

Warsaw University of Life Sciences – SGGW  
Institute of Economics and Finance  
Department of Econometrics and Statistics

**QUANTITATIVE METHODS  
IN ECONOMICS**

**METODY ILOŚCIOWE W BADANIACH  
EKONOMICZNYCH**

**Volume XXIV, No. 4**

Warsaw 2023

## **EDITORIAL BOARD**

Editor-in-Chief: Bolesław Borkowski

Vice-Editor-in-Chief: Hanna Dudek

Managing Editors: Michał Gostkowski, Grzegorz Koszela

Theme Editors:

Econometrics: Bolesław Borkowski

Multidimensional Data Analysis: Wiesław Szczesny

Mathematical Economy: Zbigniew Binderman

Analysis of Labour Market: Joanna Landmesser-Rusek

Financial Engineering: Monika Krawiec

Data Science: Michał Gostkowski

Theory of Risk: Marek Andrzej Kociński

Statistical Editor: Wojciech Zieliński

Technical Editors: Jolanta Kotlarska, Elżbieta Saganowska

Language Editor: Agata Cienkusz

Native Speaker: Yochanan Shachmurove

Editorial Assistant: Łukasz Pietrych

## **SCIENTIFIC BOARD**

Adnene Ajimi (University of Sousse, Tunisia)

Heni Boubaker (University of Sousse, Tunisia)

Peter Friedrich (University of Tartu, Estonia)

Paolo Gajo (University of Florence, Italy)

Agnieszka Gehringer (University of Göttingen, Germany)

Anna Maria Gil-Lafuente (University of Barcelona, Spain)

Jaime Gil-Lafuente (University of Barcelona, Spain)

Vasile Glavan (Moldova State University, Moldova)

Francesca Greselin (University of Milano-Bicocca, Italy)

Ana Kapaj (Agriculture University of Tirana, Albania)

Jirawan Kitchaicharoen (Chiang Mai University, Thailand)

Yuriy Kondratenko (Black Sea State University, Ukraine)

Vassilis Kostoglou (Alexander Technological Educational Institute of Thessaloniki, Greece)

Karol Kukula (University of Agriculture in Krakow, Poland)

Kesra Nermend (University of Szczecin, Poland)

Nikolas N. Olenev (Russian Academy of Sciences, Russia)

Alexander N. Prokopenya (Brest State Technical University, Belarus)

Yochanan Shachmurove (The City College of The City University of New York, USA)

Mirbulat B. Sikhov (al-Farabi Kazakh National University, Kazakhstan)

Marina Z. Solesvik (Nord University, Norway)

Ewa Syczewska (Warsaw School of Economics, Poland)

Achille Vernizzi (University of Milan, Italy)

Andrzej Wiatrak (University of Warsaw, Poland)

Dorota Witkowska (University of Lodz, Poland)

ISSN 2082-792X

e-ISSN 2543-8565

© Copyright by Department of Econometrics and Statistics WULS – SGGW  
(Katedra Ekonometrii i Statystyki SGGW w Warszawie)

Warsaw 2023, Volume XXIV, No. 4

Journal homepage: <https://qme.sggw.edu.pl>

Published by Warsaw University of Life Sciences Press

## CONTENTS

Beata Bieszk-Stolorz, Joanna Landmesser-Rusek, Iwona Markowicz – Zmiany wzorców wychodzenia z bezrobocia na przykładzie Szczecina .....	215
Urszula Grzybowska, Marek Karwański – Selekcja zmiennych metodami statystycznymi i uczenia maszynowego. Porównanie podejść na przykładzie danych finansowych .....	229
Ryszard Szupiluk – System uczenia głębokiego dla eliminacji szumów z wykorzystaniem dywergencji alpha .....	242
Dorota Witkowska, Krzysztof Kompa – Czy zmiana standingu finansowego spółek publicznych wpływa na zmiany odsetka kobiet w organach statutowych? .....	254
Robert Woźniak – Statystyczna weryfikacja „DLC” TEKKEN 7 .....	265
Katarzyna Boratyńska, Katsiaryna Yafremava – Wykorzystanie analizy dyskryminacyjnej do oceny ryzyka bankructwa przedsiębiorstw branży motoryzacyjnej w okresie pandemii Covid-19 .....	275

## ZMIANY WZORCÓW WYCHODZENIA Z BEZROBOCIA NA PRZYKŁADZIE SZCZECINA<sup>1</sup>

**Beata Bieszk-Stolorz**  <https://orcid.org/0000-0001-8086-9037>

Instytut Ekonomii i Finansów  
Uniwersytet Szczeciński  
e-mail: beata.bieszk-stolorz@usz.edu.pl

**Joanna Landmesser-Rusek**  <https://orcid.org/0000-0001-7286-8536>

Instytut Ekonomii i Finansów  
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie  
e-mail: joanna\_landmesser@sggw.pl

**Iwona Markowicz**  <https://orcid.org/0000-0003-1119-0789>

Instytut Ekonomii i Finansów  
Uniwersytet Szczeciński  
e-mail: iwona.markowicz@usz.edu.pl

**Streszczenie:** W artykule podjęto próbę oceny wpływu wybranych cech demograficznych oraz społeczno-ekonomicznych na czas trwania bezrobocia rejestrowanego w Szczecinie w latach 2011-2023. W badaniu wykorzystano metody analizy przeżycia, w tym model przyspieszonej porażki. Analiza danych w kolejnych latach posłużyła odpowiedzi na pytanie: czy wzorce wychodzenia z bezrobocia do pracy ulegały zmianom w czasie. Cel ten osiągnięto poprzez weryfikację hipotez badawczych. Założono, że przyjęte w badaniu zmienne mają wpływ na czas trwania bezrobocia oraz że wpływ ten nie był jednakowy w całym okresie badawczym.

**Słowa kluczowe:** bezrobocie, analiza przeżycia, model AFT

**JEL classification:** J64, C41

---

<sup>1</sup> Badania przeprowadzono w ramach stażu naukowego (2023) *Wykorzystanie metod analizy trwania w badaniach zjawisk społeczno-ekonomicznych na rynku pracy* – współpraca SGGW i US.

## WSTĘP

Bezrobocie jest istotnym problemem społecznym i ekonomicznym. Dotyczy całego społeczeństwa i poszczególnych osób. Bezrobocie ma negatywny wpływ na jakość życia osób, które chcą podjąć pracę. Badania skutków bezrobocia mają charakter interdyscyplinarny i dotyczą różnych aspektów. Z jednej strony analizowane są zmiany w sytuacji jednostki jako skutek bezrobocia, takie jak zdrowie fizyczne i psychiczne, sytuacja finansowa, relacje w rodzinie, sytuacja dzieci [Karwacki, Błędowski 2020]. Z drugiej strony badania dotyczą wpływu bezrobocia na kapitał ludzki i społeczny [Nichols, Mitchell, Lindner 2013], w szczególności w okresie pandemii Covid-19 [Petcu i Cismasu 2023]. Podjęty w artykule temat ma duże znaczenie, zarówno na poziomie indywidualnym jak i społecznym.

W literaturze przedmiotu coraz częściej można znaleźć prezentację wyników badań dotyczących bezrobocia z wykorzystaniem metod analizy trwania. Cooper, McCausland i Theodossiou [2006] wykazali negatywny wpływ bezrobocia na czas pozostawania w dobrym zdrowiu, stosując model przyspieszonej porażki (Accelerated Failure Time, AFT). Autorzy na podstawie metody Kaplana-Meiera wykazali też różnice w przebiegu funkcji przeżycia w dobrym zdrowiu osób bezrobotnych w poszczególnych krajach (13 krajów UE w latach 1994-2001). Dogan [2019] wykorzystał model przyspieszonej porażki do badania wpływu zasiłków dla bezrobotnych na czas trwania w bezrobociu we Francji i Polsce. Dla obu krajów wykazano wydłużenie czasu trwania bezrobocia jako skutek otrzymywania zasiłku. Inne badanie [Dogan 2020], także z zastosowaniem modelu AFT, wykazało że prawidłowość zaobserwowana we Francji i Polsce dotyczy także Hiszpanii i Turcji, a na rynku holenderskim obserwowana jest zależność odwrotna, otrzymywanie zasiłków skraca czas trwania bezrobocia. Innym przykładem zastosowania modelu AFT w analizie bezrobocia jest praca [Grzenda 2019] dotycząca czasu trwania w bezrobociu długotrwałym w Polsce. Autorka wykazała, że dłuższy czas wychodzenia z bezrobocia dotyczy kobiet i osób opiekujących się dziećmi. Natomiast w pracy [Grzenda 2023] wskazano na coraz większe trudności z powrotem na rynek pracy wśród osób starszych. Modele hazardu, a w tym również model AFT, stosuje Landmesser [2009] w analizie czasu trwania w bezrobociu ludności w miastach i na wsi, z uwzględnieniem innych cech osób bezrobotnych a także do oceny wpływu szkolenia zawodowego na czas trwania bezrobocia [Landmesser 2011]. Mała i Cabla [2022] w badaniach czasu trwania w bezrobociu również stosują modele hazardu. Wykazali oni, że w okresie kryzysowym na czeski rynek pracy szybciej wchodził krótkotrwale bezrobotni. Cabla [2012] przy zastosowaniu modelu AFT wykazał natomiast istnienie związku między czasem trwania bezrobocia a płcią, stanem cywilnym, wiekiem i statusem niepełnosprawności. Metody analizy przeżycia wykorzystał w swoich badania Nagler [2013] i stwierdził, że wprowadzenie ubezpieczenia na wypadek bezrobocia

w Chile doprowadziło do skrócenia czasu trwania zatrudnienia. W pracy Bieszk-Stolorz i Markowicz [2022] stwierdzono szybsze wychodzenie z bezrobocia w okresie kryzysu zdrowotnego związanego z pandemią Covid-19.

Przedmiotem badań jest weryfikacja wpływu czynników na czas trwania bezrobocia w Szczecinie. Analiza dotyczy lat 2011-2021. Wykorzystano metod analizy przeżycia, w tym model przyspieszonej porażki, aby wykazać wpływ wybranych czynników (wiek, płeć, wykształcenie i staż pracy badanych bezrobotnych) na czas trwania bezrobocia. Analiza danych w kolejnych latach posłużyła odpowiedzi na pytanie: czy wzorce wychodzenia z bezrobocia ulegały zmianom w czasie. Cel ten osiągnięto poprzez weryfikację postawionych dwóch hipotez badawczych:

H1: Płeć, wiek, wykształcenie i staż pracy osoby bezrobotnej miały wpływ na czas trwania bezrobocia

H2: Wpływ płci, wieku, wykształcenia i stażu pracy osoby bezrobotnej na czas trwania bezrobocia nie był jednakowy w całym okresie badawczym.

## METODYKA BADANIA

W badaniu wykorzystano metody analizy trwania. Metody te są coraz częściej stosowane w naukach ekonomicznych. Ich zaletą jest fakt, że dane mogą mieć charakter kompletnych bądź niekompletnych. Jedną z przyczyn niekompletności danych jest zjawisko cenzurowania jednostek, które wiąże się z ich eliminacją z pola obserwacji. Podczas śledzenia zbioru obiektów mogą pojawić się jednostki, którym będzie towarzyszyć zdarzenie kończące ich obserwację. Mogą też pojawić się jednostki, w przypadku których takie zdarzenie nie wystąpi do zakończenia procesu obserwacji. W tej drugiej sytuacji jednostki nazywamy cenzurowanymi. Cenzurowanie może się odbywać ze względu na czas lub ze względu na zakończenie badania w określonym terminie. Wśród metod analizy trwania wyróżnia się modele nieparametryczne, semiparametryczne i parametryczne [Bieszk-Stolorz, Markowicz 2019; Bieszk-Stolorz, Landmesser, Markowicz 2020].

Podstawą analizy trwania jest funkcja przeżycia zdefiniowana następująco [Kleinbaum, Klein 2012]:

$$S(t) = P(t > T) = 1 - F(t) \quad (1)$$

gdzie:  $T$  – czas trwania zjawiska,  $F(t)$  – dystrybuanta rozkładu zmiennej losowej  $T$ .

Funkcja przeżycia (trwania) określa prawdopodobieństwo tego, że określone zdarzenie nie zajdzie co najmniej do czasu  $t$ . Korzystając ze wzoru (1) można wyznaczyć kwartyle czasu trwania. Są to momenty czasu, dla których funkcja trwania przyjmuje odpowiednio wartości 0,75, 0,5, 0,25. Ze względu na możliwość wystąpienia obserwacji cenzurowanych kwartyle mogą nie istnieć. Wynika to z tego, że zdarzenie może nie wystąpić w przyjętym okresie obserwacji.

Najczęściej stosowanym nieparametrycznym estymatorem funkcji trwania jest estymator Kaplana-Meiera określony wzorem [Kaplan, Meier 1958]:

$$\hat{S}(t_i) = \prod_{j=1}^i \left(1 - \frac{d_j}{n_j}\right) \text{ dla } i = 1, 2, \dots, k, \quad (2)$$

gdzie:  $t_i$  – punkt czasu, w którym wystąpiło co najmniej jedno zdarzenie,  $t_1 < t_2 < \dots < t_k$ ,  $t_0 = 0$ ,  $d_i$  – liczba zdarzeń w czasie  $t_i$ ,  $n_i$  – liczba jednostek objętych obserwacją w czasie  $t_i$ ,  $n_i = n_{i-1} - d_{i-1} - z_{i-1}$ ,  $z_i$  – liczba obserwacji cenzurowanych w czasie  $t_i$ .

