Nr 4(97) 2024 Bezpieczny Bank

DOI: 10.26354/bb.4.4.97.2024

Julia Kuchno*
ORCID: 0000-0002-8362-6871
juliakuchno@gmail.com

Segmentacja wierzytelności przeterminowanych z sektora ubezpieczeń z wykorzystaniem algorytmów mieszanych danych

Streszczenie

Niniejsze opracowanie podejmuje tematykę segmentacji wierzytelności przeterminowanych z sektora ubezpieczeń, pochodzących z rynku wtórnego. Celem artykułu jest ocena efektywności zastosowania metody Fast K-Prototypes do segmentacji wierzytelności tego typu, z uwzględnieniem wpływu parametrów modelu oraz jakości danych wejściowych na jakość uzyskanych wyników. Artykuł podejmuje także tematykę spłacalności wierzytelności z sektora ubezpieczeń i oceny ryzyka przez nie generowanego. Próba badawcza zawiera 2376 roszczeń regresowych z tytułu ubezpieczeń komunikacyjnych, które były nabywane w latach 2012–2023 przez polski podmiot zajmujący się działalnością windykacyjną. Zastosowanie metody Fast K-Prototypes pozwoliło na podział wierzytelności na różne grupy ryzyka kredytowego, pod warunkiem zastosowania określonych parametrów oraz zachowania wysokiej jakości danych wejściowych poprzez odpowiednie przygotowanie i wstępną analizę. Analiza wykazała wysoki poziom ryzyka tego typu wierzytelności i ich niską spłacalność. Wyniki potwierdzają, że metoda Fast K-Prototypes może być skuteczna, ale jej efektywność zależy od jakości danych i wymaga dalszych badań w kontekście różnorodnych prób badawczych.

Słowa kluczowe: wierzytelności przeterminowane, sektor ubezpieczeń, roszczenia regresowe, ryzyko kredytowe, Fast K-Prototypes

Kod JEL: C38, G22

* Julia Kuchno – doktorantka w Szkole Głównej Handlowej w Warszawie.

Clustering overdue receivables in the insurance sector: a mixed data approach

Abstract

This study addresses the issue of overdue receivables from the secondary market. The main objective of the research is to evaluate the application of the Fast K-Prototypes algorithm to the overdue insurance receivables segmentation, considering how selected parameters and data quality influences obtained results. The article also addresses the repayment of receivables from the insurance sector and the assessment of the risks they generate. The research sample includes 2376 recourse claims which arose from motor insurance and have been acquired between 2012–2023 by a Polish debt collection company.

The application of the Fast K-Prototypes method enabled the segmentation of overdue receivables into various credit risk groups, provided that specific parameters were applied, and the input data was of high quality thanks to preliminary analysis and appropriate preparation. The analysis confirms that these assets are associated with a significant level of credit risk. The results indicate that the application of the Fast K-Prototypes method supports the debt recovery process optimization. However, the effectiveness of this method depends on the research sample and suggests the importance of further research in the context of diverse data samples.

Keywords: overdue receivables, insurance sector, recovery claims, credit risk, Fast K-Prototypes

JEL Codes: C38, G22

Wstęp

Jednym ze sposobów pozyskiwania kapitału zewnętrznego na rynku finansowym jest proces sprzedaży wierzytelności przeterminowanych. Proces ten może być szczególnie korzystny dla instytucji z sektora ubezpieczeń, zapewnia on bowiem poprawę płynności finansowej, a co za tym idzie wspiera zachowanie przez te podmioty funkcji instytucji zaufania publicznego (Śliwiński 2011, s. 467). Dzięki wyodrębnieniu części aktywów ze swojego bilansu podmioty te mają możliwość ograniczenia ryzyka związanego z ich działalnością.

W literaturze podkreśla się, że wierzytelności tego typu mogą być problematyczne z uwagi na skomplikowany charakter świadczenia ubezpieczyciela, różnorodność potencjalnych podstaw prawnych oraz kontrowersje związane z wzajemnością umowy ubezpieczenia (Gruszczyński 2018, s. 45). Podstawą powstania roszczenia wynikającego ze stosunku prawnego będącego ubezpieczeniem mogą być między innymi nieopłacone składki ubezpieczeniowe, roszczenia regresowe lub nienależne świadczenia. Złożoność prawna tych roszczeń powoduje wydłużenie się procesu zaspokojenia wierzyciela oraz wzrost ryzyka kredytowego. Zarządzanie tymi aktywami, rozumiane jako ich wycena oraz obsługa w procesie windykacji, wymaga więc zastosowania odpowiednich metod analitycznych.

Celem artykułu jest zaprezentowanie zastosowania metody Fast K-Prototypes do analizy przeterminowanych wierzytelności pochodzących z sektora ubezpieczeń. Próba badawcza zawiera 2376 przeterminowanych roszczeń regresowych z ubezpieczeń komunikacyjnych i została szerzej opisana w części metodologicznej artykułu. Kluczowym elementem zaprezentowanych badań jest możliwość podziału tych aktywów względem generowanego przez nie ryzyka kredytowego i stopy zwrotu. Metoda Fast K-Prototypes cechuje się wysoką efektywnością i elastycznością w procesie segmentacji danych mieszanych, co w przypadku analizy wierzytelności przeterminowanych ma istotne znaczenie.

Hipoteza badawcza zakłada, że: *efektywność segmentacji wierzytelności przeterminowanych za pomocą metody Fast K-Prototypes zależy od odpowiedniego doboru parametrów modelu (liczby klastrów, wartości lambda) oraz jakości danych wejściowych.*

Artykuł został podzielony na cztery części: przegląd literatury, opis metodologii badania, analiza wyników badania oraz podsumowanie. Przegląd literatury obejmuje prezentację metody Fast K- Prototypes oraz przegląd zastosowania analizy skupień w sektorach ubezpieczeniowym i bankowym. W części metodologicznej przedstawiono próbę badawczą oraz przebieg badań. Następnie omówiono wyniki analiz i ich implikacja dla wykorzystania metody Fast K-Prototypes do segmentacji wierzytelności z sektora ubezpieczeń. Artykuł kończy podsumowanie, w którym podkreślono kluczowe elementy badań oraz ich znaczenie w kontekście analizy wierzytelności przeterminowanych.

Zaprezentowane w artykule wyniki mają walor teoretyczny, uzupełniając dotychczasową literaturę o przedstawienie zastosowania metody Fast K-Prototypes do segmentacji wierzytelności przeterminowanych pochodzących z rynku wtórnego. Ponadto, sformułowane wnioski dostarczają praktycznych rekomendacji, które mogą wspierać proces zarządzania wierzytelnościami przeterminowanymi z sektora ubezpieczeń.

