad słusznością liczenia wartości narażonej na ryzyko dla zmiennej *interarrival times*. Wiemy, że czym większy jest rozmiar naruszeń, tym większe są straty dla danej firmy. W tej sytuacji interpretacja wartości narażonej na ryzyko jest oczywista. Jednakże czym większy jest czas pomiędzy zdarzeniami, tym lepiej dla tej firmy. Wtedy liczenie przekroczeń VaR może nie mieć większego znaczenia. W tym wypadku wartość VaR można zinterpretować inaczej. Mianowicie jesteśmy w stanie zobaczyć, w których momentach w przyszłości można spodziewać się ataków, a dokładniej – kiedy one nastąpią. Patrząc np. na rys. 3.22, widzimy, że w okresie od ok. 1140 do 1200 można spodziewać się zwiększonej liczby ataków.

Na koniec tego podrozdziału przejdę do wniosków z przeprowadzonej analizy na zmiennej *interarrival times*.

Wnioski

Przy rozpatrywaniu tylko i wyłącznie testów Kupca i Christoffersena okazuje się, że pierwsza metoda wypada najlepiej, a najgorzej – metoda druga. Różnice pomiędzy nimi nie są jednak

aż takie duże. Z kolei patrząc na wskaźniki MAPE oraz MedianAPE stwierdzamy, że najlepiej prezentuje się predykcja otrzymana metodą drugą, a najgorzej – metodą trzecią. W tym wypadku różnice pomiędzy dwiema pierwszymi metodami są nieznaczne, natomiast metoda trzecia jest znacznie gorsza. Z kolei subiektywne patrząc na trajektorie tych szeregów czasowych możemy stwierdzić, że trzecia metoda wydaje się być stosunkowo dobra, jednakże losowość otrzymanych wartości predykcji jest tutaj bardzo duża.

Reasumując, predykcja zmiennej *interarrival times* jest trudna. Przez niestandardowy, skokowo – malejący, ale też wykazujący pewną okresowość charakter trajektorii, ciężko jest uzyskać sensowną prognozę szeregu. Wyniki przedstawione przez autorów artykułu [24] również nie były w stanie przewidzieć przyszłych dużych skoków.
Wybór jednej metody predykcji jest zadaniem ciężkim. Jeśli jednak miałbym wybrać tylko jedną, to najłatwiejszym wyborem jest metoda pierwsza. Jest ona prosta w implementacji, a jej wyniki można uznać za wystarczające. Minusem jest jednak kształt przyszłych trajektorii. Prosty charakter tej metody nie odzwierciedla skokowej natury analizowanej zmiennej.

W kolejnym podrozdziale zaprezentuję metodę predykcji zmiennej *breach sizes* przy wykorzystaniu odpowiedniego procesu ARMA-GARCH. Następnie przedstawię wyniki swoich badań.

Predykcja w modelu ARMA-GARCH

Przejdę teraz do części dotyczącej predykcji szeregu, oraz predykcji wartości narażonej na ryzyko dla zmiennej *breach sizes*. W tym wypadku, do prognozowania zdecydowałem się wykorzystać wybrany wcześniej model ARMA$(1,1)$-GARCH$(1,1)$ z innowacjami o rozkładzie skośnym T-studenta (patrz tab. 2.2). Do szacowania parametrów modelu, predykcji przyszłych rozmiarów naruszeń jak i wartości VaR wykorzystałem wspomnianą już obszerną bibliotekę rugarch. W przeciwieństwie do wykorzystywanej przeze mnie w poprzednim podrozdziale biblioteki ACDm, która takich operacji nie umożliwia, biblioteka rugarch okazała się bardzo pomocnym narzędziem prognostycznym. Wykorzystana przeze mnie metoda predykcyjna dot. modelu ARMA-GARCH podobna jest do metody pierwszej i trzeciej (przedstawionych przeze mnie powyżej), które dotyczyły modelu ACD (uwzględniając oczywiście definicje tych modeli). Tym razem, zamiast samemu wyprowadzać wzory, tak jak w poprzednim podrozdziale, skupiłem się na skrótowym przedstawieniu biblioteki rugarch oraz na pokazaniu, w jaki sposób z niej korzystać, aby otrzymać odpowiednie wyniki.

Podobnie jak w przypadku zmiennej *interarrival times* i dopasowanego do niej modelu ACD$(1,1)$, dane (zmienna zlogarytmowana *breach sizes*) $\{y_{t_1}, y_{t_2}, ..., y_{t_n}\}$ podzieliłem na próbkę

uczącą, oraz próbkę testową. Miałem dostępne $n = 1244$ obserwacje. Wielkość próbki uczącej ustaliłem na równą 1119 i oznaczyłem ją przez $\{y_{t_1}, ..., y_{t_h}\}$, dla $h = 1119 < n$, natomiast wielkość próbki testowej na 125 i oznaczyłem ją przez $\{y_{t_h+1}, ..., y_{t_n}\}$. Do danych z próbki uczącej dopasowałem model ARMA$(1,1)$-GARCH$(1,1)$, korzystając z opisanych już przeze mnie w rozdz. 3.3.2 funkcji ugarchspec() oraz ugarchfit() (które estymują różne rodzaje modeli ARMA-GARCH). Następnie, aby dostać estymowane wartości parametrów $\mu, \phi_1, \theta_1, \omega, \alpha_1$ i β_1, których oszacowane wartości przedstawiłem w tab. 3.23, użyłem metodę *coef()* na dopasowanym przez funkcję ugarchfit() obiekcie. Dodatkowo, wykorzystana przeze mnie funkcja ugarchfit() pozwoliła mi na wyciągnięcie z niej dopasowanych do danych wektorów $\mu_t = \mathbb{E}(Y_t|\mathcal{F}_{t-1})$ (stosując metodę fitted() na dopasowanym modelu) oraz $\sigma_t = \sqrt{\text{Var}(\varepsilon_t|\varepsilon_{t-s}, \sigma^2_{t-s}, s > 0)}$ (wykorzystując metodę sigma() na dopasowanym modelu), gdzie, dla przypomnienia, $Y_t = \mu_t + \varepsilon_t = \mu_t + \sigma_t \cdot Z_t$. Wektory μ_t i σ_t odpowiadają oszacowanej średniej warunkowej oraz warunkowemu odchyleniu standardowemu (patrz równ. (3.7), (3.8) i (3.9)).

W celu predykcji tych dwóch wektorów, czyli średniej warunkowej i warunkowego odchylenia standardowego, wykorzystałem kolejną funkcję, tj. ugarchforecast() z tej samej biblioteki. Poniżej przedstawiłem opis użytych przeze mnie parametrów:

- n.ahead - oznacza horyzont czasowy predykcji. Ogólnie, prognozy 1-krokowe bazują na poprzednich wartościach z danych, natomiast prognozy n-krokowe ($n > 1$) bazują na bezwarunkowej wartości oczekiwanej modelu. Ponadto, prognoza opiera się na wartości oczekiwanej innowacji, czyli wybranej gęstości. Ponieważ wybrałem metodę prognozy 1-krokowej (podobnie jak [24, 63]), parametr ten ustaliłem na równy 1;

- n.roll - odpowiada liczbie kroków prognozy kroczącej (patrz definicja 3.9). Parametr ten kontroluje, ile razy ma być wykonana prognoza n.ahead. Domyślnie ustawiony jest na 0. W takim wypadku funkcja zwracałaby standardową prognozę n.ahead (tak jak metoda druga, przy modelu ACD). Zdecydowałem się wykonać 125 kroków prognozy kroczącej (tj. ilość elementów próbki testowej), dlatego też parametr ten ustaliłem na równy 124;

- out.sample - parametr opcjonalny, który wskazuje, ile punktów danych zachować do testów poza próbą (z ang. *out-of-sample*). Odnosi się to oczywiście do próbki testowej. Wartość tego parametru nie powinna być mniejsza, niż wartość ustawiona w parametrze n.roll. Ustaliłem go na równy 125. Wtedy funkcja ta automatycznie dzieli podany zbiór danych na próbkę uczącą i testową.

Dodatkowo, tę funkcję prognozy wywołuje się z ustalonymi parametrami modelu, otrzymanymi przy pomocy funkcji ugarchspec(). Z dopasowanego obiektu ugarchspec() najłatwiej wydobyć

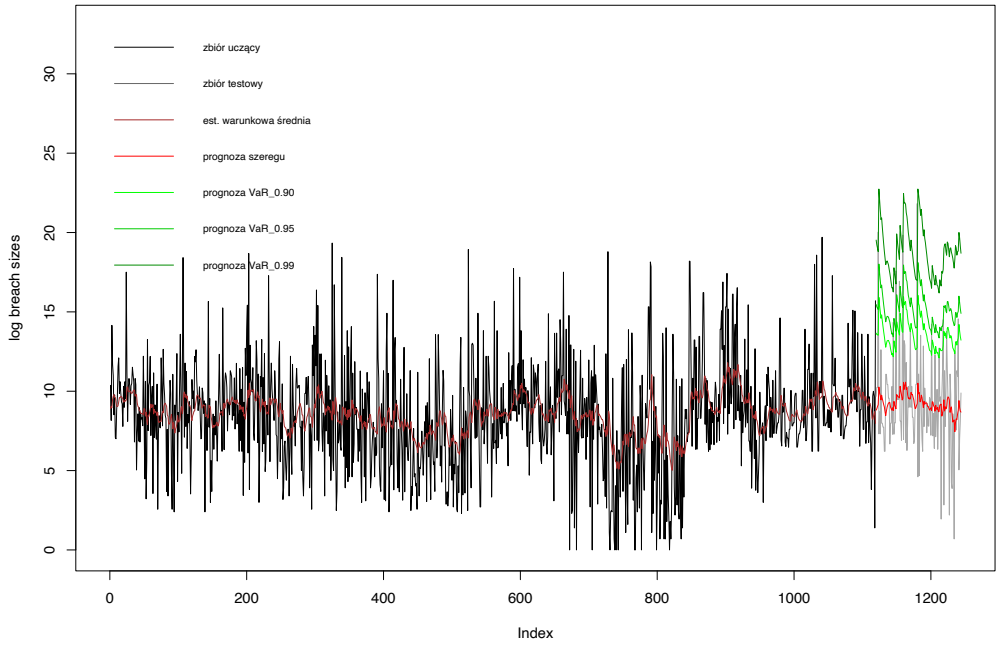
parametry korzystając z metody getspec(). Należy także pamiętać o podaniu w niej wybranego zbioru danych.