Krzywą trwania można wyznaczyć dla wszystkich obserwacji ogółem, a także dla grup wyodrębnionych ze względu na cechy jednostek. Krzywe takie można porównywać. Istnieje wiele testów do badania istotności różnic pomiędzy dwiema krzywymi przeżycia i różna jest ich moc. Weryfikowana jest hipoteza zerowa postaci  $H_0: S_1(t) = S_2(t)$ . Hipoteza alternatywna może przyjąć jedną z następujących postaci:  $H_1: S_1(t) \neq S_2(t)$ ,  $H_1: S_1(t) > S_2(t)$  lub  $H_1: S_1(t) < S_2(t)$ . Nie ma kryteriów pozwalających rozstrzygnąć, który z testów ma największą moc i powinien być wykorzystany w danym badaniu. Niektóre z nich są bardziej czułe na przebieg krzywych trwania w początkowej ich fazie, a inne w końcowej. Badania przeprowadzone przez Latta [1981] wskazują, że moc testów zmienia się również w zależności od wielkości badanej próby, od sposobu cenzurowania i rozkładu funkcji hazardu. Do porównania przebiegu krzywych przeżycia dla dwóch grup wykorzystano dwa testy: test Peto i Peto [Peto, Peto 1972] oraz test log-rank [Mantel, Haenszel 1959; Mantel 1966]. Test Peto i Peto przypisuje wagi zależne od oszacowanego percentyla rozkładu czasu trwania [Karadeniz, Ercan 2017]. Zdarzenia zachodzące wcześniej, gdy szacowana funkcja przeżycia jest duża, otrzymują większe wagi, natomiast zdarzenia znajdujące się w prawym ogonie rozkładu czasu trwania – mniejsze [Prentice, Marek 1979]. Test ten stosuje się, gdy współczynnik ryzyka pomiędzy grupami nie jest stały (Stevenson, 2009). Test log-rank jest najczęściej stosowanym testem do porównywania krzywych trwania [Karadeniz, Ercan 2017]. Przywiązuje taką samą wagę do wczesnych i późnych zdarzeń [Stevenson 2009; Allison 2010]. Test log-rank zakłada, że funkcje hazardu dla obu grup są równoległe, co oznacza, że współczynniki ryzyka dwóch grup są stałe we wszystkich okresach [Dawson, Trapp 2001; Stevenson 2009].

Wiele parametrycznych modeli przeżycia to modele przyspieszonej porażki (AFT), w których czas trwania jest wyrażany jako funkcja zmiennych objaśniających. Założenie leżące u podstaw modeli AFT mówi, że dla dowolnej ustalonej wartości funkcji przeżycia  $S(t)$  stosunek czasów trwania dla dwóch jednostek jest stały, czyli  $T_1/T_2 = const$ . Oznacza to również, że  $S_1(t) = S_2(\psi t)$  dla  $t \geq 0$ . Występujące w tych modelach zmienne objaśniające powodują odpowiednie „kurczenie się” lub „wydłużenie” czasu. Przyjmując  $\psi = \exp(-\mathbf{x}'\boldsymbol{\beta})$ , założenie AFT można wyrazić jako  $S_1(t) = S_2(\exp(-\mathbf{x}'\boldsymbol{\beta}) t)$ .

Formalnie, model AFT można zapisać w następującej postaci:

$$\ln T = \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon \quad (3)$$

gdzie:  $T$  – czas trwania,  $\mathbf{x}$  – wektor zmiennych objaśniających,  $\boldsymbol{\beta}$  – wektor parametrów modelu,  $\varepsilon$  – składnik losowy,  $\sigma$  – czynnik skalujący.

Różne rozkłady składnika losowego  $\varepsilon$  prowadzą do różnych postaci modeli AFT. Po przekształceniu równania (1) otrzymuje się  $T \exp(-\mathbf{x}'\boldsymbol{\beta}) = \exp(\sigma\varepsilon)$ , czyli  $T\psi = \exp(\sigma\varepsilon)$ . Parametr  $\psi$  nosi nazwę parametru przyspieszającego, a działanie jego polega na skalowaniu czasu. U jednostki, dla której  $\psi > 1$ , czas jest przyspieszany i zdarzenie wystąpi szybciej; z kolei  $\psi < 1$  oznacza, że czas jest spowalniany, a zdarzenie nastąpi z opóźnieniem. Podejmowana w modelach AFT parametryzacja dotyczy rozkładu składnika losowego  $\varepsilon$ . Jeśli dla  $\varepsilon$  przyjmie się rozkład wartości ekstremalnych z jednym parametrem, to czas trwania  $T$  będzie podlegał rozkładowi wykładniczemu, jeśli zaś  $\varepsilon$  będzie miał rozkład wartości ekstremalnych z dwoma parametrami, to  $T$  zyska rozkład Weibulla. Założenie dla  $\varepsilon$  rozkładu logistycznego wiąże się z rozkładem log-logistycznym dla  $T$ , zaś rozkład normalny dla  $\varepsilon$  skutkuje dla  $T$  rozkładem logarytmiczno-normalnym.

W wykładniczym modelu AFT funkcja przeżycia ma postać  $S(t) = \exp(-\exp(-\mathbf{x}'\boldsymbol{\beta})t)$ , zaś w modelu Weibulla  $S(t) = \exp(-\exp(-p\mathbf{x}'\boldsymbol{\beta})t^p)$ ,  $p > 0$ . Oba te modele spełniają założenia zarówno AFT jak i proporcjonalnego hazardu. Wiele zjawisk społecznych charakteryzuje niska początkowa intensywność procesu, rosnąca do pewnego momentu, a następnie malejąca (np. wychodzenie ze stanu bezrobocia). W takiej sytuacji zastosowanie może znaleźć rozkład logarytmiczno-normalny lub log-logistyczny. W modelu logarytmiczno-normalnym funkcja przeżycia wyraża się jako  $S(t) = 1 - \Phi\left(\frac{\ln t - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)$ , gdzie  $\Phi(z)$  jest funkcją dystrybuanty dla standardowego rozkładu normalnego. W modelu log-logistycznym funkcja przeżycia to  $S(t) = \frac{1}{1 + (\lambda t)^{1/\gamma}}$ , gdzie  $\lambda = \exp(-\mathbf{x}'\boldsymbol{\beta})$  oraz  $\gamma > 0$ . Gdy  $\gamma \geq 1$  funkcja hazardu jest monotonicznie malejąca; gdy  $\gamma < 1$  funkcja hazardu najpierw rośnie, potem maleje. Log-logistyczny model AFT nie spełnia założenia modelu proporcjonalnych hazardów, spełnia natomiast założenie modelu proporcjonalnych „odds” [Kleinbaum, Klein 2012]. Wyboru najlepszego modelu przeżycia spośród kilku alternatywnych modeli o różnych rozkładach podstawowych można dokonać na podstawie wartości kryterium informacyjnego Akaike’a (AIC).

## DANE WYKORZYSTANE W BADANIU

W badaniu wykorzystano dane indywidualne z Powiatowego Urzędu Pracy (PUP) w Szczecinie. Badanie zatem dotyczyło bezrobocia rejestrowanego. Należy tutaj zaznaczyć, że źródłem takich danych w Polsce są jedynie powiatowe urzędy pracy. Główny Urząd Statystyczny w Polsce i Eurostat dysponują jedynie danymi zagregowanymi, które są nieprzydatne w przypadku badań z wykorzystaniem



modeli analizy trwania. Z tego powodu dane wykorzystane w badaniu są unikatowe. Pozyskane dane obejmowały 193919 osób wyrejestrowanych z urzędu pracy w latach 2011-2021. Dane te zawierały datę rejestracji oraz datę wyrejestrowania, płeć, poziom wykształcenia, wiek osoby wyrejestrowanej, staż pracy, a także powód wyrejestrowania. Powody wyrejestrowania mogą być różne. Jest ich kilkadziesiąt. Ponieważ analizowano zdarzenie polegające na podjęciu pracy, więc wszystkich bezrobotnych podzielono na dwie grupy: tych, którzy podjęli pracę oraz tych, którzy zostali wyrejestrowani z pozostałych przyczyn (np. rezygnacja z pośrednictwa urzędu, wyjazd za granicę, przejście na rentę lub emeryturę). Na podstawie daty zarejestrowania i wyrejestrowania wyznaczono czas bezrobocia rejestrowanego, czyli zmienną losową  $T$ . Zdarzeniem kończącym obserwację jest podjęcie pracy. Wyrejestrowanie z innych powodów zostało przyjęte za obserwację cenzurowaną. Analizę przeprowadzono dla 11 lat wchodzących w skład okresu badawczego. 55,33% wszystkich obserwacji stanowiły obserwacje cenzurowane. W tabeli 1 przedstawiono cechy osób bezrobotnych wraz z podziałem na kategorie oraz zanzaczono grupy referencyjne.

Tabela 1. Cechy osób bezrobotnych i ich kategorie

Cecha	Kategoria	Cecha	Kategoria
Płeć	Kobieta		Co najwyżej gimnazjalne*
	Mężczyzna*		Zasadnicze zawodowe
Grupa wieku	18-24*	Poziom wykształcenia	Średnie ogólnokształcące
	25-34		Średnie zawodowe
	35-44		Wyższe
	45-54	Staż pracy	Brak stażu*
	55-59		Posiadanie stażu
	60-64		

\* – kategoria referencyjna

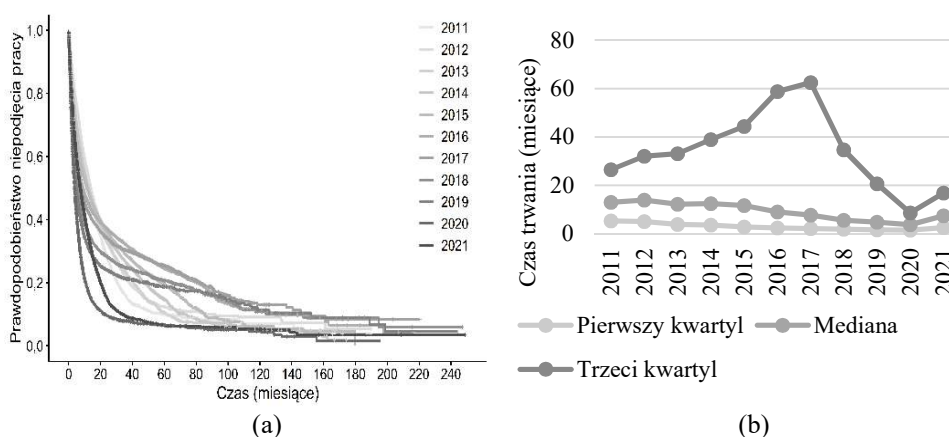
Źródło: opracowanie własne

## WYNIKI BADANIA

Badanie przeprowadzono w dwóch etapach. Pierwszy etap obejmował wstępną ocenę czasu wychodzenia z bezrobocia rejestrowanego. Na rysunku 1(a) przedstawiono oszacowania Kaplana-Meiera dla krzywych przeżycia dla 11 analizowanych lat (2011-2021). W tym przypadku krzywe przeżycia (trwania) opisują prawdopodobieństwo niewyrejestrowania z powodu podjęcia pracy, czyli inaczej mówiąc prawdopodobieństwo niepodjęcia pracy. Na rysunku 1(b) przedstawiono kwartyle czasu trwania wyznaczone na podstawie oszacowań Kaplana-Meiera. Krzywe przeżycia przedstawione na rysunku 1(a) różnią się przebiegiem. Różnice w czasach trwania są również widoczne na rysunku 1(b). Wartości pierwszego kwartyła i mediany malały do roku 2020. W 2021 roku

nastąpił wzrost ich wartości. Nieco inne były zmiany wartości trzeciego kwartyla. Jego wartości rosły do 2017 roku, następnie malały do 2020 roku. W 2021 roku nastąpił wzrost wartości trzeciego kwartyla w porównaniu do roku poprzedniego. Wynika z tego, że czas do podjęcia pracy przez osoby zarejestrowane w PUP ulegał systematycznemu skróceniu, co jest zjawiskiem pożądanym. Jednak pogarszała się sytuacja osób długotrwale bezrobotnych. Od 2017 roku widoczne jest skrócenie czasu do podjęcia pracy przez wszystkich bezrobotnych. W najlepszym, 2020 roku, 75% osób wyrejestrowanych z PUP, to osoby o czasie zarejestrowania nie dłuższym niż 8,56 miesiąca.

Rysunek 1. Oszacowania Kaplana-Meiera dla krzywych przeżycia (ocena prawdopodobieństwa niepodjęcia pracy) i kwantyle czasu trwania



Źródło: opracowanie własne w programie R (a) i programie Excel (b)

Różnice w przebiegu krzywych trwania zostały potwierdzone testem dla wielu krzywych ( $\chi^2 = 4814$  dla 10 stopni swobody,  $p < 2e-16$ ). Test ten jest rozwinięciem (lub uogólnieniem) testu Gehana, Peto i Peto oraz log-rank. Jest on oparty na statystyce chi-kwadrat. Korzystając z testu Peto i Peto oraz testu log-rank porównano parami wszystkie krzywe trwania. Wyniki przedstawiono w tabeli 2. Na ich podstawie należy stwierdzić, że w większości przypadków należy odrzucić hipotezę zerową o równości krzywych trwania. Jednak w kilku przypadkach brak jest podstaw do odrzucenia hipotezy zerowej. Wartości testu Peto i Peto wskazują na to, że przebieg krzywych trwania 2011 i 2012, 2013 i 2014, w ich początkowej fazie nie różni się istotnie od siebie. Natomiast wartości testu log-rank wskazują na to, że w całym przebiegu krzywe trwania 2014 i 2015, 2018 i 2021 nie różnią się istotnie od siebie.

Tabela 2. Porównanie krzywych przeżycia – wartości  $p$  testu Peto i Peto (log-rank)

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
2012	0,1426 (0,0010)									
2013	0,0000 (0,0001)	0,0000 (0,0000)								
2014	0,0000 (0,0061)	0,0000 (0,0000)	0,4726 (0,0240)							
2015	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,1514)	0,0000 (0,0008)						
2016	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0004)					
2017	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0004 (0,0007)				
2018	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)			
2019	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0078 (0,0002)		
2020	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	
2021	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0000 (0,0000)	0,0149 (0,0000)	0,0000 (0,4857)	0,0000 (0,0000)	0,0000 (0,0000)

Źródło: opracowanie własne

W drugim etapie badania wykorzystano modele przyspieszonej porażki (AFT). Spośród możliwych do opisu procesu podejmowania pracy przez osoby bezrobotne modeli AFT rozpatrzono cztery: wykładniczy, Weibulla, log-normalny i log-logistyczny. W pierwszym kroku oszacowano równania modeli dla całego badanego okresu czasu (2011-2021) na podstawie połączonej próby (obserwacji nie grupowano w żaden sposób) uzyskując najlepszy wynik dla modelu log-logistycznego (ten model charakteryzowała najmniejsza wartość kryterium AIC; patrz tabela 3).