1. Wybrane metody analizy skupień: teoria i zastosowania

Literatura dzieli metody klasteryzacji na hierarchiczne, niehierarchiczne oraz metody rozmytej analizy skupień (Sala 2017, s. 142). Metody hierarchiczne koncentrują się na tworzeniu hierarchii klastrów w formie dendrogramu, umożliwiając analizę zależności między grupami na różnych poziomach szczegółowości (Saxena et al. 2017, s. 666). Z kolei metody niehierarchiczne, takie jak metoda k-średnich, które dzielą zbiór danych na określoną z góry liczbę klas. Przypisanie obserwacji do klas odbywa się na podstawie wartości odległości od centrów klastrów (Sobolewski, Sokołowski 2017, s. 217). Metoda rozmytej analizy skupień polega natomiast na przypisywaniu każdego punktu danych do wszystkich klastrów z różnym stopniem przynależności zamiast jednoznacznego przypisania do jednego klastra (Saxena et al. 2017, s. 667). Z uwagi na analizowaną metodę Fast K- Prototypes, która należy do metod niehierarchicznych, dalsza analiza metodologii analiz baz danych zawęża się do tego obszaru.

Metody niehierarchiczne (*k*-średnich, *k*-modes) uznawane są za wysoce efektywne i łatwe w interpretacji. Zarzuca się im jednak wrażliwość na początkowe rozłożenie danych i lokalizację centroidów klastrów (Sala 2017, s. 143). Algorytm *k*-średnich jest skuteczny dla danych numerycznych, podczas gdy *k*-modes koncentruje się na danych kategorycznych (Huang 1998, s. 301). Połączenie tych metod umożliwia algorytm K-Prototypes, który jest szczególnie użyteczny w analizie danych mieszanych. Zoptymalizowana wersja Fast K-Prototypes oferuje lepszą skalowalność i krótszy czas obliczeń, co czyni go narzędziem wspierającym zarządzanie wierzytelnościami przeterminowanymi (Kim 2017, s. 3).

1.1. Model Fast K-Prototypes

W zaprezentowanych badaniach zastosowano zmodyfikowaną metodę K-Prototypes, zwaną Fast K-Prototypes (Kim 2017). Segmentacja danych pozwala lepiej zrozumieć specyfikę przeterminowanych wierzytelności ubezpieczeniowych, wspierając precyzyjne strategie windykacyjne. Wierzytelności wysokiego ryzyka wymagają intensywnych działań, jak egzekucja komornicza, a niskiego ryzyka – prostszych, np. automatycznych przypomnień. Taki podział może wspierać również proces alokacji zasobów, umożliwiając skoncentrowanie działań na segmentach o największym potencjale odzyskania należności.

Funkcja celu w podstawowym algorytmie K-Prototypes dąży do minimalizacji łącznego poziomu różnic (miary niepodobieństwa) pomiędzy punktami danych a centroidami, które reprezentują środki klastrów w przestrzeni wielowymiarowej. Zastała ona przedstawiona za pomocą równania 1:

Równanie 1. Funkcja kosztu modelu K-Prototypes

$$F(U,Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{i,l} d(x_i, q_l)$$

Źródło: opracowanie własne na podstawie Z. Jia, L. Song (2020, s. 2).

gdzie: $U = [u_{i,l}]$ to macierz przypisania punktów do klastrów, Q reprezentuje zbiór centroidów dla każdego klastra. Z kolei wartość $d(x_i, q_l)$ przedstawia różnicę pomiędzy punktem x_i a centroidem q_l, a sposób oszacowania tej miary został przedstawiony na równaniu 2:

Równanie 2. Miara niepodobieństwa

$$d(x_i, q_l) = \gamma \sum_{s=1}^{p} \delta(x_{i,s}, q_{l,s}) + \sum_{s=p+1}^{m} \sqrt{(x_{i,s}^N - q_{l,s}^N)^2}$$

Źródło: opracowanie własne na podstawie: Z. Jia, L. Song (2020, s. 2).

gdzie: p to liczba zmiennych kategorycznych, $m - p$ to liczba zmiennych ilościowych, a $x_{i,s}$, $q_{l,s}$ – reprezentują wartości zmiennej s w punkcie x_i, w centroidzie q_l. Parametr γ (*gamma*) odgrywa kluczową rolę w algorytmie K-Prototypes, równoważąc wpływ zmiennych kategorycznych i numerycznych na proces klasteryzacji. Wyższe wartości γ zwiększają znaczenie zmiennych kategorycznych, podczas gdy niższe wartości wzmacniają wpływ zmiennych numerycznych. Miara niepodobieństwa analizuje mieszane dane dzieląc różnice na część jakościową (hammingową) oraz ilościową (eukleidesową) (Sroka 2021, s. 49).

Algorytm K-Prototypes inicjuje centroidy jako średnie wartości zmiennych ilościowych i modalne dla zmiennych jakościowych. Do każdego klastra następuje przypisanie punktu uwzględniając minimalizację funkcji kosztu F(U,Q) (Huang 1998, s. 291–292). Następnie aktualizowana jest wartość każdego centroidu. W literaturze wskazuje się, że algorytm K-Prototypes może być zaawansowany obliczeniowo i wrażliwy na początkowe wartości centroidów, co wpływa na stabilność jego wyników (Kim 2017, s. 1).

Z tego względu w niniejszych badaniach wykorzystano algorytm Fast K-Prototypes, który modyfikuje równanie miary odległości poprzez szacowanie odległości cząstkowych (*partial distance*) (Kim 2017, s. 2) i wprowadza parametr γ (*lambda*), będący odpowiednikiem parametru *gamma* z podstawowej formuły algorytmu (równanie 1). Algorytm ten minimalizuje odległość między obiektami a centroidami klastrów, ograniczając konieczność obliczeń odległości dla wszystkich zmiennych (ilościowych i jakościowych). Maksymalna różnica między centroidami w przestrzeni zmiennych ilościowych jest wykorzystywana jako kryterium wykluczenia pewnych obliczeń. Wśród kluczowych założeń wskazuje się: podział danych na zmienne jakościowe i ilościowe, szacowanie odległości cząstkowych, iteracyjną aktualizację centroidów oraz minimalizację funkcji kosztu, czyli różnicy pomiędzy obserwacją a centroidem klastra. Oszacowanie odległości cząstkowej obejmuje wybrane zmienne, co pozwala zredukować zbędne obliczenia w sytuacjach, gdy różnice między centroidami są na tyle duże, że nie wymagają dalszej analizy przypisania punktów do klastrów. Metoda Fast K-Prototypes cechuje się wrażliwością na początkowe warunki, jak rozmieszczenie centroidów, wartość parametru λ czy liczba klastrów. Z tego względu analiza z wykorzystaniem tej metody może być bardziej złożona obliczeniowo. W niniejszych badaniach problem wrażliwości na początkowe warunki rozwiązano poprzez testowanie różnych konfiguracji parametrów, co pozwoliło na uzyskanie stabilnych i wiarygodnych wyników.