Metoda predykcji, to prognoza krocząca. Dodatkowo, model dopasowywany jest do powiększającej się próbki po każdej prognozie 1-krokowej.

Posługując się powyższą funkcją, otrzymałem prognozowane wektory średniej warunkowej $\hat{\mu}_t$ (tak jak powyżej wykorzystując metodę fitted() na dopasowanym modelu) i warunkowego odchylenia standardowego $\hat{\sigma}_t$ (analogicznie).

Podobnie jak w pierwszej metodzie predykcji z użyciem modelu ACD$(1,1)$, tutaj też autorzy artykułów [24, 63] utożsamiają średnią warunkową z y_{t_i}, czyli rozmiarami naruszeń. Chcąc znaleźć predykcję wartości narażonej na ryzyko ponownie skorzystałem z algorytmu przedstawionego na str. 98. Poniżej kroków algorytmu opisałem w skrócie, jak edytować wspomniany algorytm, aby obliczyć VaR dla modelu ARMA-GARCH. W tym wypadku, wspomniana metoda została już zaimplementowana w bibliotece rugarch. Aby otrzymać prognozę VaR_α, dla wybranych poziomów tolerancji $\alpha = 0.9$, 0.95 oraz 0.99 wystarczy na dopasowanym przez funkcję ugarchspec() obiekcie zastosować funkcję quantile() z ustalonym parametrem probs, który odpowiada poziomowi tolerancji. Chcąc liczyć predykcję VaR ręcznie musiałbym po prostu wygenerować próbki z rozkładu skośnego T-studenta, każda o długości 10000. W ten sposób otrzymałbym wektor Z_t, gdzie dla przypomnienia mamy: $Y_t = \mu_t + \varepsilon_t = \mu_t + \sigma_t \cdot Z_t$ (patrz też równ. (3.10), (3.8) i (3.9), rozdz. 3.3.2). Wektory średniej warunkowej μ_t oraz warunkowego odchylenia standardowego σ_t otrzymałem już z powyższej analizy. Wtedy analogicznie licząc kwantyl wybranego rzędu z otrzymanych próbek y_{t_i}, dla $i = h+1, ..., h+L$, otrzymałbym prognozowane wartości VaR_α.

Wyniki predykcji, do której wykorzystałem opisaną powyżej metodę, ukazałem na rys. 3.23. Przypominam, że dla lepszej przejrzystości, zmienną *breach sizes* ukazałem w skali logarytmicznej.



Rysunek 3.23: Predykcja zmiennej *breach sizes* z wykorzystaniem modelu ARMA$(1,1)$-GARCH$(1,1)$. Skala logarytmiczna.

Na rysunku 3.23 widzimy wykres szeregu czasowego w skali logarytmiecznej odp. danym z próbki uczącej (1119 obserwacji) oraz próbki testowej (125 obserwacji), dla zmiennej *breach sizes*. Zaznaczone są one odpowiednio krzywą czarną oraz krzywą szarą. Dodatkowo brązowa krzywa ukazuje oszacowaną średnią warunkową. Następnie, czerwona krzywa przedstawia prognozę szeregu. Ponadto trzema odcieniami zielonego koloru wyrysowane są predykcje wartości VaR_α, dla poziomów tolerancji kolejno $\alpha = 0.9$, 0.95 oraz 0.99 (zob. legenda). Z wykresu można odczytać, że dla rozmiarów naruszeń występuje kilka naprawdę dużych wartości (zaznaczam, że to skala logarytmiczna). W prognozie VaR widać także duże skoki, co, gdy patrzymy na kształt całego szeregu, wydaje się bardzo prawdopodobne. Powoli zmniejszające się w czasie wielkości rozmiarów naruszeń mogłyby zastanawiać, ale pod sam koniec predykcji ponownie widać gwałtowniejszy wzrost. Aby sprawdzić poprawność uzyskanej predykcji, ponownie wykorzystałem wprowadzone miary oraz analizowałem liczbę przekroczeń.

Wartość wskaźnika MAPE okazała się być niska. Wyniosła ona 0.2436383, czyli ok. 24%. Oznacza to, że średnia różnica między wartością prognozowaną, a rzeczywistą wynosi tylko 24%. W porównaniu do odpowiadających wartości dla modelu ACD$(1,1)$, ta wartość jest znacznie mniejsza. Z kolei wskaźnik MedianAPE $= 0.1711942$, czyli ok. 17%. On również jest niski. Dzięki niemu widać też, że analizowany zbiór danych miał kilka obserwacji odstających. Ponownie porównu-

jąc wartość tego wskaźnika z odpowiadającymi wartościami dla modelu ACD(1, 1), widzimy, że wartość ta jest dużo niższa. Kształt analizowanego zlogarytmowanego szeregu jest jednak mniej skomplikowany niż kształt trajektorii zmiennej *interarrival times*. Większość wartości tego szeregu waha się w przedziale ok. (4, 15). Zatem niska wartość tego wskaźnika nie jest ani trochę zaskakująca.

Przeprowadziłem również porównanie spodziewanej liczby przekroczeń z zaobserwowaną liczbą. Wykonałem także testy Kupca i Christoffersena. Wyniki przedstawiłem w tablicy 3.30.

α	Zaobs.	Spodziew.	p-val LR_{uc}	p-val LR_{cc}	Decyzja$_{uc}$	Decyzja$_{cc}$
0.90	13	12	0.88218274	0.2150243	Fail to Reject H0	Fail to Reject H0
0.95	8	6	0.4903864	0.4538441	Fail to Reject H0	Fail to Reject H0
0.99	4	1	0.05025106	0.1265971	Fail to Reject H0	Fail to Reject H0

Tablica 3.30: Wyniki testów Kupca i Christoffersena dla liczby przekroczeń przy ustalonym poziomie istotności równym 0.05. Model ARMA(1,1)-GARCH(1,1) z innowacjami o rozkładzie skośnym T-studenta.

W tablicy 3.30 ukazałem wyniki testów Kupca i Christoffersena dla liczby przekroczeń, oraz porównanie zaobserwowanych przekroczeń ze spodziewanymi. Próbka testowa liczy 125 obserwacji, zatem oczekiwana liczba przekroczeń VaR przy poziomie tolerancji $\alpha = 0.9$ to $(1 - 0.9) \cdot 125 \approx 12$; przy poziomie tolerancji $\alpha = 0.95$ to $(1 - 0.95) \cdot 125 \approx 6$; a przy $\alpha = 0.99$ to $(1 - 0.99) \cdot 125 \approx 1$. Widzimy, że dla każdego poziomu tolerancji α model niedoszacowuje wartości VaR, jednakże różnice są małe. Wyniki testów zarówno Kupca, jak i Christoffersena, przy ustalonym poziomie istotności równym 0.05, nie dają podstaw do odrzucenia hipotez zerowych. Wszystkie p – wartości testów są większe od 0.05. Najbliżej odrzucenia hipotezy zerowej jesteśmy przy poziomie tolerancji równym $\alpha = 0.99$. W tym wypadku bowiem p – wartość testu Kupca wyniosła tylko 0.050251. Ogólnie, wyniki tych testów są dobre, można więc stwierdzić, że wybrany model predykcyjny jest odpowiedni.

Wnioski

Reasumując, wskaźniki takie jak MAPE i MedianAPE mówią, że predykcja szeregu czasowego jest stosunkowo dobra. Dodatkowo, zaobserwowane liczby przekroczeń *backtestingu*, są w przybliżeniu równe oczekiwanym. Obydwa wykorzystane przeze mnie testy statystyczne nie wykazały podstaw do odrzucenia hipotezy zerowej. Wskazuje to na odpowiedniość otrzymanych prognoz. Można zatem twierdzić, że testowany przeze mnie model dobrze dopasowuje się do przeszłych rozmiarów naruszeń, a także jest odpowiedni do przewidywania przyszłych *breach sizes*.