Tabela 3. Współczynniki w modelach AFT szacowanych na podstawie połączonych danych (lata 2011-2021 razem)

Zmienna	Rozkłady			
	wykładniczy	Weibulla	log-normalny	log-logistyczny
plec	0,256***	0,222***	0,174***	0,205***
wiek2	0,378***	0,343***	0,251***	0,277***
wiek3	0,607***	0,549***	0,379***	0,404***
wiek4	0,683***	0,591***	0,334***	0,399***
wiek5	0,992***	0,935***	0,569***	0,710***
wiek6	1,749***	1,815***	1,363***	1,571***
wyksz2	-0,340***	-0,444***	-0,398***	-0,434***
wyksz3	-0,460***	-0,604***	-0,533***	-0,594***
wyksz4	-0,509***	-0,680***	-0,581***	-0,658***
wyksz5	-0,921***	-1,175***	-1,063***	-1,119***
staz	-0,261***	-0,347***	-0,349***	-0,324***

cons	2,743***	3,092***	2,605***	2,563***
p		0,740***		
sigma			1,879***	
gamma				1,047***
Obs	191884	191884	191884	191884
lnL	-232705,1	-224803,7	-224475,3	-223878,9
AIC	465434,2	449633,5	448976,6	<b>447783,8</b>

\*\*\* – poziom istotności 0,01

Źródło: opracowanie własne w programie R

Następnie wszystkie powyższe modele AFT szacowano w podokresach rocznych. W tabeli 4 przedstawiono liczbę obserwacji oraz wartości kryterium AIC dla oszacowanych 44 modeli. Można zauważyć, że monotoniczny hazard Weibulla był najbardziej adekwatny dla danych z lat 2011-2014, następnie w latach 2015-2017 najlepiej wypadł model logarytmiczno-normalny opisujący niemonotoniczny hazard, a dla lat 2018-2021 najlepszym modelem okazał się model log-logistyczny, który może opisywać hazard zarówno monotoniczny jak i niemonotoniczny.

Tabela 4. Liczba obserwacji i wartość AIC w szacowanych modelach AFT (wyłuszczone najniższe wartości dla kryterium AIC)

Rok	Liczba obserwacji	Obserwacje niepełne (niepodjęcie pracy)	Rozkłady			
			wykładniczy	Weibulla	log-normalny	log-logistyczny
2011	15194	60,82%	29484,05	<b>29457,36</b>	30436,40	29752,33
2012	15570	61,27%	31137,57	<b>30978,02</b>	31711,41	31159,57
2013	23762	53,87%	56404,34	<b>55539,87</b>	56303,12	55800,91
2014	24443	55,09%	56872,58	<b>55803,42</b>	56156,18	55920,95
2015	25568	56,72%	62502,88	59845,83	<b>59732,64</b>	59742,03
2016	23447	57,55%	59790,07	55632,09	<b>54828,85</b>	55038,44
2017	19697	59,71%	46912,42	44151,29	<b>43562,79</b>	43599,24
2018	14873	58,15%	35806,34	33752,23	33070,93	<b>33046,76</b>
2019	12680	55,50%	29829,79	28845,56	27980,29	<b>27975,22</b>
2020	7772	31,24%	22084,51	21842,82	21505,41	<b>21424,95</b>
2021	8878	36,89%	24753,41	24553,84	24623,19	<b>24512,29</b>

Źródło: opracowanie własne w programie R

W celu prześledzenia wpływu zmiennych objaśniających na wyjście osób z bezrobocia w PUP Szczecin w okresie 2011-2021, zdecydowano się dalszą analizę ograniczyć do jednego typu modeli AFT – modelu log-logistycznego. Model ten jest elastyczny pod względem opisu przebiegu hazardu oraz dla czterech spośród modelowanych 11 lat okazał się najlepszy. Tabela 5 prezentuje uzyskane w procesie estymacji wartości współczynników beta oraz gamma. Przykładowa interpretacja oceny parametru stojącego przy zmiennej płeć w modelu dla 2011 roku jest następująca: kobiety, w porównaniu z mężczyznami miały dłuższy czas wychodzenia z bezrobocia do pracy (ponieważ  $t_K = \exp(0,085)t_M = 1,089t_M$ ).

Dla wszystkich lat parametr ten przyjmował wartość dodatnią, wskazując tym samym na trudniejszą pozycję na rynku pracy kobiet niż mężczyzn, zaś w latach 2016-2019 sytuacja ta wydaje się najtrudniejsza.

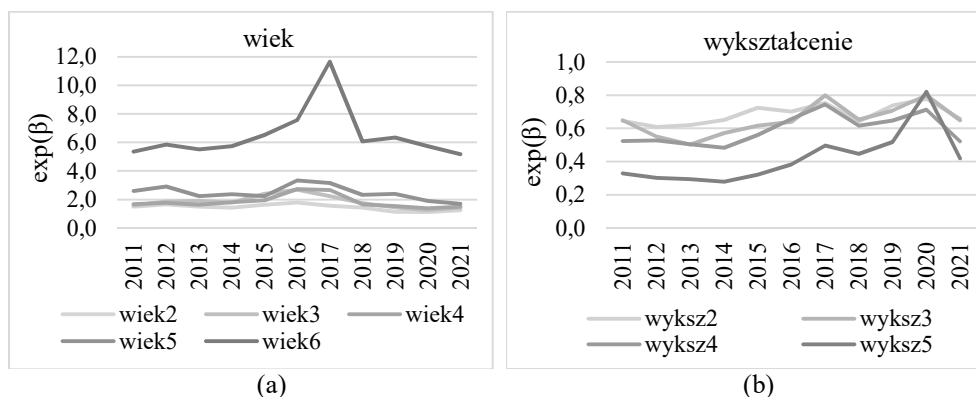
Tabela 5. Oceny parametrów w modelach log-logistycznych szacowanych dla lat 2011-2021

Zmienna	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
plec	0,085***	0,154***	0,087***	0,118***	0,216***	0,398***	0,404***	0,297***	0,339***	0,194***	0,113***
wiek2	0,408***	0,501***	0,402***	0,361***	0,490***	0,580***	0,445***	0,353***	0,127**	0,110	0,226***
wiek3	0,462***	0,623***	0,648***	0,587***	0,869***	0,989***	0,802***	0,562***	0,368***	0,228***	0,360***
wiek4	0,515***	0,568***	0,496***	0,594***	0,665***	1,004***	0,983***	0,506***	0,429***	0,321***	0,406***
wiek5	0,957***	1,072***	0,807***	0,861***	0,809***	1,207***	1,147***	0,845***	0,876***	0,645***	0,534***
wiek6	1,676***	1,765***	1,707***	1,746***	1,874***	2,023***	2,456***	1,803***	1,848***	1,746***	1,644***
wyksz2	-0,437***	-0,496***	-0,476***	-0,429***	-0,322***	-0,354***	-0,282***	-0,445***	-0,303***	-0,251***	-0,416***
wyksz3	-0,432***	-0,598***	-0,689***	-0,557***	-0,484***	-0,447***	-0,224***	-0,427***	-0,348***	-0,224***	-0,432***
wyksz4	-0,649***	-0,639***	-0,684***	-0,728***	-0,579***	-0,420***	-0,295***	-0,483***	-0,434***	-0,337***	-0,652***
wyksz5	-1,112***	-1,199***	-1,227***	-1,278***	-1,136***	-0,959***	-0,702***	-0,805***	-0,660***	-0,198***	-0,871***
staz	-0,400***	-0,462***	-0,785***	-0,666***	-0,721***	-1,012***	-0,720***	0,101**	0,190***	0,025	0,040
cons	2,870***	2,895***	3,099***	2,990***	2,661***	2,432***	1,962***	1,549***	1,371***	1,107***	1,971***
gamma	0,859***	0,918***	0,982***	0,992	1,120***	1,185***	1,124***	1,054***	0,911***	0,773***	0,843***

\*\*\* – poziom istotności 0,01, \*\* – poziom istotności 0,05

Źródło: opracowanie własne w programie R

Rysunek 2. Wpływ wieku oraz wykształcenia na czas trwania w bezrobociu (wartości  $\exp(\beta_k) > 1$  oznaczają wydłużenie czasu trwania w porównaniu z grupą referencyjną,  $\exp(\beta_k) < 1$  jego skrócenie)



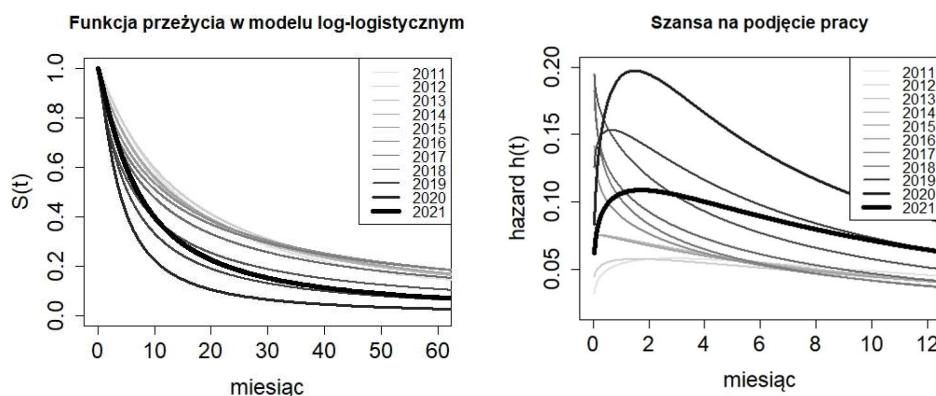
Źródło: opracowanie własne

Analiza wpływu wieku na czas trwania w bezrobociu pozwala zauważyć, że wraz ze wzrostem wieku czas bez pracy się wydłużał. W najgorszej pod tym względem sytuacji w odniesieniu do grupy referencyjnej (osoby w wieku od 18 do 24 lat) były osoby najstarsze (w wieku od 60 do 65 lat). Grupa osób najstarszych znajdowała się w wyjątkowo trudnym położeniu w latach 2016 i 2017, co dodatkowo dokumentuje rysunek 2(a). Na podstawie oszacowanych parametrów w

modelach AFT można też stwierdzić, że czas wychodzenia do pracy skracał się (świadczą o tym ujemne oceny parametrów) wraz ze wzrostem poziomu wykształcenia. Na uwagę jednak zasługuje pogorszenie się sytuacji osób z wykształceniem wyższym w roku 2020 (patrz rysunek 2(b)). Osoby z udokumentowanym stażem zawodowym miały krótszy czas wychodzenia z bezrobocia w latach 2011-2017, zaś od roku 2018 efekt stażu był przeciwny, przy czym w modelach dla lat 2020 i 2021 wystąpił brak statystycznej istotności parametru przy tej zmiennej.

Szacunki parametru gamma w modelach log-logistycznych wskazują na niemonotoniczną (najpierw rosnącą, potem malejącą w miarę trwania w bezrobociu przez kolejne miesiące) intensywność podejmowania zatrudnienia w latach 2011-2014, monotoniczną (malejącą) intensywność w latach 2015-2018 oraz ponownie niemonotoniczną w okresie 2019-2021. Przebiegi krzywych przeżycia oraz hazardu (wyrażającego intensywność wychodzenia z bezrobocia w zatrudnienie) dla poszczególnych lat wynikające z modelu AFT log-logistycznego przedstawiono na rysunku 3. Wynika z niego, że w najlepszej sytuacji bezrobotni znajdowali się w roku 2020.

Rysunek 3. Przebiegi funkcji przeżycia oraz hazardu dla zjawiska wychodzenia z bezrobocia w stan zatrudnienia w PUP Szczecin w latach 2011-2021



Źródło: opracowanie własne

## PODSUMOWANIE

Celem przeprowadzonych badań była ocena wpływu czynników (wiek, płeć, wykształcenie i staż pracy badanych bezrobotnych) na czas trwania bezrobocia w Szczecinie w latach 2011-2021. Analiza danych czasowych dała możliwość sprawdzenia, czy wzorce wychodzenia z bezrobocia ulegały zmianom. Założono, że przyjęte w badaniu zmienne mają wpływ na czas trwania bezrobocia oraz że wpływ ten nie był jednakowy w całym okresie badawczym. Przebieg

wyznaczonych krzywych przeżycia (estymator Kaplana-Meiera) wskazuje na różnice w czasie trwania bezrobocia w kolejnych badanych latach. Czas do podjęcia pracy ulegał systematycznemu skróceniu. Rozwinięciem wstępnego etapu badań było zastosowanie modelu przyspieszonej porażki z rozkładem log-logistycznym dla zmiennej  $T$ . Ogólnie uzyskane wyniki wskazują na wpływ badanych zmiennych na czas trwania bezrobocia. Okazało się, że dłuższy czas dotyczył kobiet w stosunku do mężczyzn, starszych bezrobotnych, z niższym wykształceniem, bez stażu pracy (zazwyczaj). Tym samym pozytywnie została zweryfikowana postawiona hipoteza badawcza H1. W latach 2012-2014 w Polsce zauważalny był wzrost bezrobocia, szczególnie wśród osób młodych. Był to skutek kryzysu gospodarczego z lat 2008-2009. Jednak nasze badania nie wskazują na trudniejszą sytuację młodych osób bezrobotnych w tym czasie. Ciekawym wnioskiem jest natomiast zauważona zmiana wzorców wychodzenia z bezrobocia na przestrzeni czasu, czyli pozytywnie została zweryfikowana również hipoteza badawcza H2. Wyodrębniono trzy charakterystyczne okresy: lata 2011-2014, 2015-2018 i 2019-2020. W okresach tych zmieniała się monotoniczność intensywności podejmowania zatrudnienia przez bezrobotnych. Szczególnym przypadkiem jest 2020 rok, w którym intensywność ta była duża w początkowym okresie trwania bezrobocia. Zatem kryzysowa sytuacja zdrowotna (Covid-19) nie wpłynęła na wydłużenie czasu poszukiwania pracy, a wręcz przeciwnie czas ten uległ skróceniu.