1.2. Zastosowanie klasteryzacji w sektorze ubezpieczeń i bankowości

Metody analizy skupień rozwijały się od lat 60., wspierając różne dziedziny, w tym: bankowość, ubezpieczenia oraz marketing. Choć kontekst działania tych sektorów różni się, ich wspólnym mianownikiem jest segmentacja danych, która pozwala na skuteczniejsze zarządzanie klientem oraz ryzykiem przez niego generowanym, usprawnienie procesów operacyjnych oraz poprawę wyników finansowych.

W dziedzinie ubezpieczeń metody klasteryzacji pozwalają dopasować oferowane produkty do specyficznych potrzeb klientów oraz efektywniej zarządzać portfelami ubezpieczeniowymi (Wen, Gao, Xiao 2021, s. 271). Metody te umożliwiają segmentację klientów na podstawie ich cech bez potrzeby posiadania wcześniejszych etykiet klas. Ponadto może być stosowana jako etap wstępny do bardziej zaawansowanych analiz predykcyjnych, zwiększając ich precyzję poprzez dostarczenie jednorodnych segmentów. Ta elastyczność i zdolność do analizy nieliniowych zależności sprawiają, że analiza skupień jest kluczowym narzędziem w optymalizacji strategii zarządzania ryzykiem i personalizacji ofert w sektorze ubezpieczeń (Jamotton, Hainaut, Hames 2024, s. 27–28).

W literaturze dotyczącej *credit scoringu* można także zauważyć wykorzystanie podobnych metod (Jadwal et al. 2019; 2017). Podobnie jak w ubezpieczeniach, gdzie klasteryzacja ogranicza ryzyko nadużyć, tak w *credit scoringu* poprawia ona ocenę klientów banku i identyfikuje ich prawdopodobieństwo braku spłaty (Caruso et al., 2020, s. 5). W literaturze wskazuje się, że możliwości zastosowania klasteryzacji są szerokie i może ona posłużyć ocenie ryzyka kredytowego oraz prognozowaniu spłat (Idbenjra, Coussement, De Caigny 2024 s. 2). Należy jednak zaznaczyć, że w kontekście *credit scoringu* i zarządzania ryzykiem kredytowym, metody te są częściej wykorzystywane w celu osiągnięcia innych celów niż typowe modele predykcyjne (Bijak, Thomas 2012, s. 2434–2435). Metody analizy skupień stanowią najczęściej etap wstępny, przygotowujący bazę danych do modelowania właściwego.

Zarządzanie wierzytelnościami przeterminowanymi wymaga uwzględnienia zarówno specyfiki aktywów, jak i ich heterogeniczności. Klasteryzacja, jako metoda pozwalająca na identyfikację ukrytych wzorców, może okazać się nieoceniona w tym kontekście. W przypadku NPL klasteryzacja umożliwia grupowanie wierzytelności na podstawie podobnych charakterystyk, jak wiek długu czy rodzaj zabezpieczenia (Arutjothi, Senthamarai 2022, s. 88). Dzięki temu możliwe jest dokładniejsze oszacowanie strat (LGD) oraz kalibracja modeli zgodnie z zasadami rachunkowości, np. IFRS 9 (European Central Bank 2017, s. 68). To z kolei pozwala na dostosowanie strategii windykacyjnych, lepsze prognozowanie spłat, a także bardziej precyzyjną wycenę portfeli wierzytelności, co ma kluczowe znaczenie zarówno dla wierzycieli pierwotnych, jak i wtórnych.

Na podstawie przeglądu literatury warto zauważyć, że istnieje luka badawcza w zastosowaniu metod klasteryzacji, jak Fast K-Prototypes do analizy wierzytelności przeterminowanych. Wraz z ciągłym rozwojem metod *machine learning* przedstawiane są

nowe rozwiązania, które pozwalają usprawniać procesy dotyczące segmentacji, również w kontekście wierzytelności przeterminowanych. Szczególnie ciekawe wydaje się zastosowanie metody Fast K-Prototypes w stosunku do wierzytelności z sektora ubezpieczeń, gdzie segmentacja klientów jest popularną praktyką.

2. Metodologia przeprowadzonych badań

2.1. Charakterystyka zbioru badawczego

Segmentacja wierzytelności z roszczeń regresowych z ubezpieczeń komunikacyjnych została przeprowadzona na danych pochodzących od polskiej firmy windykacyjnej. Próba badawcza zawiera 2 376 wierzytelności nabywanych w portfelach w latach 2012–2023. Początkowa wartość nominalna portfeli wyniosła 7 418 378,64 zł. We wskazanym okresie badawczym łącznie spłacono 1 264 319,60 zł. Saldo aktualne na 31 grudnia 2023 r. wierzytelności wyniosło 16 161 073,71 zł.

Zebrano dane obejmujące zarówno informacje dostępne na moment nabycia wierzytelności, jak i wybrane informacje dotyczące procesu obsługi wierzytelności z 31 grudnia 2023 r.

Wyselekcjonowano następujące zmienne ilościowe:

- wartość nominalna w momencie nabycia (*wart_nom_pocz*),
- stosunek kosztów poniesionych do momentu nabycia do początkowej wartości nominalnej (*koszty_wart_pocz*),
- koszty wierzyciela wtórnego podczas dochodzenia wierzytelności (*koszty_wierzyciel*),
- kwota spłacona po nabyciu wierzytelności (*kwota_spłacona*),
- aktualne saldo zadłużenia na 31 grudnia 2023 r. (*saldo_aktualne*),
- wskaźnik stopy zwrotu (*recovery rate*),
- okres windykacji wierzytelności przez wierzyciela wtórnego (okres_portfela),
- stosunek poniesionych kosztów przez wierzyciela wtórnego do wartości kwoty spłaconej (efektywność_koszt),
- liczba spłat komorniczych (spłat_komornik),
- liczba spłat dobrowolnych (spłat_dłużnik),

oraz zmienne kategoryczne:

- płeć,
- typ dłużnika (osoba fizyczna, działalność gospodarcza),
- wiek dłużnika w momencie uzyskania tytułu przez wierzyciela pierwotnego w przedziałach <18;24), <25;34),<35;44) <45;54), <55;64).