Jako ciekawostkę, wspomnę o funkcji ugarchpath() z tej samej biblioteki. Wykorzystuje się ją w celach symulacji przyszłych trajektorii szeregu czasowego. W funkcji tej wystarczy podać specyfikację odpowiedniego modelu ARMA-GARCH. Jej podstawowe parametry, które ustaliłem do symulacji są następujące:

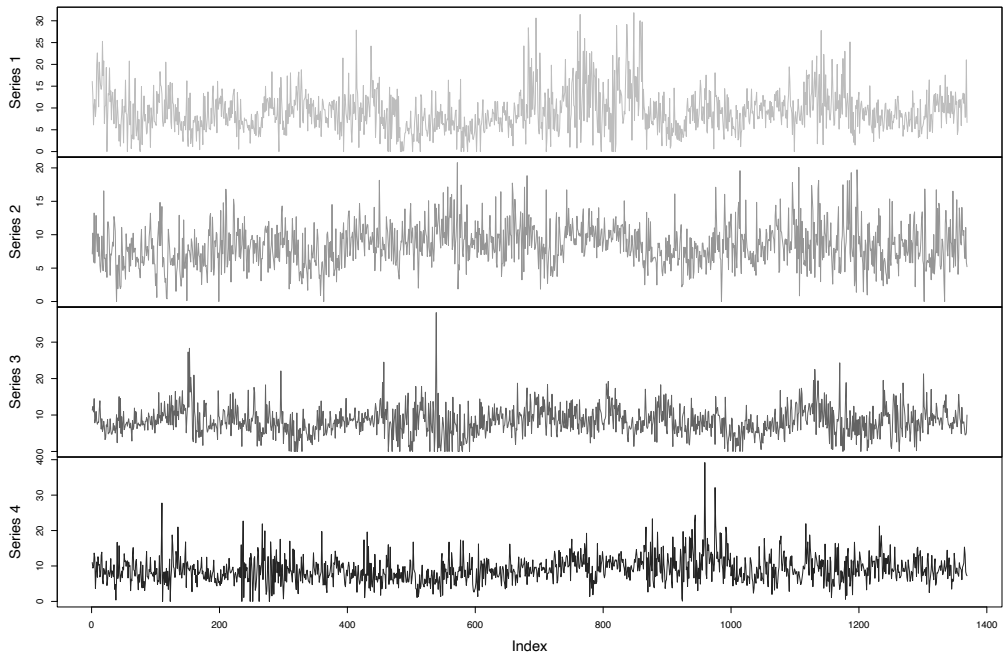
- spec, czyli obiekt klasy ugarchspec(), ze specyfikacją modelu, którą otrzymałem wykorzystując wspomnianą już metodę getspec() na dopasowanym obiekcie;

- m.sim = m, który mówi, ile razy symulacja zostanie wykonana. Parametr ten zdecydowałem się ustalić na $m = 4$;

- n.sim = n, oznacza on horyzont symulacji (por. parametr n.ahead funkcji ugarchforecast()). Zdecydowałem się ustalić go na $n = 1244 + 125 = 1369$.

Intuicyjnie mówiąc, funkcja ta przewiduje m razy n wartości. Jest to funkcja, która nie wymaga dopasowanego modelu. Zamiast tego wymagana jest tu specyfikacja odpowiedniego ustalonego modelu ARMA-GARCH.

Aby przedstawić na wykresie wyniki symulacji, wykorzystując powyższą funkcję ugarchpath(), ustaliłem otrzymane wyestymowane parametry modelu ARMA$(1,1)$-GARCH$(1,1)$, które uzyskałem z powyższej analizy. Następnie, aby dostać wektor średnich warunkowych μ_t, czyli symulacje przyszłych *breach sizes*, na tym dopasowanym obiekcie zastosowałem metodę fitted(). Wyniki symulacji przedstawiłem na rysunku 3.24.



Rysunek 3.24: Symulacja przyszłych trajektorii zmiennej zlogarytmowanej *breach sizes* z wykorzystaniem modelu ARMA$(1,1)$-GARCH$(1,1)$ z ustalonymi parametrami. Skala logarytmiczna.

Na rysunku 3.24 widzimy 4 niezależne symulacje przyszłych trajektorii zmiennej zlogarytmowanej *breach sizes*. W tym celu wykorzystałem model ARMA$(1,1)$-GARCH$(1,1)$ z ustalonymi wcześniej parametrami. Uzyskałem je przy pomocy opisanej powyżej funkcji ugarchpath(). W ogólności, taka symulacja może być przydatna do oceny losowości oraz oszacowania przyszłych wielkości *breach sizes*.

3.4. Wnioski

Podsumowując, łatwo zauważyć, że predykcja przyszłych naruszeń jest trudna. Ciężko uzyskać idealną prognozę szeregów. Wybór jedynej metody predykcyjnej również jest zadaniem kłopotliwym. Z drugiej jednak strony obydwa modele, dopasowane przeze mnie do analizowanych danych, pozytywnie przeszły oba testy statystyczne. Dodatkowo, zaobserwowane liczby przekroczeń *backtestingu*, są przynajmniej w przybliżeniu równe oczekiwanym dla obydwu prognoz. Można więc twierdzić, że rozpatrywane modele są odpowiednie do przewidywania przyszłych naruszeń oraz mogą dość efektywnie przewidywać wartości VaR zarówno dla czasów *interarrival*, jak i dla rozmiarów naruszeń. Należy jednak zwrócić uwagę na fakt, że istnieje kilka wyjątkowo dużych czasów *interarrival* oraz dużych wielkości *breach sizes*. Są one dalekie od przewidywanych

wartości VaR_α. Oznacza to, że prognoza ominęła niektóre z wyjątkowo dużych naruszeń, jakie mogłyby się okazać się bardzo niebezpieczne dla danej firmy.

4. Modele bibliometryczne

Na koniec przejdę do zwięzłego wprowadzenia teorii modeli bibliometrycznych (patrz np. [32]), a także wykorzystania jej do modelowania znanych już zmiennych *breach sizes* oraz *interarrival times*. Analizując rozkłady powyższych zmiennych zauważyłem, że do danych dobrze dopasowywały się rozkłady logarytmiczne oraz skośne. Nasuwa się więc skojarzenie z wektorami cytowań wykorzystywanymi w bibliometrii. Moim celem było zbadanie, czy modele bibliometryczne mogą mieć zastosowanie w modelowaniu tych dwóch zmiennych w obszarze cyberbezpieczeństwa. Dokładniej, interesowało mnie to, czy zastosowanie modeli, które zaproponowane są do odtwarzania wektorów cytowań (czym one są, wspomnę za chwilę), da sensowne wyniki przy modelowaniu zmiennej *breach sizes* lub *interarrival times*.

4.1. Wprowadzenie

Rozdział ten zacznę od wprowadzenia podstawowych definicji z obszaru bibliometrii, tzn. wprowadzę definicje kilku wykorzystanych przeze mnie modeli cytowań (modeli bibliometrycznych) i krótko je opiszę. Zostały one wyselekcjonowane zgodnie z dostępną mi literaturą (patrz np. [32, 33, 35]). Wybrane modele próbowałem dopasować do posiadanych danych, pobranych ze wspomnianej już strony PRC [25]. Na koniec przedstawię wyniki, które otrzymałem wykorzystując tym razem język Python.

Zazwyczaj wyznaczane są one na podstawie reprezentacji osiągnięć naukowca tj. wektora cytowań. Formalnie, wektor cytowań definiujemy w następujący sposób.

Definicja 4.1 (wektor cytowań). Wektor cytowań $x = (x_1, \ldots, x_N)$, taki, że:

$$x_1 \geqslant x_2 \geqslant \cdots \geqslant x_N \geqslant 0, \tag{4.1}$$

to wektor (najczęściej liczb całkowitych), gdzie x_i-ta wartość oznacza liczbę cytowań i-tej publikacji, natomiast N to liczba publikacji danego autora. Dokładniej, x_1 to wartość najczęściej cytowanej pracy, x_2 to wartość drugiej najczęściej cytowanej pracy, itd.

Uwaga 4.2. Odnotujmy, że rozkład wektora cytowań cechuje się dużą skośnością. Stąd, są one często modelowane np. przez rozkład Pareto II rodzaju. Innymi słowy, takie wektory mają tę cechę, że często mamy kilka dużych wartości oraz wiele wartości małych lub zerowych.

Uwaga 4.3. Na podstawie wektorów cytowań wyznaczane są tzw. wskaźniki (indeksy) bibliometryczne. Mają one realny wpływ m.in. na politykę kadrową instytucji naukowych czy otrzymywanie grantów. Przykładem takiego wskaźnika jest np. indeks h, którego definicja znajduje się poniżej.

Definicja 4.4 (Wskaźnik h). Wskaźnik h (z ang. h – *index*, wprowadzony przez J. E. Hirscha w 2005 r.) dany jest następującym równaniem:

$$h(x) = \max\{H = 1, ..., N : x_H \geqslant H\}. \tag{4.2}$$

Jest to miara, która ma na celu uwzględnienie nie tylko ogólnej jakości publikacji, ale także ich liczby. Mówiąc prościej, dany autor ma wskaźnik h równy H, jeśli H z jego N wszystkich publikacji ma co najmniej H cytowań każda, a pozostałe $(N - H)$ publikacji mają nie więcej niż H cytowań każda. Wskaźnik h jest przykładem najpopularniejszego wskaźnika bibliometrycznego.

4.2. Modele cytowań

Wskaźniki bibliometryczne często kładą nacisk na przedstawienie interpretowalnego podsumowania liczbowego. Z kolei modele cytowań mają na celu odtworzenie w całości oryginalnych wektorów, przy użyciu kilku podstawowych parametrów. Poniżej przedstawię wybrane modele, klasycznie wykorzystywane w bibliometrii. Co istotne, modele te pozwalają na wyznaczenie jawnych wzorów na przewidywaną liczbę cytowań dla k-tej najczęściej cytowanej pracy, którą oznaczyłem przez $\hat{x}_k$, dla $k = 1, ..., N$, dla danego N. Dalej, przejdę do przedstawienia kolejnych modeli cytowań.