## BIBLIOGRAFIA

- Allison P. D. (2010) *Survival Analysis Using SAS: A Practical Guide*, 2nd edition, SAS Press, North Carolina.
- Bieszk-Stolorz B., Landmesser J., Markowicz I. (2020) *Analiza trwania w badaniach ekonomicznych. Modele parametryczne*. CeDeWu, Warszawa.
- Bieszk-Stolorz B., Markowicz I. (2019) *Analiza trwania w badaniach ekonomicznych. Modele nieparametryczne i semiparametryczne*. CeDeWu, Warszawa.
- Bieszk-Stolorz B., Markowicz I. (2022) The Impact of the COVID-19 Pandemic on the Situation of the Unemployed in Poland. A Study Using Survival Analysis Methods. *Sustainability*, 14(19), 12677. <https://doi.org/10.3390/su141912677>.
- Cabla A. (2012) Modeling Unemployment Duration in the Czech Republic from LFS. *Research Journal of Economics, Business and ICT*, 7, 1-5.
- Cooper D., McCausland W. D., Theodossiou I. (2006) The health hazards of unemployment and poor education: The socioeconomic determinants of health duration in the European Union. *Economics and Human Biology*, 4, 273-297. <https://doi.org/10.1016/j.ehb.2006.06.001>.
- Dawson B., Trapp R. G. (2001) *Basic&Clinical Biostatistics*. McGraw Hill, Boston.
- Dogan F. I. (2019) Unemployment Benefits and Unemployment Duration in France and Poland. *Marmara Journal of European Studies*, 27(1), 191-216. <https://doi.org/10.29228/mjes.30>.

- Dogan F. I. (2020) Social Transfers and Unemployment Duration: an Empirical Evidence of the EU and Turkey. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 42(2), 247-264. <https://doi.org/10.14780/muiibd.854363>.
- Grzenda W. (2019) Socioeconomic Aspects of Long-Term Unemployment in the Context of the Ageing Population of Europe: the Case of Poland. *Economic Research-Ekonomska Istraživanja*, 32(1), 1561-1582. <https://doi.org/10.1080/1331677X.2019.1638289>.
- Grzenda W. (2023) Estimating the Probability of Leaving Unemployment for Older People in Poland Using Survival Models with Censored Data. *Statistics in Transition new series*, 24(3), 241-256. <https://doi.org/10.59170/stattrans-2023-046>.
- Kaplan E. L., Meier P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481. <https://doi.org/10.2307/2281868>.
- Karwacki A., Błędowski P. (2020) Bezrobocie jako współczesna kwestia społeczna – wybrane aspekty socjologiczne i ekonomiczne. *Studia Socjologiczne*, 1(236), 135-164.
- Kleinbaum D. G., Klein M. (2012) *Survival Analysis. A Self-Learning Text*. 3rd ed.; New York: Springer-Verlag. <https://doi.org/10.1007/978-1-4419-6646-9>.
- Karadeniz P. G., Ercan I. (2017) Examining Tests for Comparing Survival Curves with Right Censored Data. *Statistics in Transition*, 18(2), 311-28.
- Landmesser J. M. (2009) Econometric Analysis of Unemployment Duration Using Hazard Models. *Studia Ekonomiczne*, 1-2, 79-92. <https://www.cceol.com/search/article-detail?id=201814>.
- Landmesser J. M. (2011) The Impact of Vocational Training on the Unemployment Duration. *International Advances in Economic Research*, 17, 89-100. <https://doi.org/10.1007/s11294-010-9291-y>.
- Latta R. B. (1981) A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data. *Journal of the American Statistical Association*, 76(375), 713-719. <https://doi.org/10.2307/2287536>.
- Mala I., Cabla A. (2022) Modelling of the Unemployment Duration in the Czech Republic Based on Aggregated Complete and Individual Censored Data. *Journal of Economics/Ekonomický časopis*, 70(2), 171-187.
- Mantel N., Haenszel W. (1959) Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of National Cancer Institute*, 22 (4), 719-748. <https://www.ncbi.nlm.nih.gov/pubmed/13655060>.
- Mantel N. (1966) Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration. *Cancer Chemotherapy Reports*, 50(3), 163-170. <https://www.ncbi.nlm.nih.gov/pubmed/5910392>.
- Nagler P. (2013) How Unemployment Insurance Savings Accounts Affect Employment Duration: Evidence from Chile. *IZA Journal of Labor & Development*, 2(9). <http://www.izajold.com/content/2/1/9>.
- Nichols A., Mitchell J., Lindner S. (2013) *Consequences of Long-Term Unemployment*. Washington: The Urban Institute Report.
- Petcu M. A., Cismasu I. D. (2023) Exploring Economic Effects of Work Out of Company Office Based on Bibliometric Analysis. *Journal of Business Economics and Management*, 24(4), 676-695. <https://doi.org/10.3846/jbem.2023.20247>.



- Peto R., Peto J. (1972) Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society*, 135(2), 185-207.  
[https://www.jstor.org/stable/2344317?seq=1#page\\_scan\\_tab\\_contents](https://www.jstor.org/stable/2344317?seq=1#page_scan_tab_contents).
- Prentice R. L., Marek P. (1979) A Qualitative Discrepancy between Censored Data Rank Tests. *Biometrics*, 35(4), 861-867.
- Stevenson M. (2009) *An Introduction to Survival Analysis*, EpiCentre, IVABS. Massey Massey University.  
[http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson\\_survival\\_analysis\\_195\\_721.pdf](http://www.massey.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson_survival_analysis_195_721.pdf).

### CHANGES IN PATTERNS OF LEAVING UNEMPLOYMENT: EVIDENCE FROM SZCZECIN

**Abstract:** The article attempts to assess the impact of selected factors (age, gender, education and work experience of the unemployed) on the duration of registered unemployment in Szczecin in years 2011-2023. The study used survival analysis methods, including an acceleration failure time model. The analysis of data in subsequent years was used to assess whether the patterns of leaving unemployment for work changed. This goal was achieved by verifying the research hypotheses. It was assumed that the variables used in the study had an impact on the duration of unemployment and that this impact was not the same throughout the research period.

**Keywords:** unemployment, survival analysis, AFT model

**JEL classification:** J64, C41

## SELEKCJA ZMIENNYCH METODAMI STATYSTYCZNYMI I UCZENIA MASZYNOWEGO. PORÓWNANIE PODEJŚĆ NA PRZYKŁADZIE DANYCH FINANSOWYCH

Urszula Grzybowska  <https://orcid.org/0000-0001-7342-5382>

Marek Karwański  <https://orcid.org/0000-0001-5192-7920>

Institut Informatyki Technicznej  
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie  
e-mail: urszula\_grzybowska@sggw.edu.pl; marek\_karwanski@sggw.edu.pl

**Streszczenie:** Zgodnie z nowymi dyrektywami międzynarodowego nadzoru finansowego (MSSF9) banki powinny przyrzeć się nowemu zestawowi narzędzi analitycznych, takich jak uczenie maszynowe. Wprowadzenie tych metod do praktyki bankowej wymaga przeformułowania celów biznesowych, zarówno w zakresie trafności przewidywań, jak i definicji czynników ryzyka. W artykule porównano metody selekcji zmiennych i przypisania „ważności” w modelach statystycznych i algorytmicznych. Obliczenia przeprowadzono na przykładzie klasyfikacji danych finansowych. Na wybranych zbiorach zmiennych porównano skuteczność różnych algorytmów uczenia maszynowego. Wyniki analiz wskazują na potrzebę rewizji koncepcji „ważności” zmiennej, tak aby nie była ona zależna od struktury modelu.

**Słowa kluczowe:** selekcja zmiennych, uczenie maszynowe, ważność zmiennych

**JEL classification:** C45, C52, C55

### WSTĘP

Analizy ilościowe w finansach bardzo często korzystają z modeli regresyjnych, zarówno w podejściu indywidualnym, jak i zagregowanym. Trzy główne cele modeli to: dokładne przewidzenie wartości wyniku w oparciu o zestaw predyktorów, wyjaśnienie różnic w wartościach wyników poprzez różnice w zmiennych objaśniających, opisanie związku między zmiennymi zależnymi i niezależnymi.

<https://doi.org/10.22630/MIBE.2023.24.4.18>

Zmienne niezależne są wybierane tak, aby reprezentowały spodziewane wpływy na podstawie teorii, wcześniejszych badań i kontekstu lokalnego (w czasie i przestrzeni). W tradycyjnym podejściu przy interpretacji regresji nacisk kładzie się na wielkość i istotność statystyczną współczynników dla zmiennych niezależnych. W modelach algorytmicznych wprowadza się pojęcie ważności zmiennych (importance). Ważności zmiennych to wartości liczbowe przypisywane zmiennym w oparciu o to, jak przydatne są one w przewidywaniu zmiennej docelowej. Istnieje wiele sposobów pomiaru ważności cech. Inaczej wyznaczamy ważność zmiennych w oparciu o metody statystyczne (testowanie hipotez statystycznych), a inaczej w modelach uczenia maszynowego. Osobnym zagadnieniem jest ocena własności predykcyjnych zmiennych w modelach sieci neuronowych. Różne podejścia do tego problemu przedstawiono np. w [Olden 2004]. Uwzględnienie nadmiernej liczby zmiennych w modelu nieznacznie zwiększa dopasowanie modelu. Niestety, takie dopasowanie nie jest pożądane z punktu widzenia biznesowego ponieważ prowadzi do modeli bardziej złożonych. Wprowadzenie modeli algorytmicznych przedstawia inną filozofię, ponieważ miary dopasowania modeli statystycznych i algorytmicznych opierają się na różnych założeniach.

W pracy przedstawiamy zastosowanie różnych metod do selekcji zmiennych na przykładzie zbioru danych z 174 zmiennymi. Celem naszym jest przedstawienie i zastosowanie w kontekście ekonomicznym współczesnych metod selekcji zmiennych. Na danych finansowych pokazujemy zastosowanie metod takich jak LASSO<sup>1</sup>, miara informacji wzajemnej czy rekurencyjna eliminacja cech oraz wyznaczanie ważności zmiennych w modelach algorytmicznych: Lasów losowych, Boostingu gradientowego, XGBoost oraz stosunkowo mało jeszcze popularnej metody ważności cech w sieciach neuronowych. Dla otrzymanych zbiorów zmiennych porównujemy skuteczność metod algorytmicznych oraz regresji logistycznej. W pierwszym rozdziale pracy przedstawiamy różne sposoby doboru zmiennych do modelu wraz z przeglądem literatury dotyczącej tego zagadnienia. W drugiej części opisujemy dane. W trzecim rozdziale przedstawimy wyniki obliczeń dotyczące zarówno wyboru zmiennych jak i sprawdzenia działania modeli uczenia maszynowego na wyróżnionych zbiorach zmiennych. Ostatni rozdział poświęcony jest wnioskowi i podsumowaniu.

Obliczenia wykonano w Python ver. 3.9.

## SELEKCJA ZMIENNYCH

Selekcja zmiennych związana z redukcją wymiaru danych ma kluczowe znaczenie. Pozwala na wyeliminowanie zmiennych nieistotnych, zbędnych, pozwala uniknąć przetrenowania modelu, zwiększa prędkość obliczeniową i umożliwia lepszą interpretację wyników. Jest wiele metod selekcji zmiennych a ich przegląd

---

<sup>1</sup> Metoda wprowadzona w latach 80-tych, zwana także regularyzacją L1. Skrót oznacza Least Absolute Shrinkage and Selection Operator.

można znaleźć w wielu publikacjach (zob. [Li i in. 2017; Pudjihartono i in. 2022; Jia i in. 2022; Zebari i in. 2020]). Zwyczajowo dzielimy je na metody związane z modelem danych tzw. metody wrapper (np. rekurencyjna eliminacja cech, metody heurystyczne) lub metody embedded (Lasy losowe, LASSO, Regresja grzbietowa) [Lal i in. 2006] oraz niezależne od modelu metody filtrowania [Sánchez-Marño i in. 2007; Hopf 2021] np. oparte na korelacji zmiennych czy mierze informacji wzajemnej (MI) [Vergara 2014; Gajowniczek i in. 2022].

### **Modelowanie statystyczne**

Testy hipotez są najpopularniejszym kryterium doboru zmiennych w praktycznych problemach modelowania statystycznego. Testowanie iteracyjne modeli wykonywane jest poprzez algorytmy selekcji zmiennych do przodu (forward selection) lub eliminacji wstecz (backward selection), w zależności od tego, czy zaczyna się od modelu pustego, czy od modelu ze wszystkimi zmiennymi, które mogą być brane pod uwagę. Podczas gdy kryteria istotności są zwykle stosowane w celu włączenia lub wyłączenia zmiennych z modelu, kryteria informacyjne koncentrują się na wyborze modelu z zestawu wiarygodnych modeli. Uwzględnienie większej liczby zmiennych w modelu zwiększa dopasowanie modelu. Niestety takie dopasowanie nie zawsze jest pożądane, gdyż prowadzi do zwiększenia błędu dopasowania. Opracowano kryteria informacyjne w celu uniknięcia tego pozornego efektu dopasowania prowadzącego do wyboru bardziej złożonych modeli. Jako kryteria informacyjne używane są np. statystyki AIC bądź BIC.

Wybór modelu można również przeprowadzić, stosując strategię opartą na tzw. operatorze regularyzacji (LASSO), która polega na nałożeniu dodatkowych warunków na funkcję błędu przy wyliczaniu współczynników regresji [Hastie i in 2008; Hastie i in 2015, str. 32]. Modele LASSO są szeroko stosowane w wielowymiarowych problemach. Współczynniki regresji oszacowane w procedurach LASSO są obciążone, ale mogą mieć mniejszy średni kwadratowy błąd całkowity niż przy konwencjonalnym oszacowaniu. Ze względu na obciążenie ich interpretacja w modelach wyjaśniających lub opisowych jest trudna, a przedziały ufności oparte na procedurach ponownego próbkowania, takich jak bootstrap, nie osiągają deklarowanego poziomu nominalnego [Taylor, Tibshirani 2015].

Wielu autorów podkreśla znaczenie wykorzystywania wiedzy merytorycznej przy doborze zmiennych. W niniejszej pracy, sposób tworzenia generycznego zestawu zmiennych powoduje, że nie musi się uwzględniać empirycznego powiązania zmiennych ze zmienną celu. Innymi słowy można założyć, że merytoryczna wartość wszystkich zmiennych jest taka sama.