Na moment początkowy analizy wierzytelności zostały przeanalizowane pod kątem spłacalności. W tym celu wprowadzono zmienną reprezentującą procent spłaty całkowitej wierzytelności (równanie 3):

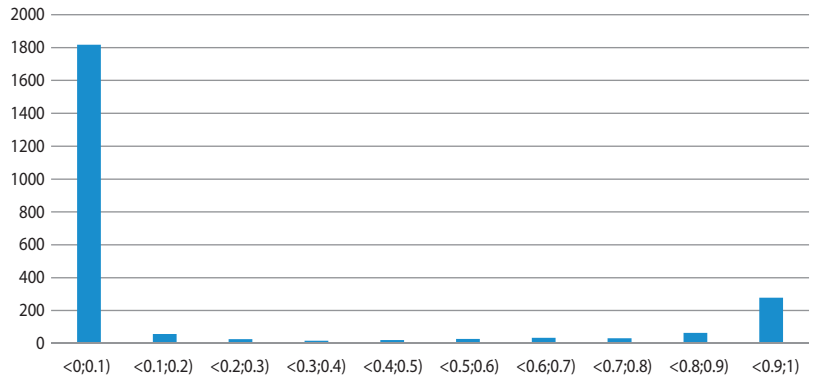
Równanie 3. Zmienna syntetyczna *Recovery Rate*

$$Recovery\ rate = \frac{Kwota\ spłacona}{Początkowe\ zadłużenie + poniesione\ koszty + naliczone\ odsetki}$$

Źródło: W. Starosta (2020, s. 196).

Analiza zmiennej *Recovery rate* (rys. 1) wykazała istotną asymetrię rozkładu, co potwierdza przypuszczenia, że wierzytelności przeterminowane z tytułu ubezpieczeń są trudne w obsłudze i obarczone wysokim ryzykiem kredytowym.

Rysunek 1. Analiza wskaźnika *Recovery rate* w sprawach



Źródło: opracowanie własne na podstawie danych pochodzących z jednego z przedsiębiorstw windykacyjnych działających na polskim rynku.

Literatura definiuje ryzyko kredytowe jako ryzyko braku spłaty zobowiązania przez dłużnika (Hull 2018, s. 52). Z tego względu relację pomiędzy zmienną *Recovery rate* a ryzykiem kredytowym specyficznym dla analizowanych wierzytelności można przedstawić za pomocą równania 4:

Równanie 4. Oszacowanie ryzyka kredytowego specyficznego dla wierzytelności

$$Ryzyko\ kredytowe = 1 - Recovery\ rate$$

Źródło: opracowanie własne na podstawie Hull (2018, s. 52).

Analizując rysunek 1 można zauważyć, że mimo iż znajduje się więcej spraw, w których nastąpiła spłata (1298) w porównaniu do liczby spraw bez jakiejkolwiek spłaty (1078), nadal można zaobserwować wysoki odsetek braku spłaty w dominującej części spraw w przedziale <0;0,1), co wskazuje na istotną liczbę spraw, w których stopa zwrotu była niższa od 10% wartości nominalnej wierzytelności. Taki rozkład

może mieć wpływ na wynik segmentacji, dlatego przewiduje się, że uzyskane klastry nie będą w pełni tych samych rozmiarów.

Próba badawcza została przeanalizowana pod kątem wybranych zmiennych ilościowych. Analiza tabeli 1 wskazuje na niską wartość nominalną w sprawach. Jednakże występują istotne różnice pomiędzy medianą a średnią i wysokim odchyleniem, co sugeruje wysokie odchylenia.

Podobną zależność można zaobserwować dla kosztów poniesionych przed sprzedażą portfela przez wierzyciela pierwotnego. Wartość mediany sugeruje, że stanowią one ok. 30% wartości nominalnej wierzytelności na moment transakcji.

Tabela 1. Podstawowe statystyki dotyczące badanych zmiennych

Nazwa zmiennej	Średnia	Mediana	Odchylenie
Wart_nom_pocz	2 867,26	939,76	7 028,49
Koszty_pocz	1 003,69	322,93	3 797,16
Koszty_wart_pocz	0,36	0,33	0,23
Koszty_wierzyciel	1 828,51	630,56	3 758,48
Kwota_spłacona	962,29	22,95	2 668,95
Liczba_spłat	3,16	1,00	7,92
Saldo_aktualne	8 934,27	1 402,77	25 540,83
Recovery_rate	0,31	0,03	0,41
Okres_portfela	7,53	10,00	4,68
Spłat_komornik	2,23	1,00	5,86
Spłat_dłużnik	0,92	0,00	5,04

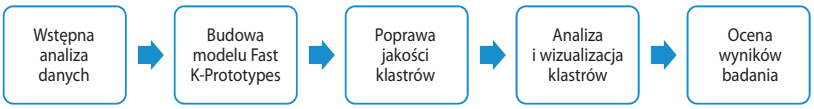
Źródło: opracowanie własne na podstawie danych pochodzących z jednego z przedsiębiorstw windykacyjnych działających na polskim rynku.

Dane finansowe na 31.12.2023 r. wskazują na istotny wzrost wartości nominalnej w wyniku naliczania odsetek i ponoszenia kosztów windykacji. Można także zauważyć wysoką efektywność windykacji przymusowej, o czym świadczy wysoka liczba spłat komorniczych. Mimo wysokiej średniej w próbie danych (31%), wartość mediany dla stopy odzysku wynosi ok. 3%, co oznacza bardzo wysokie prawdopodobieństwo braku spłaty w zbiorze, biorąc pod uwagę, że wartość mediany dla okresu prowadzonego postępowania windykacyjnego w sprawach (*okres_portfela*) wynosi 10 lat. Wstępne wyniki potwierdzają przypuszczenie, że w wyniku klasteryzacji jedna z grup może znacząco dominować liczebnie nad pozostałymi.

2.2. Plan i metodyka badawcza

Niniejsze badania zostały zaprojektowane na podstawie artykułu prezentującego metodę Fast K-Prototypes (Kim 2017) oraz przeglądu literatury dotyczącej niehierarchicznych metod analizy skupień. Badania przeprowadzono w środowisku R oraz arkuszu kalkulacyjnym EXCEL. Rysunek 2 przedstawia przebieg przeprowadzonych czynności w toku badań.

Rysunek 2. Przedstawienie procesu badawczego



Źródło: opracowanie własne na podstawie M. Walesiak (2008, s. 45).