Model Power - law

Pierwszym z modeli cytowań jest model *power - law* (patrz [33, 34]). Historycznie był on jednym z pierwszych tego typu modeli. Wzór na liczbę cytowań dla k-tej najczęściej cytowanej pracy wg. tego modelu jest określony następująco:

$$\hat{x}_k^{PowerLaw}(N, \alpha, \gamma) = \frac{\gamma}{k^\alpha}, \tag{4.3}$$

gdzie $\gamma > 0$ i $\alpha > 0$.

W praktyce parametry skali $\gamma > 0$ i eksponenta $\alpha > 0$ muszą być szacowane na podstawie danych.

Model log - normalny

Kolejny rozważany przeze mnie model to model log - normalny (z ang. *log - normal*, patrz [50, 51, 52, 53, 54]). Niech funkcja przeżycia rozkładu przesuniętego log - normalnego będzie dana wzorem:

$$S_{l,\mu,\sigma}(x) = 1 - \Phi\Big(\frac{\log(x-l)-\mu}{\sigma}\Big),$$

gdzie Φ jest dystrybuantą rozkładu standardowego normalnego z parametrami μ oraz σ. Dodatkowo, aby zwiększyć funkcjonalność modelu, l jest przesunięciem. Innymi słowy, jeśli $\log(X)$ jest zmienną losową o rozkładzie $N(0,\sigma)$ oraz $Y = (\log(X)-l)/e^{\mu}$ jest jej przesuniętą i przeskalowaną wersją, to $S_{l,\mu,\sigma}(x) = \mathbb{P}(Y > x)$. Wtedy estymowaną liczbę cytowań k-tej najczęściej cytowanej publikacji wyznacza się w oparciu o odwrotność funkcji przeżycia $S_{l,\mu,\sigma}$:

$$\hat{x}_k^{LogNormal}(N,l,\mu,\sigma) = S^{-1}_{l,\mu,\sigma}\Big(\frac{k}{N+1}\Big), \tag{4.4}$$

gdzie $l, \mu \in \mathbb{R}$, a $\sigma > 0$.
Należy wspomnieć, że nawet jeśli takie podejście opiera się na pewnych znanych obiektach z teorii rachunku prawdopodobieństwa, to nie można zakładać, że w prawdziwym świecie cytowania są niezależne i o losowym rozkładzie.

Model DGBD

Kolejnym przykładem uogólnienia modelu *power - law* jest model DGBD (z ang. *discrete generalised beta distribution*). Podejście w tym modelu polega na wykorzystaniu dyskretnego uogólnionego rozkładu beta (patrz [35, 36]). W przypadku wektorów cytowań, estymowaną liczbę cytowań k-tej najczęściej cytowanej publikacji można wyrazić przez:

$$\hat{x}_k^{DGBD}(N,A,a,b) = A \cdot \frac{(N+1-k)^b}{k^a}, \tag{4.5}$$

gdzie $A > 0$, $a > 0$ i $b \geqslant 0$.
Zauważmy, że równ. (4.5) sprowadza się do równ. (4.3) gdy $b = 0$. Ponadto, zwiększona funkcjonalność (ze względu na dodatkowy parametr) pozwala na lepsze dopasowanie do często cytowanych publikacji, jak również na lepsze dopasowanie do ogona rozkładu empirycznego.

Model 3DSI

Kolejna propozycja modelu to model 3DSI (czyli *3D model for scientific impact*) zaproponowany w [32]. Zakłada się w nim, że cytowania danego autora są rozłożone pomiędzy jego

publikacje za pomocą dwóch czynników: ze względu na regułę preferencji lub w sposób czysto losowy. Zależy on od 3 parametrów: N - liczby opublikowanych artykułów, C - całkowitej liczby cytowań i $\rho \in (0,1)$ - stosunku preferencji do przypadkowości. W szczególności, $\rho = 0$ oznacza, że wszystkie publikacje otrzymują cytowania zupełnie losowo, a $\rho = 1$ oznacza, że wszystkie cytowania są zgodne z regułą "bogaci stają się bogatsi" (z ang. *rich-get-richer*, patrz [32]).

Wtedy estymowana liczba cytowań dla k-tej najczęściej cytowanej publikacji jest dana jest wzorem:

$$\hat{x}_k^{\mathrm{3DSI}}(N,C,\rho) = \frac{1-\rho}{\rho}\frac{C}{N}\left(\prod_{i=k}^{N}\frac{i}{i-\rho}-1\right) = \frac{1-\rho}{\rho}\frac{C}{N}\left[\frac{\Gamma(k-\rho)}{\Gamma(k)}\frac{\Gamma(N+1)}{\Gamma(N+1-\rho)}-1\right], \tag{4.6}$$

gdzie $\rho \in (0,1)$, oraz $\Gamma(k) = \int_0^\infty t^{k-1}\,e^{-t}\,dt$, $k > 0$ to funkcja gamma.

Zaletą tego modelu jest fakt, że wszystkie jego parametry są łatwe w interpretacji, przy czym $C = \sum_{k=1}^{N} \hat{x}_k$ i ρ kontroluje stopień skośności rozkładu cytowań.

4.3. Analizy empiryczne

4.3.1. Wykorzystane dane

W niniejszym podrozdziale przeprowadzę analizę empiryczną. W tym celu wykorzystam wspomniane powyżej modele. Dane, z których skorzystałem do analizy w swojej pracy, pobrałem ze wspomnianej strony PRC [25]. Przygotowałem je w sposób podobny do tego, jaki opisałem w podrozdz. 3.1. Oczyściłem dane, tzn. usunąłem wartości 0 i NA dla zmiennej *breach sizes*. Dodatkowo, pamiętając o równaniu (4.1), wartości w tym wektorze posortowałem malejąco. Pozostało mi dokładnie 6822 obserwacji. Ponadto, dla lepszej przejrzystości wyników na prezentowanych przeze mnie wykresach, rozpatrywałem też zmienną zlogarytmowaną *breach sizes*. W tym wypadku danych zostało mniej (tj. dokładnie 6705), gdyż niektóre rozmiary naruszeń były równe 1. Po zlogarytmowaniu były już równe 0, więc usunąłem je z danych.

Druga rozpatrywana przeze mnie zmienna to *interarrival times*. Wektor ten oczyściłem analogicznie jak rozmiary naruszeń, następnie uporządkowałem malejąco jego wartości.

Dodatkowo cały zbiór danych podzieliłem na mniejsze podzbiory odpowiadające poszczególnym typom naruszeń oraz typom organizacji. Na każdym z podzbiorów przeprowadziłem analogiczne analizy.

Moim celem było zbadanie, czy modele bibliometryczne mogą mieć zastosowanie w modelowaniu zmiennych *interarrival times* oraz *breach sizes* w przypadku tematu cyberbezpieczeństwa.

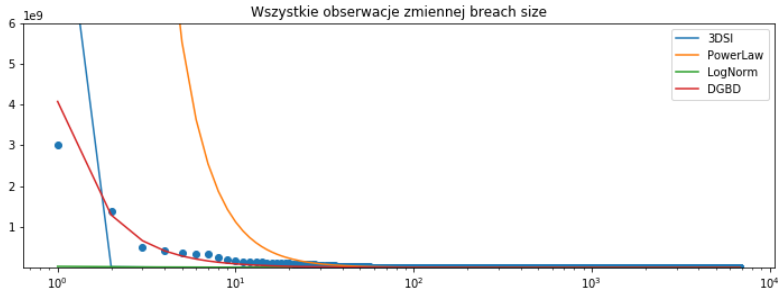
4.3.2. Wyniki

Jak już wspomniałem w poprzednim podrozdziale, do analizy zdecydowałem się wykorzystać język Python. Główne biblioteki, z których w tym celu skorzystałem to: numpy, matplotlib, pandas, seaborn, json oraz scipy.stats. Dodatkowo wykorzystałem udostępniony mi zbiór modułów nazwany model_fitting utworzony przez M. Gągolewski i B. Żogała-Siudem w 2021 r. (patrz [79]). Każdy z wymienionych przeze mnie modeli został tu już zaimplementowany jako klasa języka Python.

Zacznę od przedstawienia wyników analizy, którą przeprowadziłem na zmiennej *breach sizes*, a następnie przejdę do ukazania wyników dot. zmiennej *interarrival times*.

Zmienna *Breach sizes*

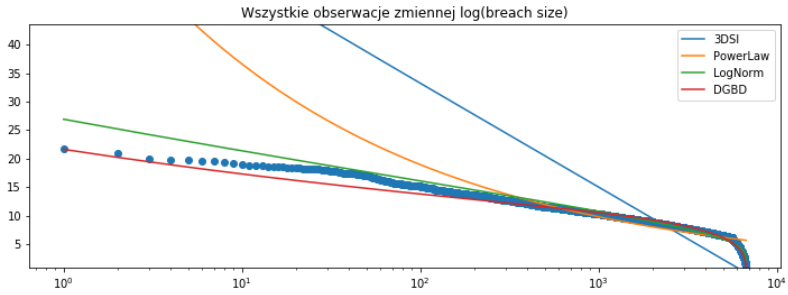
Na rysunku 4.1 przedstawiłem wyniki dopasowania proponowanych modeli na całym wektorze *breach sizes*.



Rysunek 4.1: Wyniki dopasowania dla zmiennej *breach sizes*.