### **Modelowanie algorytmiczne**

W przypadku modeli algorytmicznych tworzone są dedykowane sposoby przypisywania ważności zmiennym predykcyjnym. [Elith i in 2008; Adler, Painsky 2022]. Wartości mierzące wagi zmiennych pomagają w interpretacji danych, a także

pozwalają na dokonanie rankingu zmiennych i ułatwiają dobór zmiennych do modelu. W ramach modeli uczenia maszynowego takich jak Drzewa decyzyjne, Lasy losowe, Boosting gradientowy czy XGBoost miary ważności zmiennych są oparte na liczbie przypadków, w których zmienna jest wybierana do podziału, ważonej kwadratową poprawką do modelu dla każdego podziału i uśrednioną dla wszystkich drzew [Elith i in 2008; Ben Jabeur i in. 2023]. W ramach modelowania algorytmicznego możemy także stosować jedną z metod opakowanych (wrapper), rekurencyjną eliminację cech (RFE). W tej metodzie uwzględniane są coraz mniejsze podzbiory zmiennych, najmniej ważne zmienne są usuwane z bieżącego zestawu na podstawie miary ważności, aż do osiągnięcia zadanej z góry liczby zmiennych [Kohavi, John 1997]. Dla sieci neuronowych jednym ze sposobów wyboru zmiennych jest metoda VIANN oparta na modyfikacji algorytmu Welforda [De Sa 2019]. W celu oceny ważności zmiennej  $x_s$  używamy miary opartej na uśrednionej wariancji zmian wag-parametrów sieci pierwszej warstwy ukrytej podłączonej do zmiennej wejścia  $x_s$  w całym procesie propagacji wstecznej. Oznacza to, że końcowy wynik ważności zmiennej będzie zależał zarówno od wag-parametrów końcowych, jak i od odchylenia ich podczas treningu. Zakłada się, że im bardziej waga  $w_{(a,b)}$  połączenia  $(a, b)$  zmienia się w fazie uczenia, tym większe znaczenie węzła  $a$  w procesie predykcji. Korzystając z VIANN, musimy określić, na jakich etapach uczenia aktualizujemy wariancję. Można rozważyć kilka opcji, ze względu na iterację (po każdej partii), na epokę lub interwał zdefiniowany przez użytkownika. Dla uproszczenia w pracy aktualizujemy wariancję wag w każdej epoce.

Obok sposobu wyboru zmiennych w ramach metod uczenia maszynowego, istnieją także metody hybrydowe, które np. łączą metody filarycyjne z metodami opartymi na uczeniu maszynowym. W ostatnim czasie dla dużych zbiorów stosowane są metody heurystyczne wyboru zmiennych do modelu [Jia i in. 2022].

W pracy zastosujemy model hybrydowy. Osobno dokonamy selekcji zmiennych kategorycznych i ciągłych. Pierwszym krokiem będzie filtacja zmiennych. Dla zmiennych kategorycznych stosujemy miarę  $\chi^2$  oraz miarę informacji wzajemnej (MI), która jest znana także jako przyrost informacji lub entropia krzyżowa. Dla zmiennych ciągłych dokonamy eliminacji zmiennych silnie skorelowanych oraz zmiennych z dużym VIF<sup>2</sup>. Dla pozostałych zmiennych ciągłych dokonamy wyboru zmiennych metodą LASSO, Regresji grzbietowej oraz obliczając ważność zmiennych w modelach uczenia maszynowego.

## OPIS DANYCH

Kredyty w rachunku bieżącym dla segmentu małych i średnich przedsiębiorstw (MŚP) są kredytami odnawialnymi ze stałym limitem pieniężnym.

---

<sup>2</sup> Wskaźnik inflacji wariancji służący do badania współliniowości zmiennych.

Informacje potrzebne do budowy modeli analitycznych zbierane są w kilku systemach informatycznych. Dla celów niniejszego artykułu wykorzystano dane jednego z polskich banków. Dane zostały poddane modyfikacjom tak aby niemożliwa była ich identyfikacja a jednocześnie zachowane relacje pomiędzy atrybutami. W tym celu stworzony został „model danych kredytowych” – zdefiniowano atrybuty transakcji/klienta, a następnie przeliczono sztuczne atrybuty stworzone na podstawie pierwotnych danych.

Na potrzeby budowy modeli parametrów ryzyka dokonano wstecznego przeglądu defaultów i oznaczono defaulty zgodnie z powszechnie stosowanymi procedurami. Zastosowano metodologię stałego horyzontu czasowego [Engelmann i Rauhmeier 2011]. Zaletami stosowania tej metodologii są: (1) rozproszenie dat referencyjnych po całym okresie zbierania danych, (2) zastosowanie wspólnego horyzontu T przyczynia się do jednorodności realizowanej straty ekonomicznej, (3) łatwość uzyskania informacji z RDS (Reference Data Set) dostępnego w banku. Metoda stałego horyzontu czasowego ma również wady, które należy rozważyć: (1) ustalony horyzont czasowy T nie pozwala na kalkulację w wielu horyzontach, (2) nie można uwzględnić bezpośrednio transakcji przeterminowanych, gdy wiek transakcji jest mniejszy niż T, (3) nie uwzględnia wszystkich istotnych informacji, ponieważ w przypadku niewykonania zobowiązania w okresie obserwacji, wykorzystuje się obserwację tylko z tego okresu, (4) nie uwzględnia się możliwości, że bieżące ekspozycje mogą nie zostać zrealizowane w dowolnym momencie następnego okresu.

W procesie zbierania danych wykorzystano informacje z kilku źródeł wykorzystywanych w banku w ciągu długiego okna czasowego (modele TTC). Aby stworzyć Tabelę Analityczną (Analytical Base Table) wykorzystano dane z następujących systemów:

1. Oceny wniosków kredytowych (Applicatins evaluations)
2. Historia klienta (Customer relationship history)
3. Spłat (Repayment history)
4. Inne informacje kredytowe (Other loan informations)
5. Informacje o wykorzystaniu produktów (General information on product usage)
6. Segmentacja klienta (Customer segmentation)
7. Operacje gotówkowe (Cash based transactions)
8. Rozliczenia (Settelments)
9. Zabezpieczenia (Collarerals)
10. Ratingi BIK (Credit Bureau ratings)
11. Informacje o właścicielach (SME owners information)

Model danych kredytowych zebrano w tzw. *modelu generycznym* – zaproponowanym „z góry” układzie atrybutów, który zawierał definicje czynników ryzyka podzielone na dwie grupy: podstawowe informacje o klientach i informacje dodatkowe o kontaktach. Model ten został zbadany przez ekspertów biznesowych. Główną ideą modelu generycznego był wybór jak największej liczby atrybutów -

„słabych uczniów” i stworzenie odpowiedniej reguły klasyfikacji opartej na wnioskach statystycznych, a nie biznesowych. Zgodnie z tym podejściem procedura budowy modelu polegała na określeniu modelu maksymalnego, tj. modelu z największą liczbą zmiennych, a następnie utworzenie finalnego modelu poprzez usunięcie zmiennych z modelu maksymalnego.

Ostatecznie zebrane dane zawierały 6590 obserwacji wśród których było 1014 defaultów. Każda obserwacja opisywana była przez 174 atrybuty (zmienne). W zbiorze pierwszych 14 zmiennych stanowiły zmienne mierzone w skali nominalnej pozostałe mierzone w skali interwałowej. Zmienną targetową była zmienna binarna default = 1, non-default = 0.

## WYNIKI OBLICZEŃ

### Selekcja zmiennych kategorycznych

W pierwszym kroku wyboru zmiennych usunięto zmienne o zbyt dużej liczbie kategorii ( $X_4$ - 31 kategorii,  $X_7$ - 55 kategorii), zmienną o 2 kategoriach, gdzie druga liczyła 1 obserwację ( $X_6$ ). Dla pozostałych zmiennych kategorie z małą liczbą obserwacji zostały scalone. Ostatecznie wykorzystano 11 zmiennych nominalnych. Dla tego zbioru zbudowano model regresji logistycznej z selekcją zmiennych metodą krokową w przód. Trzy zmienne nie zostały włączone do modelu. Istotne okazały się zmienne  $X_{11}$ ,  $X_2$ ,  $X_{12}$ ,  $X_{13}$ ,  $X_5$  i  $X_{10}$ . Następnie dla 11 zmiennych obliczono miarę informacji wzajemnej, która wyróżniła zmienne  $X_{11}$ ,  $X_5$ ,  $X_{12}$  i  $X_1$ . Z kolei miara oparta na teście  $\chi^2$  wskazała, że nie ma zależności między zmienną  $y$  a zmiennymi  $X_3$ ,  $X_8$  i  $X_{13}$ . Największy wpływ na  $y$  mają zmienne:  $X_{11}$ ,  $X_{14}$  i  $X_{12}$ . Metody uczenia maszynowego pozwalają na wyznaczenie ważności zmiennych (zobacz tabela 1).

Tabela 1. Ranking zmiennych kategorycznych uzyskany metodami uczenia maszynowego

	RF		XGBoost		GB	
	Variable	Importance	Variable	Importance	Variable	Importance
1	$X_{11}$	0,4684	$X_{11}$	0,5005	$X_{11}$	0,7232
2	$X_2$	0,0964	$X_2$	0,1249	$X_2$	0,1280
3	$X_3$	0,0880	$X_{12}$	0,0751	$X_{12}$	0,0515
4	$X_5$	0,0801	$X_1$	0,0428	$X_{13}$	0,0191
5	$X_8$	0,0620	$X_{14}$	0,0418	$X_5$	0,0187
6	$X_9$	0,0479	$X_{10}$	0,0410	$X_{10}$	0,0173
7	$X_{10}$	0,0446	$X_{13}$	0,0384	$X_9$	0,0158
8	$X_{12}$	0,0440	$X_9$	0,0382	$X_{14}$	0,0141
9	$X_{13}$	0,0387	$X_5$	0,0365	$X_8$	0,0056
10	$X_{14}$	0,0210	$X_8$	0,0309	$X_1$	0,0041
11	$X_1$	0,0089	$X_3$	0,0300	$X_3$	0,0028

Źródło: opracowanie własne

Wszystkie 3 rozważane metody wskazały zmienne  $x_{11}$  i  $x_2$  jako najważniejsze. Kolejną pozycję zajmowała zmienna  $X_{12}$  dla Boostingu gradientowego (GB) i algorytmu XGBoost, zaś Lasy losowe (RF) wskazały jako najważniejszą zmienną  $X_{11}$ .

### Wybór zmiennych ciągłych

W bazie występowało 160 zmiennych typu ciągłego. W pierwszym kroku wyeliminowano zmienne skorelowane, przyjmując jako wartość progową 0,9. Usunięto także zmienne z wysokim VIF, przyjmując jako próg wartość 20. Pozostały 73 zmienne. Usunięto zmienne quasi-stałe. Do dalszej analizy pozostało 68 zmiennych. Dalszej selekcji zmiennych dokonano metodami uczenia maszynowego a także za pomocą regularyzacji L1 (LASSO) i L2 (regresja grzbietowa, Ridge) w regresji logistycznej, które są metodami typu osadzonego (embedded methods). Wyznaczono ważność zmiennych metodami uczenia maszynowego tj. przy użyciu Lasów losowych (RF), algorytmu XGBoost (XGB), Boostingu gradientowego (GB) oraz sztucznych sieci neuronowych (ANN). Wartości miar ważności dla najlepszych 11 zmiennych przedstawiono w tabeli 2.

Tabela 2. Ranking zmiennych uzyskany metodami uczenia maszynowego

	RF		XGBoost		GB		ANN	
1	$X_{152}$	0,29	$X_{152}$	0,879	$X_{152}$	0,993	$X_{152}$	1,000
2	$X_{153}$	0,11	$X_{74}$	0,009	$X_{103}$	0,004	$X_{150}$	0,721
3	$X_{149}$	0,08	$X_{50}$	0,008	$X_{153}$	0,001	$X_{147}$	0,398
4	$X_{161}$	0,05	$X_{125}$	0,007	$X_{50}$	0,001	$X_{38}$	0,252
5	$X_{147}$	0,05	$X_{103}$	0,007	$X_{66}$	0,001	$X_{153}$	0,175
6	$X_{118}$	0,05	$X_{39}$	0,007	$X_{38}$	0,000	$X_{97}$	0,163
7	$X_{154}$	0,03	$X_{47}$	0,006	$X_{44}$	0,000	$X_{145}$	0,142
8	$X_{129}$	0,03	$X_{90}$	0,005	$X_{47}$	0,000	$X_{143}$	0,132
9	$X_{103}$	0,03	$X_{153}$	0,005	$X_{150}$	0,000	$X_{57}$	0,108
10	$X_{39}$	0,03	$X_{38}$	0,005	$X_{46}$	0,000	$X_{120}$	0,107
11	$X_{76}$	0,03	$X_{91}$	0,005	$X_{147}$	0,000	$X_{160}$	0,105

Źródło: opracowanie własne

Każda ze stosowanych metod uczenia maszynowego wyróżniła inne zmienne, choć niektóre zmienne np.  $X_{152}$  i  $X_{153}$  powtarzają się w każdym rankingu. Obliczenie korelacji rang dla wszystkich 68 zmiennych zamieszczone w tabeli 3 pokazuje brak korelacji między rangami zmiennych, których kolejność, poprzez przyporządkowanie ważności, została wyznaczona w poszczególnych modelach uczenia maszynowego. Wizualizacja związku między rangami zamieszczona została na rysunku 1.



Tabela 3. Współczynniki korelacji rang Kendalla wraz z odchyleniami standardowymi

	FR	XGBoost	GB	ANN
FR	1,0000	0,1624	-0,1229	0,1484
		0,0502	0,1383	0,0736
XGBoost	0,1624	1,0000	0,2388	-0,1062
	0,0502		0,0040	0,2002
GB	-0,1229	0,2388	1,0000	-0,0667
	0,1383	0,0040		0,4211
ANN	0,1484	-0,1062	-0,0667	1,0000
	0,0736	0,2002	0,4211	

Źródło: opracowanie własne

Następnie zbudowano model Regresji logistycznej (LR) dla wszystkich 68 zmiennych i dla tego modelu zastosowano regularyzację L1 (LASSO) z parametrem  $c = 0,01$ . Regularyzacja wyróżniła następujące zmienne jako istotne:  $X_{24}, X_{25}, X_{32}, X_{43}, X_{44}, X_{45}, X_{46}, X_{49}, X_{59}, X_{66}, X_{74}, X_{76}, X_{84}, X_{87}, X_{103}, X_{114}, X_{115}, X_{116}, X_{118}, X_{129}, X_{133}, X_{134}, X_{136}, X_{143}, X_{147}, X_{149}, X_{150}, X_{152}, X_{153}, X_{160}, X_{162}, X_{171}$ .