Wstępna analiza danych stanowiła jeden z kluczowych etapów badań, ponieważ pozwoliła zrozumienie specyfiki danych, dzięki czemu zidentyfikowano kluczowe zmienne ilościowe i kategoryczne (jakościowe). W trakcie badania usunięto obserwacje charakteryzujące się brakami w analizowanych danych.

Następnie za pomocą wizualizacji danych oraz analizy głównych składowych (PCA) dokonano wymiarowości bazy danych oraz identyfikacji obserwacji odstających na podstawie odległości Mahalanobisa (Hubert, Debruyne 2010, s. 38). Obserwacje te zostały usunięte. W kolejnym kroku dokonano analizy korelacji (wskaźnik Pearsona) oraz współliniowości zmiennych ilościowych [wskaźnik wariancji inflacji, (Welfe 2018, s. 39,149]. Analiza korelacji została dokonana za pomocą współczynnika Pearsona (Welfe 2018, s. 39). Ponadto według testu ANOVA sprawdzono wpływ zmiennych jakościowych (Wu, Hu, Zheng 2021 s. 5407) na zmienną *Recovery_rate*.

Wstępna analiza danych została zakończona normalizacją danych ilościowych za pomocą metody min max (Walesiak 2008, s. 45). Zmienne kategoryczne zostały przekształcone na odrębne poziomy, co pozwoliło algorytmowi na poprawne obliczenie miary niepodobieństwa między obserwacjami (Kim 2017, s. 4).

W kolejnym etapie przystąpiono do budowy modelu Fast K-Prototypes zgodnie z opisem zaprezentowanym w rozdziale 1.1. W algorytmie Fast K- Prototypes centroidy zmiennych kategorycznych i numerycznych zostały wylosowane, a odległości obliczono dla obu typu zmiennych. Algorytm minimalizował funkcję celu, gdzie parametr lambda (λ) równoważył wpływ zmiennych ilościowych i jakościowych. Obserwacje zostały przypisane do klastrów z najmniejszą odległością od centroidów. Te z kolei były aktualizowane w każdej iteracji. Aby znaleźć optymalny model przetestowano różne konfiguracje parametrów *k* i *lambda*.

Przedział dla parametru *k* ustalono na poziomie <2;6>, aby ograniczyć złożoność obliczeniową (Kaminskyi, Nehrey 2021; Caruso et al. 2020). Z kolei parametr lamb-

da testowany był w przedziale <0.1;2> (Kim 2017, s. 4). Po zakończonej segmentacji dokonano oceny jakości wyników za pomocą wartości kosztu funkcji kosztu oraz wskaźników zaprezentowanych w tabeli 2.

Tabela 2. Prezentacja miar oceny efektywności algorytmu Fast K-Prototypes

Nazwa wskaźnika	Równanie	Opis
Indeks Calinski-Harabasz (CH):	$CH = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}}$	*SSB* – suma kwadratowych odległości między klastrami *SSW* – wariancja wewnątrz klastrów. Wyższe wartości indeksu *CH* wskazują na lepszą jakość procesu klasteryzacji, z uwagi na wysoką separację między nimi (*SSB*) lub niską wariancję wewnątrz klastra (*SSW*).
Mean Silhouette Score	$S(i) = \frac{b(i) - a\,(i)}{\max(a(i), b(i))}$	$a(i)$ – średnia odległość obserwacji *i* do punktów w tym samym klastrze. $b(i)$ – średnia odległość obserwacji *i* do punktów w najbliższym sąsiednim klastrze Wartość wskaźnika mieści się w przedziale <-1;1>. Im wyższe od zera są wartości wskaźnika, tym punkt jest lepiej przypisany do klastra.
Dunn indeks	$D = \frac{min_{i \neq j} d(C_i, C_j)}{max_k \Delta(C_k)}$	$d(C_i, C_j)$ – odległość między klastrami *i* i *j*, $\Delta(C_k)$ – maksymalna odległość między dwoma punktami w klastrze. Im wyższa jest wartość tego wskaźnika, tym lepsza jest jakość dokonanej segmentacji.

Źródło: opracowanie własne na podstawie M. Walesiak (2008, s. 8).

Badania literaturowe wskazują, że rozłożenie obserwacji w grupach może być nierównomierne w sektorze bankowości lub ubezpieczeń (Jadwal et al. 2022; Kaminskyi, Nehrey 2021; Abolmakarem, Abdi, Khalili-Damghani 2016). Biorąc pod uwagę specyfikę wierzytelności przeterminowanych pochodzących z sektora ubezpieczeń, liczbę obserwacji w próbie, a także strukturę danych przedstawioną w punkcie 2.1, zdecydowano, iż segmentacja zostanie uznana za efektywną, jeżeli średnia wartość wskaźnika *Mean sihouette score* będzie wyższa od zera, a klastry będą się różnić między sobą pod kątem średnich poziomów ryzyka oraz stopy zwrotu.

3. Analiza wyników badania

W tej części artykułu zaprezentowano wyniki badań nad oceną efektywności zastosowania algorytmu Fast K-Prototypes do segmentacji wierzytelności pochodzących z sektora ubezpieczeń. Szczególną uwagę poświęcono wpływie doboru parametrów algorytmu na uzyskane wyniki segmentacji. W dalszej części przedstawiono segmentację próby badawczej dla najlepszego scenariusza parametrów oraz zaprezentowano kluczowe obserwacje z zastosowania tej metody do segmentacji wierzytelności przeterminowanych.

3.1. Opis przebiegu badania i analiza jakości segmentacji

W trakcie wstępnej analizy danych zidentyfikowano i usunięto obserwacje odstające (41). Eliminacja obserwacji pozwoliła na poprawę jakości dalszych analiz i wyników klasteryzacji.

W kolejnym kroku przeprowadzono analizę korelacji zmiennych kategorycznych, obliczono wskaźnik VIF oraz dokonano testu ANOVA. Na podstawie wstępnej analizy usunięto ze zbioru zmienne *okres_portfela* oraz *typ_dłużnika*. Analiza korelacji zmiennych wykazała silne korelacje pomiędzy zmienną *recovery_rate* a pozostałymi zmiennymi związanymi z procesem zaspokojenia roszczeń (*kwota spłacona*, *spłat_komornik* oraz *spłat_dłużnik*). Pozostałe zmienne miały wyniki istotne statystycznie.

Po zakończeniu doboru zmiennych przystąpiono do uruchomienia algorytmu. Za pomocą algorytmu Fast K-Prototypes oszacowano wyniki zaprezentowane w tabeli 3, w której pokazano wartości wskaźników jakości segmentacji dla różnych konfiguracji liczby klastrów (*k*) i parametru *lambda*.