Widzimy tu wyniki dopasowania modeli dla zmiennej *breach sizes*. Jest to wykres punktowy oryginalnych danych. Dodatkowo kolorem niebieskim oznaczony został model 3DSI, kolorem pomarańczowym – model *power - law*, kolorem czerwonym – model DGBD oraz kolorem zielonym – model Log - normalny (podpisany jako LogNorm).

Widzimy, że najlepiej do danych pasuje model DGBD.



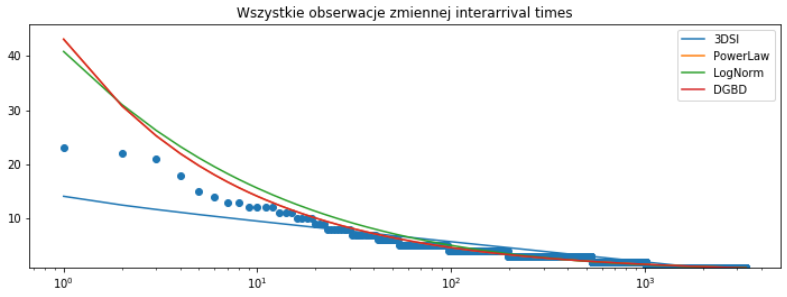
Rysunek 4.2: Wyniki dopasowania dla zmiennej zlogarytmowanej *breach sizes*.

Rysunek 4.2 przedstawia wyniki dopasowania modeli dla zmiennej *breach sizes*, ale zlogarytmowanej. Ponownie najlepiej do całych danych pasuje model DBGD. Przy czym sprawdził się tu również model Log - normalny.

Wykonałem także analogiczne badania dla różnych typów naruszeń oraz typów organizacji. Otrzymałem wyniki prowadzące do wniosków podobnych do powyższych. Jednak ze względu na przejrzystość i objętość pracy zdecydowałem się nie umieszczać wykresów z tych badań w tym miejscu.

Zmienna *Interarrival times*

Przejdę teraz do analizy dotyczącej zmiennej *interarrival times*. Na rysunku 4.3 przedstawiłem wyniki dopasowania tych samych modeli, ale dla czasów *interarrival*.



Rysunek 4.3: Wyniki dopasowania dla zmiennej *interarrival times*.

Tym razem widzimy, że potencjalnie dopasowanych modeli mogłoby być więcej. Po pierwsze, model Power Law nie jest widoczny na tym rysunku, z uwagi na to, że nachodzi się on z modelem DBGD. Ogólnie widzimy, że wszystkie 3 modele nie są dopasowane do danych wystarczająco dobrze. Model 3DSI dla większych wartości dopasowuje się dość dobrze, ale dla początkowych mocno odstaje. Pozostałe modele łapią kształt w tym sensie, że odstają w początkowych wartościach, jednakże wyłapują krzywiznę widoczną dla analizowanych danych.

Wnioski

Wykorzystanie narzędzi bibliometrycznych, a konkretnie wybranych przeze mnie: Power law, Log - normalny, 3DSI oraz DBGD, doprowadziło do interesujących wniosków. Po pierwsze, w przypadku zmiennej *breach sizes* widzimy, że model DBGD dopasowuje się dość dobrze (nawet dla początkowych wartości). Wskazuje to obiecujący, nowy kierunek badań związanych z modelowaniem tej zmiennej. Dalej, w przypadku zmiennej *interarrival times* modele DBGD, Power law, jak i Log - normalny pasują do danych, choć dla początkowych wartości odstają od dopasowanych krzywych. Interpretacyjnie ciekawy jest fakt, że w przypadku danych określających różnicę w czasach ataków otrzymujemy dość dobre wyniki dla wybranych modeli bibliometrycznych.

Zakończenie

Celem niniejszej pracy był przegląd oraz usystematyzowanie aktualnego stanu wiedzy z zakresu ubezpieczeń w obszarze cyberbezpieczeństwa. W związku z tym wykonałem studium literaturowe dotyczące ubezpieczeń w tym obszarze, tzn. porównałem istniejące definicje problemu oraz przygotowałem wykaz cyber zagrożeń podlegających ubezpieczeniu, np. eksfiltracja danych, naruszenie poczty e-mail, "oszustwo na CEO", różnorakie infekcje złośliwym oprogramowaniem lub ataki DDoS. Następnie przeanalizowałem wybrane oferty cyberubezpieczeń takich firm jak: PZU S.A., Findia Insurance, AXA XL Insurance lub Travelers Indemnity Company, dostępne zarówno na polskim jak i amerykańskim rynku.

Wybrałem również i opisałem zaproponowane w literaturze sposoby modelowania w obszarze cyberubezpieczeń. Ogólnie rozróżniłem dwa podejścia:

- modelowanie zmiennych określających rozmiar naruszeń oraz czasy między naruszeniami (odpowiednio, zmienna *breach size* oraz *interarrival times*);
- modelowanie całej sieci (np. sieci komputerowej, sieci oddziałów przedsiębiorstwa itd.).

W przypadku podejścia modelowania całej sieci, przedstawiłem proponowaną metodologię opartą na pracy [1]. Ze względu jednak na wymóg posiadania informacji dotyczących topologii całej sieci oraz szczegółowych założeń dotyczących np. wykorzystywanych modeli Markowa, podejście to jest trudne do odtworzenia i analizy. Dlatego też pozostawiam je jako przyszły kierunek badań.

Jeśli chodzi o modelowanie zmiennych *breach sizes* oraz *interarrival times*) udało mi się znaleźć otwarte źródła danych dotyczących cyberubezpieczeń. Dzięki temu miałem możliwość przeprowadzenia szerokiej analizy porównawczej omówionych metod. Początkowo wykorzystałem podejście oparte o modelowanie przy użyciu rozkładów prawdopodobieństwa. W pracy [23] przeprowadzona przez autorów analiza uwzględniała wykorzystanie metody skalowania wielowymiarowego (MDS). Otrzymali oni wyniki wskazujące, iż różne typy naruszeń danych muszą być modelowane jako odrębne kategorie ryzyka. W oparciu o powyższe wnioski, rozpatrywałem zarówno dopasowanie rozkładu do całego zestawu danych jak i do grup wyznaczonych przez różne typy naruszeń oraz różne typy organizacji (tzn. klasyfikację tematyczną przedsiębiorstw).

Otrzymane wyniki prowadzą do następujących wniosków:

- W przypadku zmiennej *breach sizes* najbardziej obiecujące jest wykorzystanie rozkładu skośnego log - normalnego lub skośnego T-studenta po wcześniejszym przekształceniu logarytmicznym zmiennej. Wniosek ten pokrywa się z wynikami publikacji [23], w której autorzy zdecydowali się wybrać właśnie ten rozkład obok rozkładu skośnego - normalnego.

- Z kolei dla zmiennej *interarrival times* próby zbadania zgodności z wybranymi rozkładami prawdopodobieństwa nie prowadzą do żadnych sensownych wniosków. Jedynie według testu Chi - kwadrat rozkład „pasujący" do danych, to ujemny dwumianowy dla typów naruszeń CARD oraz STAT.

Widzimy zatem, że dla obydwu zmiennych modelowanie rozkładami prawdopodobieństwa jest trudne i daje wyniki niejednoznaczne w interpretacji. Warto również wspomnieć, że weryfikacja formalna hipotez o zgodności wymaga często szacowania parametrów rozkładów. To z kolei prowadzi do konieczności wykorzystania modyfikacji znanych testów statystycznych.

Nie jest to jednak jedyne znane podejście w literaturze (patrz rozdz. 2.2). Przejdę teraz do omówienia wyników z podejścia stochastycznego. Przede wszystkim zbadałem autokorelacje dla rozpatrywanych zmiennych. Zaobserwowałem dość duże wartości dla *interarrival times* oraz nieco mniejsze, ale wciąż wyraźne, dla zmiennej *breach size*. Otrzymane wyniki potwierdzają to, co zasugerowali autorzy [24], że czasy między zdarzeniami oraz rozmiary naruszeń powinny być modelowane przez procesy stochastyczne, a nie przez rozkłady.

W przypadku zmiennej *breach sizes* najlepiej dopasowanym modelem był model ARMA$(1,1)$-GARCH$(1,1)$ z innowacjami o rozkładzie skośnym T-studenta. Dla zmiennej *interarrival times* modelem, który najlepiej pasował do danych był model ACD$(1,1)$ z innowacjami pochodzącymi z uogólnionego rozkładu Gamma. Rozważałem dopasowania modeli także w przypadku grup wyznaczonych przez różne typy organizacji oraz różne typy naruszeń. Wyniki uzyskane w tych grupach w kilku przypadkach prowadziły do wyboru innych modeli np. dla typów DISC, STAT oraz PHYS wartości wykorzystanych kryteriów informacyjnych były najmniejsze dla modelu LACD_2. Jednakże ponieważ różnica między modelem ACD, a LACD_2 była bardzo mała, to zdecydowałem się wybrać prostszy model. Ponadto dla typu CARD najlepszym modelem okazał się LACD_1. Należy jednak pamiętać, że dla tego właśnie typu dostępne są tylko 32 obserwacje.

Dalej, w celu sprawdzenia zależności pomiędzy zmiennymi *interarrival times*, a *breach sizes*, obliczyłem empiryczne współczynniki korelacji ρ Spearmana oraz τ Kendalla pomiędzy obydwiema zmiennymi. Dla każdego typu naruszenia, ich wartości prowadziły do tych samych wniosków. Podobnie, nieparametryczne testy rangowe, zarówno dla korelacji Spearmana, jak i Ken-

dalla, nie dały podstaw do odrzucenia hipotezy zerowej. Okazało się więc, że wyniki dla każdego z typów naruszeń nie wskazują na istnienie zależności pomiędzy analizowanymi zmiennymi.