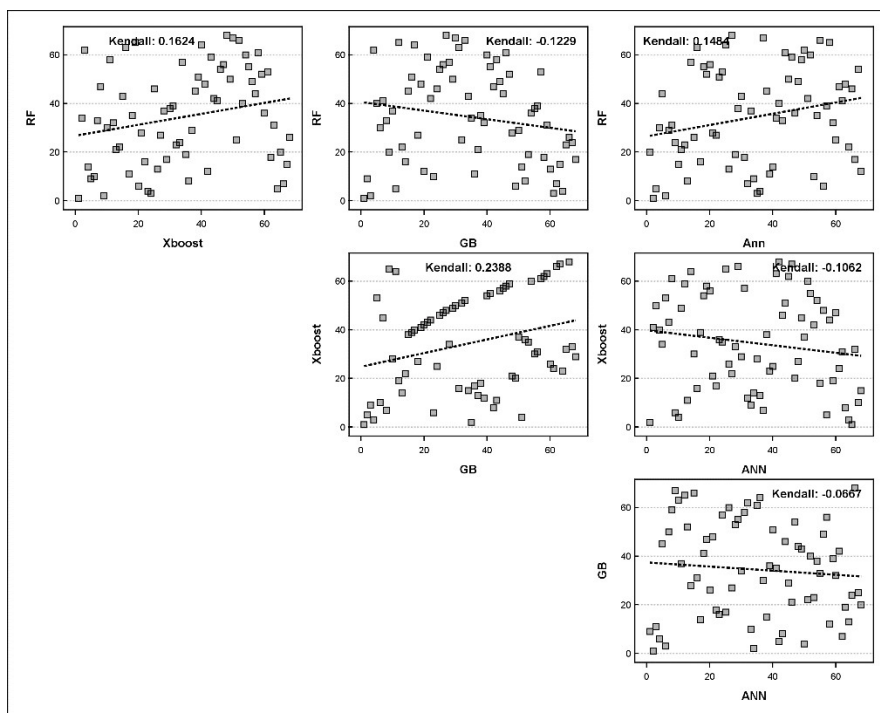
Regularyzacja L2 (grzbietowa) z parametrem  $c = 0,05$  wskazała następujące zmienne jako istotne:  $X_{44}, X_{45}, X_{46}, X_{59}, X_{66}, X_{87}, X_{103}, X_{114}, X_{118}, X_{129}, X_{147}, X_{149}, X_{150}, X_{152}, X_{153}, X_{162}$ . Te same zmienne uzyskano dla regularyzacji L2 z parametrem  $c = 0,01$ .

Przeprowadzono także rekurencyjną eliminację cech (RFE). Jest to metoda kosztowana obliczeniowo. W pierwszej kolejności przeprowadzono rekurencyjną eliminację cech biorąc jako wyjściowy zbiór wszystkie 68 zmiennych. Przy wyborze 5 zmiennych stosując jako algorytm drzewo decyzyjne, otrzymano zbiór zmiennych  $X_{38}, X_{50}, X_{103}, X_{152}, X_{153}$ . Las losowy wskazał zbiór  $X_{118}, X_{149}, X_{152}, X_{153}$  oraz  $X_{161}$ . W sytuacji, gdy RFE przeprowadzono dla zbioru 16 zmiennych wyróżnionych przez regularyzację L2, Las losowy wskazał zmienne  $X_{118}, X_{129}, X_{149}, X_{152}$  i  $X_{153}$ , zaś Boosting gradientowy (GB) wyróżnił zmienne  $X_{87}, X_{103}, X_{118}, X_{152}, X_{153}$ . Zauważmy, że choć wyróżnione przez różne sposoby selekcji zbiory są różne, to niektóre zmienne powtarzają się, np.  $X_{103}, X_{152}, X_{153}$ . Powtarzające się zmienne oznaczają odpowiednio: liczbę dni do pierwszego zamknięcia (zapadalności) ACA na rachunku OVERDRAFT, wpływy /zaangażowanie kredytowe (%) - wartość średnia z ostatnich - 2 lat, wzrost sprzedaży (%) - wartość średnia z ostatnich - 2 lat<sup>3</sup>.

W tabeli 4 przedstawiono zbiory zmiennych uzyskane różnymi metodami, na których następnie przeprowadzono klasyfikację z wykorzystaniem wybranych metod uczenia maszynowego (LR, RF, GB oraz XGB). Zbiory w prawej kolumnie różnią się od zbiorów w kolumnie lewej dodaniem zmiennych nominalnych  $X_2$  i  $X_{11}$ .

<sup>3</sup> Dokładny opis zmiennych wykorzystanych w analizie jest dostępny u autorów.

Rysunek 1. Wykres zależności między rangami zmiennych w różnych modelach



Źródło: opracowanie własne

Tabela 4. Wyróżnione zbiory zmiennych użyte w analizach porównawczych

A	$X_{38}, X_{50}, X_{103}, X_{152}, X_{153}$	A1	$X_2, X_{11}, X_{38}, X_{50}, X_{103}, X_{152}, X_{153}$
B	$X_{118}, X_{149}, X_{152}, X_{153}, X_{161}$	B1	$X_2, X_{11}, X_{118}, X_{149}, X_{152}, X_{153}, X_{161}$
C	$X_{118}, X_{129}, X_{149}, X_{152}, X_{153}$	C1	$X_2, X_{11}, X_{118}, X_{129}, X_{149}, X_{152}, X_{153}$
D	$X_{87}, X_{103}, X_{118}, X_{152}, X_{153}$	D1	$X_2, X_{11}, X_{87}, X_{103}, X_{118}, X_{152}, X_{153}$
E	$X_{118}, X_{147}, X_{149}, X_{152}, X_{153}, X_{161}$	E1	$X_2, X_{11}, X_{118}, X_{147}, X_{149}, X_{152}, X_{153}, X_{161}$
F	$X_{39}, X_{50}, X_{74}, X_{103}, X_{125}, X_{152}$	F1	$X_2, X_{11}, X_{39}, X_{50}, X_{74}, X_{103}, X_{125}, X_{152}$
G	$X_{50}, X_{66}, X_{103}, X_{152}, X_{153}$	G1	$X_2, X_{11}, X_{50}, X_{66}, X_{103}, X_{152}, X_{153}$
H	$X_{38}, X_{97}, X_{143}, X_{145}, X_{147}, X_{150}, X_{152}, X_{153}$	H1	$X_2, X_{11}, X_{38}, X_{97}, X_{143}, X_{145}, X_{147}, X_{150}, X_{152}, X_{153}$

Źródło: opracowanie własne

W tabeli 5 przedstawiono miarę jakości klasyfikacji, trafność, która oznacza odsetek poprawnych klasyfikacji dokonywanych przez model LR, RF, GB lub XGB i wyróżnionych zbiorów zmiennych ciągłych. W tabeli 6 przedstawiono trafność wyznaczoną na zbiorach zmiennych uzupełnionych o zmienne nominalne  $X_2$  i  $X_{11}$ .

Tabela 5. Wyniki trafności modeli wraz z odchyleniami standardowymi

	A	B	C	D
LR	0,9116 (0,0312)	0,9512 (0,0100)	0,9267 (0,0097)	0,9021 (0,0103)
RF	0,9994 (0,0008)	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,0009)
GB	0,9990 (0,0009)	0,9990 (0,0009)	0,9990 (0,0009)	0,9992 (0,0009)
XGB	0,9990 (0,0009)	0,9991 (0,0009)	0,9992 (0,0009)	0,9988 (0,0012)
	E	F	G	H
LR	0,9531 (0,0099)	0,7813 (0,0364)	0,8810 (0,0097)	0,8467 (0,0129)
RF	0,9982 (0,0012)	0,9990 (0,0009)	0,9992 (0,0009)	0,9990 (0,0009)
GB	0,9990 (0,0009)	0,9992 (0,0009)	0,9994 (0,0008)	0,9990 (0,0009)
XGB	0,9992 (0,0009)	0,9990 (0,0009)	0,9992 (0,0009)	0,9992 (0,0009)

Źródło: opracowanie własne

Tabela 6. Wyniki trafności modeli wraz z odchyleniami standardowymi

	A1	B1	C1	D1
LR	0,9387 (0,0304)	0,9673 (0,0090)	0,9558 (0,009488)	0,9472 (0,0080)
RF	0,9990 (0,0009)	0,9994 (0,0008)	0,9990 (0,000948)	0,9992 (0,0009)
GB	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,000929)	0,9992 (0,0009)
XGB	0,9994 (0,0008)	0,9992 (0,0009)	0,9992 (0,000929)	0,9994 (0,0008)
	E1	F1	G1	H1
LR	0,9669 (0,0085)	0,7813 (0,0364)	0,9275 (0,0224)	0,8467 (0,0129)
RF	0,9988 (0,0009)	0,9996 (0,0007)	0,9994 (0,0008)	0,9992 (0,0009)
GB	0,9992 (0,0009)	0,9994 (0,0008)	0,9994 (0,0008)	0,9992 (0,0009)
XGB	0,9992 (0,0009)	0,9990 (0,0009)	0,9996 (0,0007)	0,9992 (0,0009)

Źródło: opracowanie własne

Otrzymane wyniki pokazują, że wybór zestawu zmiennych nie wpływa znacząco na jakość klasyfikacji mierzoną trafnością dla takich metod uczenia maszynowego jak Lasy losowe (RF), Boosting gradientowy (GB), czy algorytm XGB. Dla tych metod włączenie do zbioru zmiennych kategorycznych nie poprawiło jakości klasyfikacji, która jest bardzo wysoka. Dla Regresji logistycznej (LR) obserwujemy zależność jakości klasyfikacji od użytego zbioru zmiennych. Dodatkowo, dla wszystkich zestawów zmiennych z wyjątkiem F oraz H, włączenie zmiennych kategorycznych poprawiło znacząco jakość klasyfikacji dla Regresji logistycznej.

## PODSUMOWANIE

W podejściu statystycznym modele regresyjne w naturalny sposób pozwalają definiować ważność zmiennych poprzez wartości parametrów/wag (przy zastosowaniu transformacji standaryzującej) oraz wykorzystują statystyczną istotność pozwalającą wyeliminować parametry, których ważność jest bliska zeru. Dzięki temu interpretacja ważności jest intuicyjna i zgodna z fizyczną interpretacją teoretyczną. Coraz częściej zaczyna się wprowadzać do modelowania procedury regularyzacji, które są mniej oczywiste, ale pozwalają na bardziej elastyczne podejście do procesu selekcji zmiennych zastępując regułę 0/1 (jest/nie ma) regułą włączania/wyłączania częściowego. Warto zauważyć, że takie podejście ma na celu identyfikację związku przyczynowego będącego głównym przedmiotem badań. Poznanie i zrozumienie przyczyny pozwala na opracowanie skutecznego sposobu reagowania w oparciu o model.

W modelach algorytmicznych proces selekcji zmiennych zależy od wewnętrznej struktury modelu. I tak np. w przypadku Lasów losowych powiązany jest z funkcją zanieczyszczenia co powoduje, że interpretacja wyników zależy bardzo silnie od wewnętrznych algorytmów stosowanych w modelu i nie musi odpowiadać znaczeniu jakie intuicyjnie nadajemy pojęciu ważności zmiennej. Jeszcze silniej widać to w przypadku modeli sztucznych sieci neuronowych, w których ważność zmiennej powiązana jest ze ścieżką zmian wewnętrznych parametrów w procesie trenowania modelu.

Wyniki przeprowadzonych analiz dowodzą, że badając strukturę zależności między czynnikami przy pomocy modeli regresyjnych bardzo silnie uzależniamy jej interpretację od używanych narzędzi.

Należy podkreślić, że nie ma jednej optymalnej metody selekcji zmiennych. Każdy z przedstawionych sposobów wyróżnił inny zestaw. Pewne zmienne pojawiały się w każdym z nich. Niestety, korelacja (Kendalla) pomiędzy wynikami różnych metod nie pozwala na potwierdzenie, że ważność zmiennej jest cechą zewnętrzną przypisaną do niej. Widać potrzebę dokonania rewizji pojęcia „ważność” tak, aby nie była ona zależna od struktury modelu.

Wyniki prezentowane w niniejszej pracy wskazują, że powinniśmy zmienić punkt widzenia: zamiast opisywać ważność poszczególnych zmiennych należy interpretować grupy zmiennych czyli inaczej rating zmiennych. Oczywiście zmienia to sposób interpretacji i wykorzystania modeli, ale pozwala lepiej uchwycić związek między zmiennymi zależnymi i niezależnymi.

## BIBLIOGRAFIA

- Adler A. I., Painsky A. (2022) Feature Importance in Gradient Boosting Trees with Cross-Validation Feature Selection. *Entropy*, 24(5), 687. <https://doi.org/10.3390/e24050687>.