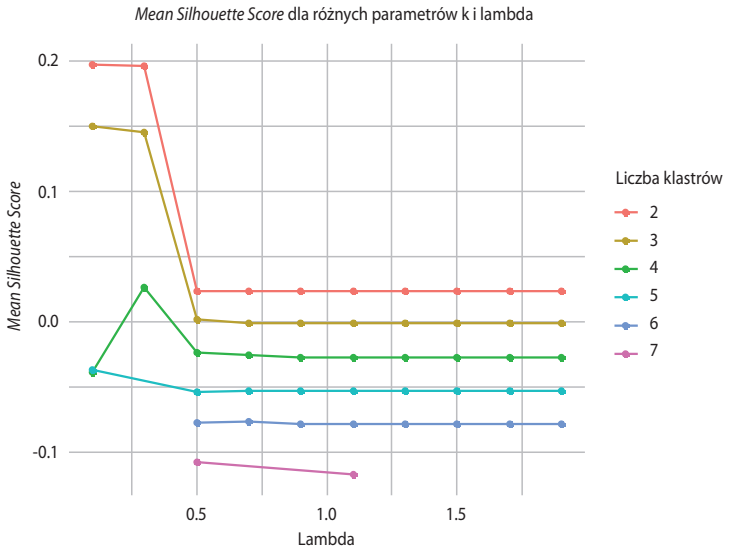
Tabela 3. Analiza jakości segmentacji metodą Fast K-Prototypes

k	lambda	Mean Silhouette	Dunn Index	Funkcja kosztu	CH Index
2	0,1	0,198	0,00000403	1294,22	15,91
2	0,3	0,196	0,00000403	1505,51	16,21
3	0,1	0,151	0,00000403	1291,62	10,24
3	0,5	0,002	0,00000403	1559,15	6,83
4	0,3	0,026	0,00000403	1458,81	10,71
5	0,7	-0,053	0,00000403	1545,42	6,75
6	0,5	-0,077	0,00000403	1442,03	5,98

Źródło: opracowanie własne na podstawie badań własnych.

Analiza prezentowanych miar efektywności klasteryzacji wskazuje, że najlepszy możliwy podział został dokonany dla dwóch klastrów i parametru *lambdy* na poziomie 0.1, co zostało potwierdzone przez najwyższą wartość współczynnika Mean Silhouette i niską wartość funkcji kosztu. Wniosek ten potwierdza także analiza rysunku 3 prezentującego wartości wskaźnika *Mean Silhouette Score* dla różnych konfiguracji liczby klastrów (*k*) oraz parametru *lambda*.

Rysunek 3. Analiza wartości wskaźnika *Mean Silhouette Score*

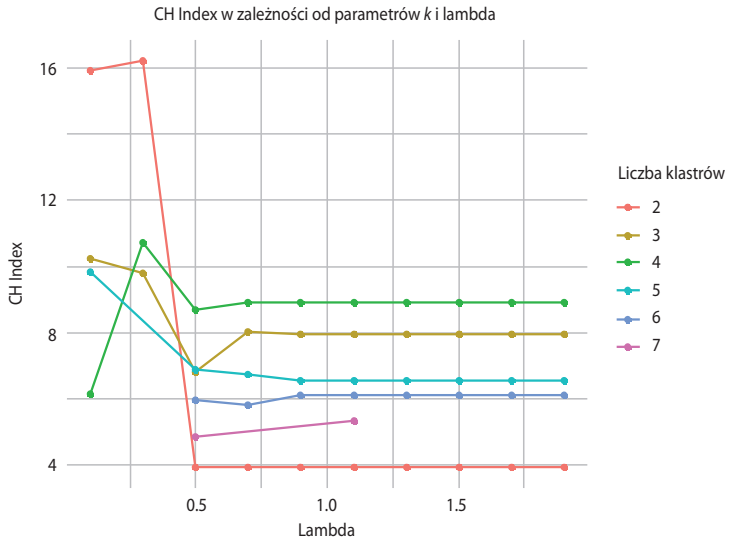


Źródło: opracowanie własne na podstawie badań własnych w środowisku R.

Dla większej ilości klastrów (np. 5 lub 6) wyniki wskaźnika są ujemne i bliskie zeru, co sugeruje, że granice klastrów mogą na siebie nachodzić. Jest to zgodne z obserwacjami z tabeli, gdzie dla tych konfiguracji wartości wskaźnika *Mean Silhouette* były znacznie niższe.

Analizując tabelę 3 widzimy, że optymalny indeks CH został osiągnięty dla konfiguracji 4 klastrów i parametru lambdy równej 0.3. Na podstawie tej wartości można stwierdzić, że największa wewnętrzna spójność oraz separacja pomiędzy klastrami została osiągnięta dla tego zestawu parametrów. Można także zauważyć, że najwyższą wartość CH uzyskano dla scenariusza, w którym występują dwa klastry, a lambda plasuje się na poziomie 0.1. Jednak wraz ze wzrostem parametru lambda wartość indeksu spada, co wskazuje na pogorszenie się jakości klasteryzacji (por. rys. 4).

Rysunek 4. Analiza wskaźnika CH Index



Źródło: opracowanie własne na podstawie badań własnych.

Wyższa liczba zmiennych ilościowych w próbie sprzyja niskim wartościom parametru lambda, co zmniejsza ich wpływ na funkcję kosztu. Wartości lambda istotnie wpływają także na stabilność wskaźnika Dunn'a. Wyniki analizy wskazują, że przybiera on takie same wartości dla wszystkich zestawów parametrów, co również może sugerować niski poziom separacji pomiędzy klastrami. Wyniki analizy indeksów jakości segmentacji metodą Fast K- Prototypes wskazują, że model przypisuje większą wagę zróżnicowanym zmiennym, jednakże może być konieczna w tym zakresie dalsza eliminacja obserwacji odstających.

3.1. Analiza segmentacji dla optymalnego scenariusza

W dalszej części badania wygenerowano wyniki dla scenariusza parametrów *k* oraz lambda, który charakteryzował się najlepszymi wynikami wskaźników (2;0.1). Analiza została przeprowadzona dla wybranych zmiennych względem *Recovery rate*: *saldo_aktualne* oraz *efektywność_koszt*. Analizę klasteryzacji rozpoczęto od oceny średnich wartości zmiennej *Recovery rate* oraz odpowiadającym im średnich wartości ryzyka kredytowego.