Porównując swoje wyniki z tymi przedstawionymi w pracy Xu i inni [24] starałem się dopasować kopułę do danych. Zbadałem zarówno podejście parametryczne, jak i nieparametryczne do dopasowania kopuły, estymacji jej parametrów i parametrów rozkładów brzegowych. Wyniki potwierdzają jednak powyższe wnioski o braku zależności obydwu zmiennych. Okazało się więc, że, w przeciwieństwie do wyników z pracy [24], uzyskane przeze mnie wyniki nie wskazują na istnienie zależności pomiędzy *interarrival times*, a *breach sizes*.

Następnie przeprowadziłem analizę związaną z wartością narażoną na ryzyko. By wyznaczyć odpowiednie wartości konieczne było uzyskanie predykcji na podstawie modelu. Zadanie to jednak nie jest oczywiste w przypadku modelu ACD$(1,1)$. Rozważyłem więc trzy podejścia zaproponowane w literaturze oraz zaproponowałem do nich pewne modyfikacje. W pierwszej metodzie dopasowanie modelu do danych odbywa się tylko raz – na samym początku. Następnie otrzymane predykcje średnich warunkowych to kolejne wartości prognozy. Druga metoda jest podobna do pierwszej, jednakże w tym wypadku, w kolejnych krokach prognozy kroczącej wykorzystywane były dane z próbki testowej. Trzecia metoda była bardziej skomplikowana obliczeniowo. Metoda ta różni się jednak liczbą dopasowań modelu. Tutaj dopasowywałem model do danych po każdej prognozie 1-krokowej. Metoda symulacji jaką wykorzystałem w celu predykcji przyszłych *interarrival times* to symulacja metodą Monte – Carlo.

Przy rozpatrywaniu tylko i wyłącznie testów Kupca i Christoffersena okazuje się, że pierwsza metoda wypada najlepiej, a najgorzej – metoda druga. Różnice pomiędzy nimi nie są jednak aż takie duże. Z kolei patrząc na wskaźniki MAPE oraz MedianAPE stwierdzamy, że najlepiej prezentuje się predykcja otrzymana metodą drugą, a najgorzej – metodą trzecią. W tym wypadku różnice pomiędzy dwiema pierwszymi metodami są nieznaczne, natomiast metoda trzecia jest znacznie gorsza. Z kolei patrząc subiektywne na trajektorie tych szeregów czasowych możemy stwierdzić, że trzecia metoda wydaje się być stosunkowo dobra, jednakże losowość otrzymanych wartości predykcji jest tutaj bardzo duża. Reasumując, predykcja zmiennej *interarrival times* jest trudna. Przez niestandardowy, skokowo – malejący, ale też wykazujący pewną okresowość charakter trajektorii, ciężko jest uzyskać sensowną prognozę szeregu. Wyniki przedstawione przez autorów artykułu [24] również nie były w stanie przewidzieć przyszłych dużych skoków.

Wybór jednej metody predykcji jest zadaniem ciężkim. Jeśli jednak miałbym wybrać tylko jedną, to najłatwiejszym wyborem jest metoda pierwsza. Jest ona prosta w implementacji, a jej wyniki można uznać za wystarczające. Minusem jest jednak kształt przyszłych trajektorii. Prosty charakter tej metody nie odzwierciedla skokowej natury analizowanej zmiennej.

W przypadku zmiennej *breach size* wskaźniki takie jak MAPE i MedianAPE mówią, że predykcja szeregu czasowego jest stosunkowo dobra. Dodatkowo, zaobserwowane liczby przekroczeń *backtestingu* są w przybliżeniu równe oczekiwanym. Obydwa wykorzystane przeze mnie testy statystyczne nie wykazały podstaw do odrzucenia hipotezy zerowej. Wskazuje to na odpowiedniość otrzymanych prognoz. Można zatem twierdzić, że testowany przeze mnie model dobrze dopasowuje się do przeszłych rozmiarów naruszeń, a także jest odpowiedni do przewidywania przyszłych *breach sizes*.

Podsumowując, model ACD$(1,1)$ dla *interarrival times* oraz ARMA$(1,1)$-GARCH$(1,1)$ dla *breach size*, dopasowane przeze mnie do analizowanych danych, pozytywnie przeszły oba testy statystyczne. Dodatkowo, zaobserwowane liczby przekroczeń *backtestingu*, były przynajmniej w przybliżeniu równe oczekiwanym dla obydwu prognoz. Można więc twierdzić, że rozpatrywane modele są odpowiednie do przewidywania przyszłych naruszeń oraz mogą dość efektywnie przewidywać wartości VaR zarówno dla czasów *interarrival*, jak i dla rozmiarów naruszeń. Należy jednak zwrócić uwagę na fakt, że istnieje kilka wyjątkowo dużych czasów *interarrival* oraz dużych wielkości *breach sizes*. Są one dalekie od przewidywanych wartości VaR_α. Oznacza to, że prognoza ominęła niektóre z wyjątkowo dużych naruszeń, jakie mogłyby się okazać bardzo niebezpieczne dla danej firmy.

Dodatkowo w rozdz. 4 wykorzystałem wybrane modele bibliograficzne, by sprawdzić ich użyteczność w przypadku modelowania zmiennych *breach size* oraz *interarrival times*. Okazało się, że w przypadku zmiennej *breach sizes* model DBGD dopasowuje się dość dobrze (nawet dla początkowych wartości). Dalej, w przypadku zmiennej *interarrival times* modele DBGD, Power law, jak i Log - normalny pasują do danych, choć dla początkowych wartości odstają od dopasowanych krzywych.

Przeprowadzone przez mnie badania wskazują na ciekawe kierunki badań w obszarze cyberbezpieczeństwa. Przede wszystkim dogłębnej analizy wymaga zaproponowany w literaturze sposób modelowania przy użyciu całej sieci, ze szczególnym uwzględnieniem przewidywania wartości VaR. Zadanie predykcji dla modelu ACD$(1,1)$ również wymaga dalszych analiz. Uzyskane wyniki w przypadku zastosowania modeli bibliometrycznych otwierają nowy i obiecujący kierunek badań związany z wykorzystaniem tych narzędzi w zupełnie nowym kontekście.

A. Dodatek

W tym dodatku zamieściłem tabelki dot. analizy zgodności rozkładów (patrz podrozdz. 3.3.1) z wynikami przeprowadzonych przeze mnie badań dla różnych typów naruszeń i organizacji. Poniższe wyniki prowadziły do podobnych wniosków, które zaprezentowałem we wspomnianym podrozdziale.

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.17129	0.90181
Normal	< 0.01	0.07289
Weibull	< 0.01	0.11743
Skew-Normal	< 0.01	0.25278
Log-Logistic	0.58837	0.00373
Skew-Student	< 0.01	0.00060
GPD	0.86241	0.44796

Tablica 1.1: Wyniki analizy na danych oryginalnych: typ CARD (32 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.09113
Normal	< 0.01	0.00058
Weibull	< 0.01	0.00058
Skew-Normal	< 0.01	0.00058
Log-Logistic	< 0.01	0.58808
Skew-Student	< 0.01	0.00058
GPD	< 0.01	0.20102

Tablica 1.2: Wyniki analizy na danych oryginalnych: typ DISC (1553 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00059
Gamma	< 0.01	0.00059
Log-Normal	< 0.01	0.08238
Normal	< 0.01	0.00059
Weibull	< 0.01	0.19767
Skew-Normal	< 0.01	0.00059
Log-Logistic	< 0.01	0.00059
Skew-Student	< 0.01	0.00059
GPD	< 0.01	0.10673

Tablica 1.3: Wyniki analizy na danych oryginalnych: typ HACK (1603 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00057
Gamma	< 0.01	0.00057
Log-Normal	0.07448	0.91368
Normal	< 0.01	0.00057
Weibull	< 0.01	0.74135
Skew-Normal	< 0.01	0.00057
Log-Logistic	0.15940	0.77703
Skew-Student	< 0.01	0.00057
GPD	0.02603	0.88168

Tablica 1.4: Wyniki analizy na danych oryginalnych: typ INSD (376 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.34764
Normal	< 0.01	0.00058
Weibull	< 0.01	0.04661
Skew-Normal	< 0.01	0.00058
Log-Logistic	< 0.01	0.00058
Skew-Student	< 0.01	0.00058
GPD	< 0.01	0.75331

Tablica 1.5: Wyniki analizy na danych oryginalnych: typ PHYS (1474 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	< 0.01	0.62512
Normal	< 0.01	0.00060
Weibull	< 0.01	0.39810
Skew-Normal	< 0.01	0.00060
Log-Logistic	< 0.01	0.00062
Skew-Student	< 0.01	0.00060
GPD	0.22313	0.15651

Tablica 1.6: Wyniki analizy na danych oryginalnych: typ PORT (874 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.05773	0.14537
Normal	< 0.01	0.00065
Weibull	< 0.01	0.37447
Skew-Normal	< 0.01	0.03162
Log-Logistic	0.58819	0.94660
Skew-Student	< 0.01	0.00060
GPD	0.41598	0.90030

Tablica 1.7: Wyniki analizy na danych oryginalnych: typ STAT (184 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00058
Gamma	< 0.01	0.00058
Log-Normal	< 0.01	0.26657
Normal	< 0.01	0.00058
Weibull	< 0.01	0.27620
Skew-Normal	< 0.01	0.00058
Log-Logistic	< 0.01	0.11021
Skew-Student	< 0.01	0.00058