- Ben Jabeur S., Stef N., Carmona P. (2023) Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. *Comput Econ*, 61, 715-741. <https://doi.org/10.1007/s10614-021-10227-1>.
- Breiman L. (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199-215.
- De Sa C.R. (2019) Variance-Based Feature Importance in Neural Networks. [in:] Kralj Novak P., Šmuc T., Džeroski S. (eds) *Discovery Science, Lecture Notes in Computer Science*, 11828, Springer, Cham. [https://doi.org/10.1007/978-3-030-33778-0\\_24](https://doi.org/10.1007/978-3-030-33778-0_24).
- Engelmann B., Rauchmeier R. (2011) *The Basel II: Risk Parameters. Estimation, Validation, Stress Testing - with Applications to Loan Risk Management*. Springer Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-16114-8>.
- Elith J., Leathwick J. R. and Hastie T. (2008) A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology*, 77, 802-813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>.
- Gajowniczek K., Wu J., Gupta S., Bajaj C. (2022) HOFS: Higher Order Mutual Information Approximation for Feature Selection in R. *SoftwareX*, 19, 1-9. <https://doi.org/10.1016/j.softx.2022.101148>.
- Hastie T., Tibshirani R., Friedman J. (2008) *The Elements of Statistical Learning* (2nd ed.), Springer.
- Hastie T., Tibshirani R., Wainwright M. (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York Chapman & Hall/CRC. <https://doi.org/10.1201/b18401>.
- Hopf K., Sascha R. (2021) Filter Methods for Feature Selection in Supervised Machine Learning Applications - Review and Benchmark. arXiv preprint arXiv:2111.12140, 2021.
- Jia W., Sun M., Lian J. et al. (2022) Feature Dimensionality Reduction: A Review. *Complex Intell. Syst.*, 8, 2663-2693. <https://doi.org/10.1007/s40747-021-00637-x>.
- Kohavi R, John G. H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2), 273-324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
- Lal T. N., Chapelle O., Weston J., Elisseeff A. (2006) *Embedded Methods*. [in:] Guyon I., Nikravesh M., Gunn S., Zadeh L. A. (eds) *Feature Extraction. Studies in Fuzziness and Soft Computing*, 207, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-35488-8\\_6](https://doi.org/10.1007/978-3-540-35488-8_6).
- Li J., Cheng K., Wang S., Morstatter F., Trevino R. P., Tang J., Liu H. (2017) Feature Selection: A Data Perspective. *ACM Computing Surveys*, 50(6), Article 94, 1-45. <https://doi.org/10.1145/3136625>.
- Olden J., Joy M., Death R. (2004) An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks using Simulated Data. *Ecological Modelling*, 178(3-4), 389-397. <https://doi.org/10.1016/j.ecolmodel.2004.03.013>.
- Pudjihartono N., Fadason T., Kempa-Liehr A. W., O'Sullivan J. M. (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinform*, 2(927312). doi: 10.3389/fbinf.2022.927312.
- Sánchez-Marño N., Alonso-Betanzos A., Tombilla-Sanromán M. (2007) Filter Methods for Feature Selection – A Comparative Study. [in:] Yin H., Tino P., Corchado E., Byrne W., Yao X. (eds) *Intelligent Data Engineering and Automated Learning - IDEAL 2007*. IDEAL 2007. *Lecture Notes in Computer Science*, 4881, Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-77226-2\\_19](https://doi.org/10.1007/978-3-540-77226-2_19).

- Taylor J., Tibshirani R. J. (2015) Statistical Learning and Selective Inference. Proc Natl Acad Sci U S A, 112(25), 7629-34. doi: 10.1073/pnas.1507583112.
- Vergara J. R., Estévez P.A. (2014) A Review of Feature Selection Methods Based on Mutual Information. Neural Comput & Applic, 24, 175-186. <https://doi.org/10.1007/s00521-013-1368-0>.
- Zebari R., Abdulazeez A., Zeebaree D., Zebari D., Saeed J. (2020) A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), 56-70. <https://doi.org/10.38094/jastt1224>.

### VARIABLE SELECTION BY STATISTICAL AND MACHINE LEARNING METHODS. COMPARISON OF APPROACHES USING FINANCIAL DATA AS AN EXAMPLE

**Abstract:** In line with new international financial supervision directives (IFRS9), banks should look at a new set of analytical tools, such as machine learning. The introduction of these methods into banking practice requires reformulation of business goals, both in terms of the accuracy of predictions and the definition of risk factors. The article compares methods for selecting variables and assigning "importance" in statistical and algorithmic models. The calculations were carried out using the example of financial data classification. The effectiveness of various machine learning algorithms on selected sets of variables was compared. The results of the analyzes indicate the need to revise the concept of the "importance" of a variable so that it does not depend on the structure of the model.

**Keywords:** variable selection, machine learning, variable importance

**JEL classification:** C45, C52, C55

## SYSTEM UCZENIA GŁĘBOKIEGO DLA ELIMINACJI SZUMÓW Z WYKORZYSTANIEM DYWERGENCJI ALPHA

Ryszard Szupiluk  <https://orcid.org/0000-0002-4416-7131>

Kolegium Analiz Ekonomicznych  
Szkola Główna Handlowa w Warszawie  
e-mail: rszupi@sgh.waw.pl

**Streszczenie:** W niniejszym artykule przedstawimy system eliminacji szumów z wielowymiarowych szeregów czasowych. Elementami tego systemu będą bloki: dekompozycji PCA/EVD, grupowania oparty na funkcji dywergencji alpha, dekompozycji odwrotnej PCA oraz filtracji indywidualnej. W ramach etapu grupowania przedyskutowane będą typowe problemy tego zadania, motywujące do poszukiwania nowych rozwiązań. Zaprezentujemy także eksperymenty potwierdzające słuszność opracowanej koncepcji.

**Słowa kluczowe:** grupowanie, filtracja szumów, eliminacja szumów, funkcje dywergencji, finansowe szeregi czasowe, uczenie głębokie

**JEL classification:** C02, C50

### WPROWADZENIE

Eliminacja szumów i zakłóceń jest jednym z typowych zadań analizy danych. Motywowane to jest licznymi praktycznymi zastosowaniami jak: identyfikacja trendów ekonomicznych, separacja sygnałów medycznych czy rekonstrukcja sygnałów mowy. Stosowane są tu różne techniki bazujące m.in. na wygładzaniu średnimi ruchomymi, zastosowaniu transformacji Fouriera, czy rozkładzie szeregów na komponenty o określonych charakterystykach [Shiryayev 1999, Hamilton 1994, Vaseghi 1997, Richter et al. 2022].

W niniejszym opracowaniu przedstawimy system eliminacji szumów, który ze względu na swoją złożoność i charakter stosowanych metod, może być określony jako system uczenia głębokiego. Jednym z jego elementów jest analiza składowych głównych (PCA-ang. Principal Component Analysis), co można

<https://doi.org/10.22630/MIBE.2023.24.4.19>

bezpośrednio powiązać z dekompozycją względem składowych własnych (EVD – ang. Eigenvalue Value Decoposition), stąd będziemy stosować akronim PCA/EVD [Jolliffe 2002]. Prezentowana koncepcja eliminacji szumów zakłada podejście wielowymiarowe, w którym zestaw szeregów czasowych poddajemy dekompozycji PCA/EVD otrzymując ukryte komponenty. Zakładamy przy tym, że niektóre z nich są szumami, których identyfikacja a następnie eliminacja pozwoli na usunięcie wspólnych zakłóceń o niejako fizycznym charakterze, tj. takich gdzie występuje pewien obiektywny/odrębny składnik (lub składniki) identyczny w kilku szeregach. Jest to podejście nawiązujące do metod ślepej separacji sygnałów [Comon i Jutten 2010]. Kluczowym elementem jest ocena i wybór komponentów, które można uznać za szumy. Rozważymy tą kwestię jako problem grupowania.

Zagadnienie grupowania, określane także jako klasteryzacja lub segmentacja, należy do podstawowych zagadnień uczenia maszynowego [Mitchell 1997, Richter et al. 2022]. Mimo ideowej prostoty niektórych rozwiązań, a także możliwych wygodnych praktycznych przypadków, w ogólnym problemie grupowania występuje wiele ograniczeń i niejednoznaczności istotnie wpływających na jakość otrzymywanych rozwiązań. Stąd zastosowanie konkretnych algorytmów powinno uwzględniać specyfikę danego problemu, zwłaszcza charakterystykę danych [Everitt 2011].

W prezentowanym systemie zastosujemy algorytm grupowania oparty na funkcji dywergencji alpha [Amari 1985], ze względu na jej właściwości możliwe do wykorzystania w przypadku danych o charakterystyce szumowej lub posiadających istotny składnik losowy. Należy przy tym zaznaczyć, że ta metoda grupowania (lub oceny podobieństwa) może także stanowić samodzielne podejście w identyfikacji szumów.

## WIELOWARSTWOWY SYSTEM ELIMINACJI SZUMÓW

Założmy, że analizowane szeregi czasowe, generowane są w pewnym wspólnym środowisku, jak np. rynki finansowe, a w związku z tym, składają się zarówno z niepożądanych szumów, zakłóceń lub krótkoterminowych fluktuacji własnych, jak i użytecznych składników wspólnych, jak trendy i cykle średnio i długoterminowe. Naszym celem jest w pierwszym etapie wydzielenie z bazowych szeregów czasowych wspólnych komponentów szumowych a następnie ich eliminacja oraz dokonanie transformacji odwrotnej. Dla tak uzyskanych „oczyszczonych” szeregów czasowych możliwa jest przejście do etapu eliminacji szumów indywidualnych, jedną ze standardowych metod filtracji. System eliminacji szumów, przedstawiony na rysunku 1, składa się z następujących warstw.

**Wejście.** Zestaw szeregów czasowych  $\mathbf{x}_i$ ,  $i=1,\dots,N$ . zbieramy w jednej zmiennej wielowymiarowej  $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_N]$ .



**Warstwa 1.** Dekompozycja PCA/EVD. Zmienną  $\mathbf{X}$  poddajemy transformacji PCA otrzymując składowe główne  $\mathbf{Y}=[\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N]$ . Można przyjąć  $\mathbf{Y}=\mathbf{U}^T\mathbf{X}$ , gdzie  $\mathbf{U}$  jest macierzą kolumnowych wektorów własnych macierzy kowariancji  $\mathbf{X}$  otrzymanych z rozkładu EVD  $\mathbf{X}\mathbf{X}^T/N=\mathbf{U}\mathbf{S}\mathbf{U}^T$ , macierz  $\mathbf{S}$  oznacza diagonalną macierz wartości własnych rozkładu.

**Warstwa 2.** Grupowanie i identyfikacja szumu. Otrzymane składowe główne grupujemy, wydzielając grupę komponentów związanych z zaszumieniem lub zakłóceniami. Ten aspekt omówimy dokładniej w dalszej części artykułu.

**Warstwa 3.** Eliminacja szumów. Grupę komponentów szumowych eliminujemy. W praktyce procesowej oznacza to zastąpienie danego sygnału zerami lub przyjęcie w transformacji odwrotnej odpowiedniej kolumny zerowej.

**Warstwa 4.** Transformacja odwrotna. Dokonujemy transformacji odwrotnej do transformacji z warstwy pierwszej, otrzymując szeregi czasowe pozbawione komponentów szumowych.

**Warstwa 5.** Eliminacja szumów indywidualnych. Może się odbywać różnymi technikami, w tym zwykła średnią kroczącą.

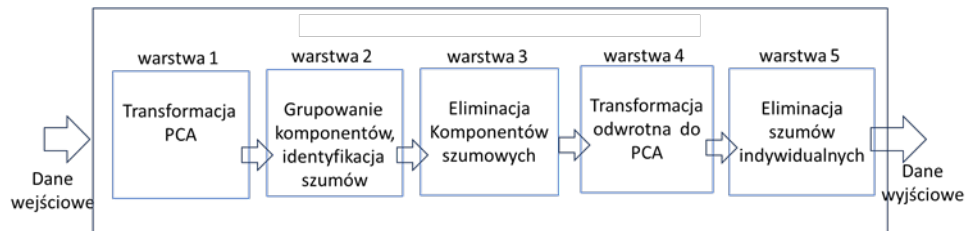
**Wyjście.** Otrzymujemy sygnały przefiltrowane.

Prezentowany układ, określa pewien ogólny schemat, który może być w szerokim zakresie modyfikowany lub dopasowywany do konkretnej sytuacji. W powyższej procedurze, w zasadzie jednym istotnym wymogiem/założeniem jest odwracalność macierzy transformacji. Przykładowo, w przypadku pracy z danymi typu Biga Data (m.in. duży wolumen, duża zmienność) wskazanym może być skorzystanie z adaptacyjnych algorytmów PCA [Oja 1992]. Należy przy tym zauważyć, iż w prezentowanym podejściu, bazując na EVD, nie musimy czynić żadnych dodatkowych założeń co do charakterystyki analizowanych szeregów bazowych, jak również odnośnie charakterystyki komponentów w nich zawartych. W naszym przypadku, przyjmujemy, że nie są dla nas istotne założenia związane z istnieniem statystyk drugiego rzędu, stacjonarnością lub ergodycznością sygnałów. Oczywiście można uwzględniać te kwestie w miarę potrzeb.

Zadania warstwy piątej, nie będą szerzej omawiane w niniejszej pracy, gdyż mogą być wykonane standardowymi technikami licznie ujętymi w literaturze [Therrien 1992, Oppenheim i Schafer 2014, Richter 2022].

Należy jednak zaznaczyć, iż eliminacja szumów indywidualnych powinna się odbywać po wcześniejszym usunięciu szumów wspólnych i zadania warstwy piątej nie powinny być wykonywane wcześniej. Zmiana tej kolejności oznaczać będzie konwolucyjne zmieszanie szumu wspólnego i indywidualnego danego sygnału z jego komponentami użytecznymi. Skuteczne wydzielenie takich komponentów jest zadaniem dekonwolucji wielowymiarowej, będącej nadal aktualnym wyzwaniem badawczym [Levin et al. 2009].

Rysunek 1. System uczenia głębokiego eliminacji szumów



Źródło: opracowanie własne

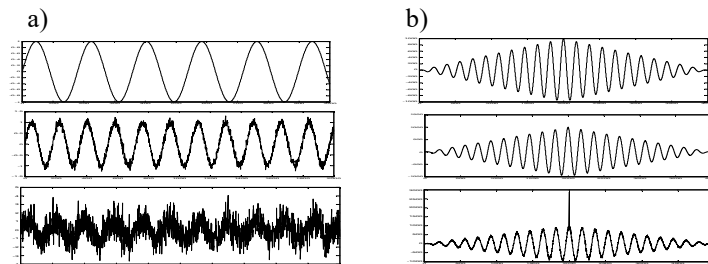
W przypadku zmiany postaci dekompozycji mogą się pojawić dodatkowe założenia. Niemniej, można zauważyć, że w instrumentalnej optyce uczenia maszynowego, istotność przyjętych założeń warunkowana jest specyfiką użycia danej metody w konkretnej sytuacji.

Jedną z modyfikacji może być przyjęcie innych dekompozycji wielowymiarowych jak: analiza składowych niezależnych (ICA – ang. Independent Component Analysis), nieujemna faktoryzacja macierzy (NMF – ang. Nonnegative Matrix Factorization) lub nieliniowe PCA [Comon i Jutten 2010]. Obecny wybór PCA/EVD motywowany jest doskonale zbadanymi jej właściwościami teoretycznymi oraz potwierdzonymi licznymi praktycznymi zastosowaniami.