Zaprezentowane wartości wskaźników Mean Silhouette w tabeli 4 wskazują, że pierwszy klaster charakteryzował się dużo niższą efektywnością segmentacji w porównaniu do drugiego klastra. Może to sugerować większą heterogoniczność obserwacji lub sygnalizować obecność obserwacji odstających w tej grupie. Tak jak przypuszczano na początku badania klastry różnią się między sobą w stosunku pod kątem liczby przypisanych obserwacji.

Tabela 4. Analiza wyników segmentacji metodą Fast K-Prototypes

Numer klastra	1 (w %)	2 (w %)
Mean Sillhouette	-4	22
Liczba obserwacji	908	1427
Średnia wartość stopy odzysku	21,44	16,69
Mediana wartości stopy odzysku	0,00	0,00
Odchylenie standardowe	37,05	33,46
Minimalna stopa odzysku	0	0
Maksymalna stopa odzysku	100	100
Średnie ryzyko kredytowe	78,56	83,31

Źródło: opracowanie własne na podstawie badań własnych.

Zaobserwowana średnia wartość odzysku w klastrze pierwszym osiągnęła wyższy próg w porównaniu do wartości oszacowanych dla klastra drugiego. Jednakże należy wskazać, że wartość mediany dla obu klastrów plasuje się na poziomie 0%, co oznacza, iż większość obserwacji w obu zbiorach jest obarczona podobnym ryzykiem kredytowym. Można także zaobserwować wyższy poziom zmienności stopy zwrotu i ryzyka kredytowego w klastrze pierwszym (niski poziom Mean Sillhouette, wysokie odchylenie standardowe). Wstępna analiza dokonanej segmentacji sugeruje, że klaster pierwszy obejmuje obserwacje o wyższym potencjale spłaty, ale bardziej zróżnicowanych cechach w porównaniu do klastra drugiego. Z kolei klaster drugi charakteryzuje się większą jednorodnością i stabilnością, jednak średnie ryzyko kredytowe oszacowane dla tej grupy jest odpowiednio wyższe.

W celu sformułowania wniosków dla efektywnego zarządzania wierzytelnościami, analiza została uzupełniona o zbadanie efektywności poniesionych kosztów w sprawach rozumianych jako stosunek poniesionych przez wierzyciela wtórnego kosztów do kwoty spłaconej (tabela 5).

W przypadku efektywności kosztowej tendencja osiąganego przez nie ryzyka kredytowego w sprawach jest podobna. Wraz ze wzrostem wskaźnika spada ryzyko kredytowe. Wzrost wskaźnika efektywności kosztowej oznacza, z jednej strony, że w sprawie nastąpiła spłata, a także iż poniesione przez wierzyciela wtórnego koszty stanowiły pewien odsetek tej spłaty. W klastrze pierwszym o wyższym potencjale spłaty wyższe poniesione koszty powodują generowanie przez te sprawy niższego średniego ryzyka kredytowego. Z kolei w klastrze drugim, mimo podobnej tendencji wzrostowej stopy odzysku wraz ze wzrostem kosztów, rentowność tych spraw jest stosunkowo niższa.

Tabela 5. Analiza wskaźnika efektywności kosztów

	Klaster 1			Klaster 2		
Saldo aktualne	**L**	**RR (w %)**	**RK (w %)**	**L**	**RR (w %)**	**RK (w %)**
0	478	19	81	756	15	85
(0; 0,2>	214	21	79	321	13	87
>0,2	216	27	73	348	23	77
Suma końcowa	908			1 425		

Źródło: opracowanie własne na podstawie badań własnych.

W kontekście zarządzania wierzytelnościami przeterminowanymi wyniki z tabeli 4 i 5 mają przełożenie na wnioski dotyczące zarządzania portfelem wierzytelności przeterminowanych z sektora ubezpieczeń. Skoro proces windykacji dąży do maksymalizacji stopy odzysku, można stwierdzić, że alokowane zasoby w wierzytelności z grupy pierwszej mają wyższą efektywność w stosunku do grupy drugiej. To oznacza, że dla tych spraw należałoby wdrożyć intensywne działania windykacji przymusowej. Z kolei wskaźnik stopy zwrotu w klastrze drugim przyjmuje stosunkowo wyższą wartość dla spraw, w których nie zostały poniesione koszty. W takiej sytuacji należałoby rozważyć wdrożenie czynności windykacji polubownej w tych sprawach.

Podsumowanie

Wyniki przeprowadzonych badań potwierdzają tezę, że: *efektywność segmentacji wierzytelności przeterminowanych za pomocą metody Fast K-Prototypes zależy od odpowiedniego doboru parametrów oraz jakości danych wejściowych.* Najwyższe wskaźniki jakości segmentacji zostały oszacowane dla dwóch klastrów i niskiej wartości lambda. Oznacza to, że metoda Fast K-Prototypes wymaga precyzyjnego dopasowania do charakterystyki zbioru danych. Ponadto, zrozumienie specyfiki danych poprzez analizę głównych składowych (PCA) oraz analizę współliniowości i korelacji pomiędzy zmiennymi pozwala na poprawę jakości tej metody analizy skupień. Takie wnioski podkreślają, że skuteczne zastosowanie metody wymaga starannego przygotowania danych i optymalizacji parametrów, a także testowania różnych scenariuszy parametrów dla zbioru danych.

Jednocześnie zastosowanie metody Fast K-Prototypes do segmentacji wierzytelności przeterminowanych z sektora ubezpieczeń pozwoliło na podział badanego zbioru danych na dwie grupy o zróżnicowanej średniej stopie zwrotu i odmiennym średnim wskaźniku ryzyka kredytowego. Oznacza to, że mimo niskich wskaźników

jakości separacji klastrów i w miarę jednorodnego zbioru danych (duży odsetek spraw niespłaconych, ta sama podstawa powstania roszczenia) metoda Fast K-Prototypes podzieliła zbiór analizowanych danych zgodnie z początkowymi oczekiwaniami, na grupę o niższym potencjale stopy odzysku i wyższym poziomie ryzyka kredytowego oraz grupę obarczoną niższym poziomem ryzyka kredytowego oraz charakteryzującą się wyższą rentownością. Taki podział na klastry o różnej rentowności znajduje także potwierdzenie w literaturze (Caruso et al. 2020, s. 5). Na podstawie segmentacji możliwe było sformułowanie wniosków o efektywności procesu windykacji w tych grupach. Wydzielona w trakcie badań grupa o wyższym potencjale stopy zwrotu i niższym wskaźniku ryzyka kredytowego wydaje się być bardziej rentowna w kontekście windykacji sądowo-egzekucyjnej. Z kolei w grupie o wyższym ryzyku kredytowym zaobserwowano niższą efektywność kosztową, co oznacza, że nakłady na windykację mogą nie przynosić proporcjonalnych korzyści. Na podstawie przeprowadzonych badań i analiz można wysnuć ogólny wniosek, że wierzytelności regresowe z sektora ubezpieczeń charakteryzują się wysokim ryzykiem kredytowym. Specyfika wierzytelności przeterminowanych z sektora ubezpieczeń wiąże się z wysoką złożonością prawną spraw, co przekłada się na wydłużenie procesu windykacji oraz niższą stopę zwrotu. To potwierdza także niski odsetek spraw zamkniętych (spłaconych) w obu grupach sugeruje trudności w szybkim zaspokojeniu wierzyciela. Na tej podstawie można wysnuć ogólne spostrzeżenie, że dla wierzyciela pierwotnego sprzedaż takich roszczeń jest efektywna z punktu widzenia zarządzania aktywami i posiadanym ryzykiem, ze względu na generowanie niższych kosztów.