Tablica 1.8: Wyniki analizy na danych oryginalnych: typ UNKN (637 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00057
Gamma	< 0.01	0.00057
Log-Normal	0.35615	0.34556
Normal	< 0.01	0.00060
Weibull	< 0.01	0.01342
Skew-Normal	< 0.01	0.00057
Log-Logistic	0.21041	0.08301
Skew-Student	< 0.01	0.00057
GPD	0.14033	0.44721

Tablica 1.9: Wyniki analizy na danych oryginalnych: typ BSF (81 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.09036	0.84711
Normal	< 0.01	0.00063
Weibull	< 0.01	0.00470
Skew-Normal	< 0.01	0.00063
Log-Logistic	0.21733	0.80566
Skew-Student	< 0.01	0.00060
GPD	0.18652	0.46025

Tablica 1.10: Wyniki analizy na danych oryginalnych: typ BSO (180 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00057
Gamma	< 0.01	0.00057
Log-Normal	< 0.01	0.26179
Normal	< 0.01	0.00057
Weibull	< 0.01	0.00057
Skew-Normal	< 0.01	0.00060
Log-Logistic	0.01630	0.03604
Skew-Student	< 0.01	0.00057
GPD	0.20516	0.71990

Tablica 1.11: Wyniki analizy na danych oryginalnych: typ BSR (107 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.47977	0.93815
Normal	< 0.01	0.00060
Weibull	< 0.01	0.12506
Skew-Normal	< 0.01	0.00060
Log-Logistic	0.51842	0.41985
Skew-Student	< 0.01	0.00060
GPD	< 0.01	0.58197

Tablica 1.12: Wyniki analizy na danych oryginalnych: typ EDU (212 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.76151	0.79268
Normal	< 0.01	0.00060
Weibull	< 0.01	0.02544
Skew-Normal	< 0.01	0.00060
Log-Logistic	< 0.01	0.01038
Skew-Student	< 0.01	0.00060
GPD	0.02187	0.29877

Tablica 1.13: Wyniki analizy na danych oryginalnych: typ GOV (83 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	< 0.01	0.10646
Normal	< 0.01	0.00060
Weibull	< 0.01	0.00061
Skew-Normal	< 0.01	0.00060
Log-Logistic	< 0.01	0.00060
Skew-Student	< 0.01	0.00060
GPD	< 0.01	0.06626

Tablica 1.14: Wyniki analizy na danych oryginalnych: typ MED (553 obserwacji).

Model	**Kolmogorov.Smirnov.Test***	**Anderson.Darling.Test**
Exponential	< 0.01	0.00060
Gamma	< 0.01	0.00060
Log-Normal	0.88245	0.92098
Normal	< 0.01	0.05760
Weibull	< 0.01	0.55528
Skew-Normal	< 0.01	0.15734
Log-Logistic	0.94553	0.49966
Skew-Student	< 0.01	0.00060
GPD	0.48604	0.86425

Tablica 1.15: Wyniki analizy na danych oryginalnych: typ NGO (21 obserwacji).

Podstawowa analiza

Poniżej przedstawiłem tablice analogiczne do 3.20 oraz 3.17. Zawierają one zestawienie podstawowych statystyk dotyczących rozmiarów naruszeń oraz czasów między zdarzeniami.

	Min	Median	Mean	SD	Max	Total
BSF	7	2078	1483462.22	9736884.3	1.300e+08	334
BSO	2	4845	18107900.21	173407485.3	3.000e+09	382
BSR	8	883	2005144.90	11473986.0	1.016e+08	241
EDU	12	2493	38706.02	323344.6	7.500e+06	620
GOV	8	3000	409663.78	3820229.9	7.600e+07	539
MED	1	2012	72531.65	1444691.3	7.880e+07	3207
NGO	13	1871	71171.96	378180.0	3.000e+06	69
Suma	1	2147	1553831.71	46483791.1	3.000e+09	5392

Tablica 1.16: Tablica przedstawiającą analogiczne statystyki (patrz tab. 3.20) dla rozmiarów naruszeń, łącznie dla wszystkich typów naruszeń (kolejno: minimum, mediana, średnia, odchylenie standardowe, maksimum i liczba obserwacji).

	Min	Median	Mean	SD	Max	Total
BSF	0.80914	21600.00000	47738.90708	86704.28548	801948.9520	333
BSO	0.01335	6.16098	11.61744	14.58943	150.0000	381
BSR	0.03020	7.00000	18.15298	39.94262	444.7154	240
EDU	0.00251	4.00000	7.16561	12.87056	167.7334	619
GOV	0.00563	4.00000	8.04833	12.42315	131.5024	538
MED	1.30395	1620.84585	4973.26739	13069.09042	327600.0000	3206
NGO	0.28415	25.45354	56.50143	79.08660	456.7810	68
Suma	0.00022	0.42582	0.82866	1.18376	21.0000	5390

Tablica 1.17: Tablica przedstawiająca analogiczne statystyki (patrz tab. 3.17) dla czasów między zdarzeniami, łącznie dla wszystkich typów naruszeń (jednostka: dzień), (kolejno: minimum, mediana, średnia, odchylenie standardowe, maksimum i liczba obserwacji).

Spis rysunków

Spis tablic

Bibliografia

[1] Xu M., Hua L., *Cybersecurity Insurance: Modeling and Pricing*, North American Actuarial Journal 23(2), 2019, 220–249,

[2] Camillo M., *Cyber risk and the changing role of insurance*, JOURNAL OF CYBER POLICY 2(1), 2017, 53-63,

[3] Elnagdy S. A., Qiu M., Gai K., *Understanding Taxonomy of Cyber Risks for Cybersecurity Insurance of Financial Industry in Cloud Computing*, 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing, 2016, 295-300,

[4] `https://www.cisa.gov/cybersecurity-insurance` [dostęp: 03.03.2021r.],

[5] `https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa_cyber_risk_for_insurers_sept2019.pdf` [dostęp: 03.03.2021r.],

[6] `https://www.soa.org/resources/research-reports/2015/research-emerging-risks-survey-reports/` [dostęp: 03.03.2021r.],

[7] `https://www.iaisweb.org/page/supervisory-material/application-papers//file/77763/application-paper-on-supervision-of-insurer-cybersecurity` [dostęp: 03.03.2021r.],

[8] Van Mieghem, P., *Performance Analysis of Complex Networks and Systems*, Cambridge: Cambridge University Press. doi:10.1017/CBO9781107415874, 2014,

[9] Maochao Xu, Gaofeng Da & Shouhuai Xu, *Cyber Epidemic Models with Dependences*, Internet Mathematics, 11:1, DOI: 10.1080/15427951.2014.902407, 2015, 62-92,

[10] `https://www.zurich.com/en/knowledge/topics/digital-data-and-cyber/what-did-we-learn-about-cyber-risks-in-2020` [dostęp: 27.04.2021r.],

[11] `https://www.pzu.pl/dla-firm-i-pracownikow/majatek-firmy-i-oc/majatek/ubezpieczenie-od-ryzyk-cybernetycznych` [dostęp: 25.03.2021r.],

[12] https://colonnade.pl/dla-firm/financial-lines/ubezpieczenie-ryzyk-cybernetycznych [dostęp: 26.04.2021r.],

[13] https://findia.pl/ubezpieczenie-cyber [dostęp: 27.04.2021r.],

[14] https://axaxl.com/insurance/products/cyber-insurance [dostęp: 27.04.2021r.],

[15] https://www.zurich.com/en/products-and-services/protect-your-business/what-we-protect/cyber-risk [dostęp: 28.04.2021r.],

[16] https://www.travelers.com/cyber-insurance [dostęp: 28.04.2021r.],

[17] https://www.esecurityplanet.com/products/cyber-insurance-companies/ [dostęp: 25.04.2021r.],

[18] https://www.insurancebusinessmag.com/us/news/cyber/top-10-cyber-insurance-companies-in-the-us-195463.aspx [dostęp: 25.04.2021r.],

[19] https://www.soa.org/resources/research-reports/2021/14th-annual-survey/ [dostęp: 03.03.2021r.],

[20] https://www.soa.org/resources/research-reports/2021/14th-annual-survey/ [dostęp: 03.03.2021r.],

[21] https://www.fsb.org/wp-content/uploads/P131017-2.pdf [dostęp: 29.04.2021r.],

[22] Wheatley, S., Maillart, T. i Sornette, D. *The extreme risk of personal data breaches and the erosion of privacy.* Eur. Phys. J. B 89, 7 https://doi.org/10.1140/epjb/e2015-60754-4, 2016,

[23] Martin Eling, Nicola Loperfido, *Data breaches: Goodness of fit, pricing, and risk measurement*, Insurance: Mathematics and Economics, Volume 75, 2017, 126-136,

[24] Xu, Maochao; Schweitzer, Kristin; Bateman, Raymond; Xu, Shouhuai, *Modeling and Predicting Cyber Hacking Breaches*, IEEE Transactions on Information Forensics and Security, 10.1109/TIFS.2018.2834227, 2018, 1-1,

[25] https://privacyrights.org/data-breaches [dostęp: 05.05.2021r.],

[26] J. Koronacki, J. Mielniczuk, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwo Naukowo - Techniczne, Warszawa, 2001,