Bibliografia

Wydawnictwa zwarte

Hull J.C., *Risk Management and Financial Institutions*, John Wiley & Sons, Inc., New Jersey 2018.

Śliwiński A., *Ryzyko ubezpieczyciela i windykacja ubezpieczeniowa*, [w:] K. Kreczmańska-Gigol (red.), *Windykacja należności, ujęcie interdyscyplinarne*, Difin, Warszawa 2011.

Welfe A., *Ekonometria*, Polskie Wydawnictwo Ekonomiczne S.A., Warszawa 2018.

Artykuły prasowe i okolicznościowe

Abolmakarem S., Abdi F., Khalili-Damghani K., *Insurance customer segmentation using clustering approach*, „International Journal of Knowledge Engineering and Data Mining", vol. 4, no. 1, 2016.

Arutjothi G., Senthamarai C., *Assessment of Probability Defaults Using K-Means Based Multinomial Logistic Regression*, „International Journal of Computer Theory and Engineering", vol. 14, no. 2, 2022.

Bijak K., Thomas L.C., *Does segmentation always improve model performance in credit scoring?*, „Expert Systems with Applications", vol. 39, no. 3, 2012.

Caruso G., Gattone S.A., Fortuna F., Di Battista T., *Cluster analysis for mixed data: An application to credit risk evaluation*, "Socio-Economic Planning Sciences", vol. 73, no. 100850, 2020.

Gruszczyński A., *Przeniesienie wierzytelności z umowy ubezpieczenia majątkowego*, „Wiadomości Ubezpieczeniowe", nr 2, 2018.

Hubert M., Debruyne M., *Minimum covariance determinant*, „Wiley Interdisciplinary Reviews: Computational Statistics", vol. 2, 2010.

Huang Z., *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, „Data Mining and Knowledge Discovery", vol. 2, 1998.

Idbenjra K., Coussement K., De Caigny A., *Investigating the beneficial impact of segmentation-based modelling for credit scoring*, „Decision Support Systems", vol. 179, no. 114170, 2024.

Jadwal P.K., Jain S., Gupta U., Khanna P., *K-Means clustering with neural networks for ATM cash repository prediction*, [w:] S. Satapathy, A. Joshi (eds), *International Conference on Information and Communication Technology for Intelligent Systems*, Springer, Cham 2017.

Jadwal P.K., Jain S., Gupta U., Khanna P., *Clustered support vector machine for ATM cash repository prediction*, [w:] B. Pati, C. Panigrahi, S. Misra, A. Pujari, S. Bakshi (eds), *Progress in Advanced Computing and Intelligent Engineering*, Springer, Singapore 2019.

Jadwal P.K., Pathak S., Jain S., *Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis*, „Microsystem Technologies", vol. 28, 2022.

Jamotton C., Hainaut D., Hames T., *Insurance Analytics with Clustering Techniques*, „Risks", vol. 12, no. 9, doi: 10.3390/risks12090141, 2024.

Kaminskyi A., Nehrey M., *Clustering approach to analysis of the credit risk and profitability for nonbank lenders*, CEUR Workshop Proceedings, Machine Learning Methods and Models, Predictive Analytics and Applications – 13th Workshop on the International Scientific Practical Conference Modern Problems of Social and Economic Systems Modelling, MPSESM-W, 2021, https://ceur-ws.org/Vol-2927/paper10.pdf (dostęp 5.12.2024).

Kim B., *A Fast K-prototypes Algorithm Using Partial Distance Computation*, „Symmetry", vol. 9, no. 58, 2017.

Sala K., *Przegląd technik grupowania danych i obszarów zastosowań*, „Społeczeństwo i Edukacja", vol. 25, nr 2, 2017.

Saxena A., Prasad M., Gupta A., Bharill N., Patel O.P., Tiwari A., Joo E.M., Weiping D., Lin C.T., *A review of clustering techniques and developments*, „Neurocomputing", vol. 267, 2017.

Sobolewski M., Sokołowski A., *Grupowanie metodą k-średnich z warunkiem spójności*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu" 2017, nr 468.

Sroka Ł., *Wykorzystanie algorytmu k-prototypów w segmentacji klientów przedsiębiorstwa w marketingu wielopoziomowym*, „Wiadomości Statystyczne. The Polish Statistician" 2021, vol. 66, nr 7.

Starosta W., *Modelling Recovery Rate for incomplete defaults using time-varying predictors*, „Central European Journal of Economic Modelling and Econometrics", no. 12, 2020.

Walesiak M., *Procedura analizy skupień z wykorzystaniem programu komputerowego ClusterSim i środowiska R*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 7 (1207), 2008.

Wen Ch., Gao K., Xiao Y., *Data-Driven Market Segmentation in Insurance Industry and Other Related Sectors,* „Journal of Finance and Accounting", vol. 9, no. 6, 2021.

Wu S., Hu X., Zheng W. et al., *Effects of reservoir water level fluctuations and rainfall on a landslide by two-way ANOVA and K-means clustering*, „Bulletin of Engineering Geology and the Environment", vol. 80, 2021.

Zhang X., Yu L., *Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods*, „Expert Systems with Applications", vol. 237 (A), 2024.

Zhou L., Zhang N., *Customer Segmentation and Optimal Insurance Compensation Ratio: Decision-making Analysis in Financial Institutions*, „International Journal of Multimedia and Ubiquitous Engineering", vol. 10, no. 8, 2015.

Jia Z., Song L., W*eighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient*, „Mathematical Problems in Engineering", 2020.

Materiały internetowe

European Central Bank, *Guidance to banks on non-performing loans*, 2017, https://www.bankingsupervision.europa.eu/ecb/pub/pdf/guidance_on_npl.en.pdf (dostęp 5.12.2024).