[27] Adam W. Grace, Ian A. Wood, *Approximating the tail of the Anderson–Darling distribution*, Computational Statistics and Data Analysis, Volume 56, Issue 12, 2012, 4301-4311,

[28] C. Forbes, M. Evans, N. Hastings, B. Peacock, *Statistical Distributions* Fourth Edition, John Wiley & Sons, Inc., 2011,

[29] L. Bauwens, P. Giot, J. Grammig, and D. Veredas, *A comparison of financial duration models via density forecasts*, Int. J. Forecasting, vol. 20, no. 4, 2004, 589–609,

[30] Karadağ Atas, Özge & Aktas Altunay, Serpil, *Goodness of fit tests for generalized gamma distribution*, 10.1063/1.4952355, 1738, 2016,

[31] P. R. Hansen and A. Lunde, *A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)?*, J. Appl. Econ., vol. 20, no. 7, 2005, 873–889,

[32] G. Siudem, B. Żogała-Siudem, A. Cena, M. Gągolewski, *Three dimensions of scientific impact*, National Academy of Sciences, 2020, 13896–13900,

[33] Leo Egghe & Ronald Rousseau, *An informetric model for the Hirsch-index*, Scientometrics, 2006, 121-129,

[34] Leo Egghe, *An informetric model for the Hirsch-index*, Scientometrics, 2009, 232-129,

[35] G. G. Naumisa, G. Cocho, *Tail universalities in rank distributions as an algebraic problem: The beta-like function*, Physica A: Statistical Mechanics and its Applications, Volume 387, Issue 1, 2008, 84-96,

[36] Gustavo Martínez-Mekler, Roberto Alvarez Martínez, Manuel Beltrán del Río, Ricardo Mansilla, Pedro Miramontes, Germinal Cocho, *Universality of Rank-Ordering Distributions in the Arts and Sciences*, Public Library of Science, PLOS ONE, 2009, 1-7,

[37] E. Widz, T. Bar, *Autokorelacja stóp zwrotu w badaniu słabej efektywności polskiego rynku kapitałowego*, Annales Universitatis Mariae Curie-Skłodowska. Sectio H, Oeconomia 43, 2009, 223-232,

[38] S. Csorgo, J. J. Faraway, *The Exact and Asymptotic Distributions of Cramer-von Mises Statistics*, Wiley, Journal of the Royal Statistical Society. Series B (Methodological) Vol. 58, No. 1, 1996, 221-234,

[39] Rosie Shier, *Statistics: 1.4 Chi-squared goodness of fit test*, Mathematics Learning Support Centre, 2004, 1-2,

[40] J. Mielniczuk, *Analysis of Time Series : Theory*, Instytut Podstaw Informatyki PAN, Warszawa, 2015, 131-134,

[41] Shapiro, S. S., and M. B. Wilk, *An Analysis of Variance Test for Normality (Complete Samples)*, Biometrika, vol. 52, no. 3/4, 1965, 591–611,

[42] J. Jakubowski, R. Sztencel, *Wstęp do teorii prawdopodobieństwa*, Wydawnictwo SCRIPT, Wydanie 4, 2010, 104-107,

[43] Paweł Dygas, *Zarządzanie ryzykiem w ubezpieczeniach - wykłady semestr zimowy 2020/2021*, Departament ds. Kontroli, Bezpieczeństwa, Ryzyka, Reasekuracji i Antyfraudu UNIQA TU S.A. / UNIQA TU na Życie S.A.

[44] Marek Lusztyn, *Weryfikacja historyczna modeli wartości zagrożonej – zastosowanie wybranych metod dla rynku polskiego w okresie kryzysu finansowego*, Wydawnictwo Uniwersytetu Ekonomicznego we Wroclawiu, Ekonometria, 2013, 117-129,

[45] Alexios Ghalanos, *Introduction to the rugarch package (Version 1.4-3)*, `https://rdrr.io/cran/rugarch/f/inst/doc/Introduction_to_the_rugarch_package.pdf`, 2020

[46] Peter F. Christoffersen, *Evaluating Interval Forecasts*, Wiley, International Economic Review, 39(4), 1998, 841-862,

[47] Piotr Mazur, *Pomiar ryzyka rynkowego za pomocą miary Value at Risk – podejście dwuetapowe*, Autoreferat rozprawy doktorskiej napisanej pod kierunkiem dr hab., prof. UW Ryszarda Kokoszczyńskiego, WNE UW, 2014,

[48] Marek Kwas, Michał Rubaszek, *Skrypt do przedmiotu: Modelowanie Ryzyka Finansowego z R*, Zakład Modelowania Rynków Finansowych, Instytut Ekonometrii SGH, 2020

[49] Eom, Y.H., Fortunato, S., *Characterizing and modeling citation dynamics*, PLOS ONE 6(9): e24926, 2011, 1-7,

[50] Thelwall, M., Wilson, P., *Distributions for cited articles from individual subjects and years*, Journal of Informetrics 8, 2014, 824-839,

[51] Thelwall, M., *Are the discretised lognormal and hooked power law distributions plausible for citation data?*, Journal of Informetrics 10, 2016, 454-470,

[52] Thelwall, M., *The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression*, Journal of Informetrics 10, 2016, 336-346,

[53] Néda, Z., Varga, L., Biró, T.S., *Science and Facebook: The same popularity law!*, PLOS ONE 12(7): e0179656, 2017, 1-11,

[54] Brito, R., Navarro, A.R., *The inconsistency of h-index: A mathematical analysis*, Journal of Informetrics, Elsevier, vol. 15(1), 2021, 1-11,

[55] Kanti Mardia, J. Kent, J. Bibby, *Multivariate Analysis, 1st Edition*, Academic Press, Harcourt Brace & Company, 1979, 521

[56] Hubert W. Lilliefors, *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown*, Journal of the American Statistical Association, 62(318), 1967, 399–402,

[57] Hubert W. Lilliefors, *On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown*, Journal of the American Statistical Association, 64(325), 1969, 387-389,

[58] F.G Parsons, P.H Wirsching, *A Kolmogorov - Smirnov goodness-of-fit test for the two-parameter weibull distribution when the parameters are estimated from the data*, Microelectronics Reliability, Volume 22, Issue 2, 1982, 163-167,

[59] B. Edwards, S. Hofmeyr, S. Forrest, *Hype and heavy tails: A closer look at data breaches*, Journal of Cybersecurity, Volume 2, Issue 1, 2016, Pages 3–14,

[60] Jaworski, P., Durante, F., Härdle, W., & Rychlik, T., *Copula Theory and Its Applications*, Springer; 2010th edition, 2010,

[61] M. Xu, L. Hua, and S. Xu, *A vine copula model for predicting the effectiveness of cyber defense early-warning*, Technometrics, vol. 59, no. 4, 2017, 508–520,

[62] Lilian Cheung, *High Frequency Data: Modeling Durations via the ACD and Log ACD Models*, Honors Scholar Theses. 394, 2014,

[63] `https://rpubs.com/ionaskel/VaR_Garch_market_risk` [dostęp: 16.09.2021r.],

[64] T. Nagler, *kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities*, Journal of Statistical Software, 84(7), 2018, 1-22,

[65] Brechmann, Eike & Schepsmeier, Ulf, *Modeling dependence with C- and D-Vine Copulas: The R package CDVine*, Journal of Statistical Software, 52(3), 2013, 1-27,

[66] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY, 2009, 745

[67] J. Ćwik, J. Mielniczuk, *Statystyczne systemy uczące się: Ćwiczenia w oparciu o pakiet R*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2009,

[68] Ryszard Szekli, *Matematyka ubezpieczeń majątkowych i osobowych*, Skrypt do wykładu, Uniwersytet Wrocławski, 2016

[69] Ryszard Magiera, *Modele i metody statystyki matematycznej*, Oficyna Wydawnicza GiS, 2002

[70] H. Braun, *A simple method for testing goodness-of-fit in the presence of nuisance parameters*, Journal of the Royal Statistical Society, 42, 1980, 53-63,

[71] Peter J. Brockwell, Richard A. Davis, *Introduction to Time Series and Forecasting*, Springer-Verlag New York, 2002, 437,

[72] Wojciech Otto, *Ubezpieczenia majątkowe: Część I Teoria ryzyka*, Wydawnictwo WNT, Warszawa, 2015, 348,

[73] Aldirawi, H., Yang, J., & Metwally, A.A., *Identifying Appropriate Probabilistic Models for Sparse Discrete Omics Data*, IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2019, 1-4,

[74] Gordon, Lawrence & Loeb, Martin & Sohail, Tashfeen, *A framework for using insurance for cyber-risk management*, Communications of the ACM 46(3), 2003, 81-85,

[75] Thomas Nagler, Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Tobias Erhardt, *Statistical Inference of Vine Copulas*, `https://CRAN.R-project.org/package=VineCopula`, 2021,

[76] Christian Genest, Anne-Catherine Favre, *Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask*, Journal of Hydrologic Engineering 12, 2007, 347-368,

[77] J. Jakubowski, *Modelowanie rynków finansowych*, SCRIPT, 2006, 260,

[78] Gallego Escudero, H. F., & Ríos Saavedra, O. A., *Use of the Autoregressive Conditional Duration Model to predict the dollar fall in the colombian Exchange Market*, Revista de Economía del Rosario. Vol. 23. No. 2, 2020, 1-21,,

[79] M. Gągolewski, B. Żogała-Siudem `https://github.com/cenka/Predicting-Bibliometric-Indices-with-Citation-Models`, 2021, [dostęp: 16.11.2021r.]