## PROBLEM GRUPOWANIA DANYCH SZUMOWYCH

Zasadniczym elementem prezentowanej koncepcji jest ocena komponentów uzyskanych z PCA/EVD, czyli rozróżnienie między komponentami użytecznymi a szumowymi. W przypadku finansowych szeregów czasowych, takimi jak np. logarytmiczne stopy zwrotu, sytuacja może być o tyle kłopotliwa, że komponenty szumowe mogą mieć charakterystykę daleką od czystych białych szumów losowych [Shiryaev 1999]. W efekcie ocena na bazie funkcji autokorelacji może nie być zbyt użyteczna, funkcja ta nie jest też wysoce czuła na szum, co pokazuje rysunek 1, gdzie dopiero stosunkowo duży poziom szumu jest widoczny w postaci piku na funkcji autokorelacji ( $SNR \approx -2$  dB). Dlatego raczej będziemy oczekiwać, że komponenty zakłócające są w pewnym stopniu podobne do białych szumów, niż nimi faktycznie są. Stąd wyznaczenie komponentów szumowych potraktujemy jako zadanie przypisania do grupy o charakterze szumowym niż stricto spełniające warunki białego szumu.

Rysunek 2. Zaszumione sygnały a) i odpowiadające im funkcje autokorelacji b), gdzie zerowe opóźnienie odpowiada obserwacji środkowej.



Źródło: opracowanie własne

Zagadnienie grupowania można przedstawić jako identyfikację w posiadanym zbiorze: danych, obiektów, obserwacji lub wektorów do siebie podobnych lub sobie bliskich. Przynależność do danej grupy bazuje zwykle na wybranym matematycznym kryterium podobieństwa lub odległości. Można wyróżnić takie klasy metod grupowania jak: metody oparte na metrykach, grupowanie hierarchiczne lub metody oparte na rozkładach [Hennig et al. 2015, Richter et al. 2022]. Możliwe jest także grupowanie w oparciu o modele danych i ich parametry, a także grupowanie uwzględniające kontekst praktycznych zastosowań. W szerszym rozumieniu grupowanie można określić jako identyfikację wewnętrznej struktury wielowymiarowych danych [Maharaj et al. 2019, Meila 2003].

W każdym z wymienionych podejść występują algorytmy w różnych wariantach i lokalnych modyfikacjach. Jednak, mimo, że zasadniczą formułą implementacji oraz realizacji algorytmów grupowania jest uczenie maszynowe, gdzie użytkownik metody musi zwykle określić ile grup (klastrow) ma uzyskać, lub zaakceptować regułę to określającą, co można potraktować jako tzw. hiperparametr [Agrawal et al. 2005, Amorim i Hennig 2015]. Określanie pewnych cech odnoszących się do modelu, procesu uczenia czy otrzymanych wyników jest typowe w całokształcie metod uczenia maszynowego. Jednak w przypadku grupowania danych wielowymiarowych mamy z reguły znacznie ograniczone możliwości oceny słuszności przyjętych założeń lub sposobu parametryzacji algorytmów, a w konsekwencji, trudności weryfikacji czy kontroli uzyskanych rezultatów [Hüllermeier 2010]. Nie tylko różne metody, ale także te same, tylko inaczej sparametryzowane (np. k-średnich z różnymi metrykami lub początkowymi centrami) mogą prowadzić do całkowicie odmiennych wyników [Celebi et al. 2013].

W przypadku szeregów czasowych o charakterze szumowym lub ze znaczącym komponentem szumowym, problemem może być sposób oceny ich podobieństwa na bazie charakterystyk indywidualnych bądź wzajemnych relacji opartych na standardowych metrykach lub statystykach drugiego rzędu. Z tego

względu, zastosujemy funkcję dywergencji alpha jako metodę oceny podobieństwa, wykorzystując jej interesujące właściwości analityczne, w szczególności cechę asymetrii miary podobieństwa [Amari i Nagaoka 2000].

## FUNKCJA DYWERGENCJI ALPHA

Dywergencja  $D(\mathbf{y}||\mathbf{z})$ , to funkcja dwuargumentowa, określona na nieujemnych zmiennych  $z$  i  $y$ , która spełnia warunek  $D(\mathbf{y}||\mathbf{z}) \geq 0$ , gdzie  $D(\mathbf{y}||\mathbf{z}) = 0$  wtedy i tylko wtedy, gdy  $\mathbf{y} = \mathbf{z}$ . Dywergencja nie musi spełniać nierówności trójkąta  $D(\mathbf{y}||\mathbf{z}) \leq D(\mathbf{y}||\mathbf{x}) + D(\mathbf{x}||\mathbf{z})$ . Dywergencje są z reguły niesymetryczne  $D(\mathbf{y}||\mathbf{x}) \neq D(\mathbf{x}||\mathbf{y})$ . Dla niektórych dywergencji konieczny jest warunek sumowania się wartości zmiennych  $z$  i  $y$  do jedności. Dywergencje mogą być zdefiniowane dla wielkości ciągłych jak i dyskretnych. Obecnie funkcje dywergencji stosowane są do oceny podobieństwa (lub jego braku) między rozkładami oraz/lub nieujemnymi zmiennymi, wektorami, macierzami lub funkcjami [Amari 1985, Cichocki et al. 2009, Csiszar 1974].

Dywergencja, może być też przyjmowana lub interpretowana, jako miara quasi-odległości. Miary dywergencji, uzyskiwane mogą być zarówno na drodze aksjomatycznej, tzn. jako charakterystyki o pewnych interesujących właściwościach matematycznych [Csiszar 2008], jak i wynikać z badania praktycznych problemów [Cichocki et al. 2009].

Większość z dość licznej listy dywergencji można zaklasyfikować lub uznać za szczególne przypadki kilku zasadniczych rodzin. Dwie szerokie klasy stanowią dywergencje Ciszar oraz Bregman'a.

Dywergencje Ciszar, zwane także  $f$ -dywergencjami, wyrażają się ogólną formułą [Csiszar 1974]:

$$D_c(\mathbf{y} || \mathbf{z}) = \sum_{i=1}^N \left( z_i f \left( \frac{y_i}{z_i} \right) \right), \quad (1)$$

gdzie elementy wektorów  $\mathbf{x}$  oraz  $\mathbf{y}$  spełniają warunki  $x_i \geq 0, y_i \geq 0$  zaś  $f(u)$ ,  $u = \frac{y}{z}$  jest funkcją wypukłą w  $(0, \infty)$  oraz ciągłą w punkcie 0,  $N$  oznacza ilość elementów wektorów  $\mathbf{y}$  oraz  $\mathbf{z}$ . Przyjmuje się, że jeśli funkcja dywergencji (1) spełnia dodatkowe warunki, takie że,  $f(1) = 0$  oraz jest ona ściśle wypukła w 1, to może ona być miarą podobieństwa.

Dywergencje Bregmana stanowią szeroką klasę określaną przez [Bregman 1967, Dhillon i Tropp 2007]:

$$D_\phi(\mathbf{y} || \mathbf{z}) = \sum_{t=1}^N \left( \phi(y_t) - \phi(z_t) - \phi'(z_t)(y_t - z_t) \right), \quad (2)$$

gdzie  $\phi(u)$  jest ściśle wypukłą funkcją klasy  $C^1$  (wszędzie różniczkowalną) posiadającą ciągłą pochodną  $\phi'(u)$ .

W dalszych rozważaniach skoncentrujemy się na dywergencji alpha, która jest klasą dywergencji należąca zarówno do rodziny dywergencji Bregmana jak i f-dywergencji Ciszar'a [Amari 1985, Amari i Nagaoka 2000]. Dywergencja ma postać:

$$D_A^\alpha(\mathbf{y}||\mathbf{z}) = \sum_{i=1}^N \frac{y_i^\alpha z_i^{1-\alpha} + (\alpha - 1)z_i - \alpha y_i}{\alpha(\alpha - 1)}. \quad (3)$$

Dla parametru  $\alpha=2$  stanowi odległość Pearson'a, dla  $\alpha=0.5$  odległość Hellingera, dla  $\alpha=-1$  odległość chi-kwadrat Neyman'a. Dla  $\alpha \rightarrow 1$  otrzymuje się uogólnioną dywergencję Kullback'a-Leibler'a:  $D_A^{\alpha \rightarrow 1}(\mathbf{y}||\mathbf{z}) = D_{GKL}(\mathbf{y}||\mathbf{z})$  zaś dla  $\alpha \rightarrow 0$  otrzymuje się dualną uogólnioną dywergencję Kullback'a-Leibler'a:  $D_A^{\alpha \rightarrow 0}(\mathbf{y}||\mathbf{z}) = D_{DGKL}(\mathbf{y}||\mathbf{z})$  [Cichocki et al. 2009]. Alpha dywergencja Amarięgo może zostać otrzymana z dywergencji Ciszar'a za pomocą nieliniowości:

$$f(u) = \frac{u(u^{\alpha-1} - 1)}{\alpha^2 - \alpha} + \frac{1-u}{\alpha}. \quad (4)$$

Należy tu zaznaczyć, że dywergencja alpha nie wymaga sumowania wartości zmiennych do jedności. My jednak będziemy dokonywać wstępnego przetworzenia danych tak, aby ten warunek został spełniony.

## GRUPOWANIE Z FUNKCJĄ DYWERGENCJI ALPHA

W zadaniu grupowania z wykorzystaniem dywergencji alpha, wykorzystamy jej cechę niesymetryczności wraz z typową charakterystyką szumów losowych, jaką jest ich samopodobieństwo [Mandelbrot 1997]. Możemy oczekiwać, że te dwie właściwości skutkować będą otrzymywaniem relatywnie niedużych wartości symetrycznych dla „czystych” szumów  $D^\alpha(\mathbf{x}||\mathbf{y})=D^\alpha(\mathbf{y}||\mathbf{x})$ , i wyników niesymetrycznych  $D^\alpha(\mathbf{x}||\mathbf{y}) \neq D^\alpha(\mathbf{y}||\mathbf{x})$  o większych wartościach dla danych o mniejszej zawartości szumu. Cały algorytm grupowania można przedstawić następująco:

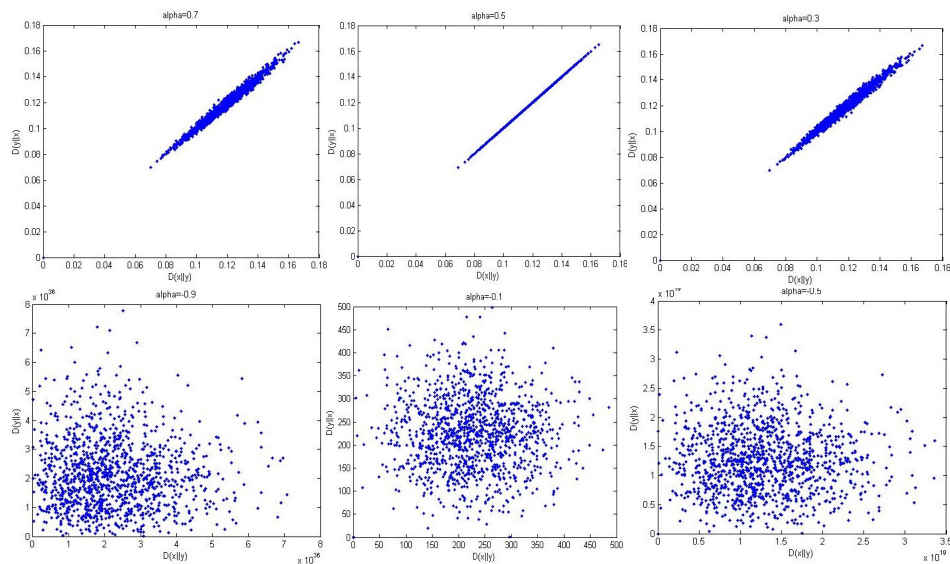
1. Dla komponentów  $\mathbf{Y}$  dokonaj wstępnego przygotowania danych, postaci (zapis Matlaba)  $\mathbf{y}_i = \mathbf{y}_i - \min(\mathbf{y}_i)$  oraz  $\mathbf{y}_i = \mathbf{y}_i / (\sum(\mathbf{y}_i) + \exp(-100))$ .
2. Dla wszystkich par komponentów otrzymanych z PCA oraz przyjętego  $\alpha$  obliczamy dywergencje  $D^\alpha(\mathbf{y}_i||\mathbf{y}_j)$  oraz  $D^\alpha(\mathbf{y}_j||\mathbf{y}_i)$ .
3. Znajdujemy takie pary komponentów, że  $D^\alpha(\mathbf{y}_i||\mathbf{y}_j) \approx D^\alpha(\mathbf{y}_j||\mathbf{y}_i) \approx 0$ .
4. Weryfikujemy otrzymane wyniki przy różnych wartościach  $\alpha$ .

Typowy zakres  $\alpha$  należy do przedziału  $(0,1)$ , zmiana tego parametru w tym zakresie nie powinna zasadniczo zmieniać wyników grupowania. Można także oprzeć się na uśrednionym pomiarze wartości dywergencji przy różnych wartościach  $\alpha$ .

## BADANIE PRAKTYCZNE

Obecnie przedstawimy powyższą procedurę na przykładzie symulowanych szeregów czasowych. Na początek sprawdzimy słuszność koncepcji oceny szumów (sygnałów losowych) dla wygenerowanych szeregów z rozkładu gaussowskiego ( $x=\text{randn}(50,1000)$  w zapisie Matlaba). Rysunek 3 przedstawia rozrzut punktów we współrzędnych  $D(y|x)$ ,  $D(x|y)$ . Celem badania jest test efektu symetryczności oraz niedużych wartości miar dywergencji dla sygnałów losowych. Badanie przeprowadzono dla parametrów  $\alpha$ : 0.7; 0.5, 0.3, -0.9, -0.1, -0.5. Jak widać, dla parametrów dodatnich uzyskujemy efekty zgodne z oczekiwaniami. Ujemne parametry  $\alpha$  wykazują w zasadzie całkowity brak zależności co do kierunku i wartości, w zakresie tych parametrów metoda nie działa. Jest to o tyle znamienne, że przy  $\alpha \rightarrow -1$ , alpha dywergencja zbiega do popularnej miary chi-kwadrat Neyman'a.

Rysunek 3. Wartości  $D(y|x)$  względem  $D(x|y)$  dla sygnałów losowych (szumy gaussowskie) dla parametrów  $\alpha$ : 0.7, 0.5, 0.3 w górnym wierszu oraz -0.9, -0.1, -0.5 w dolnym

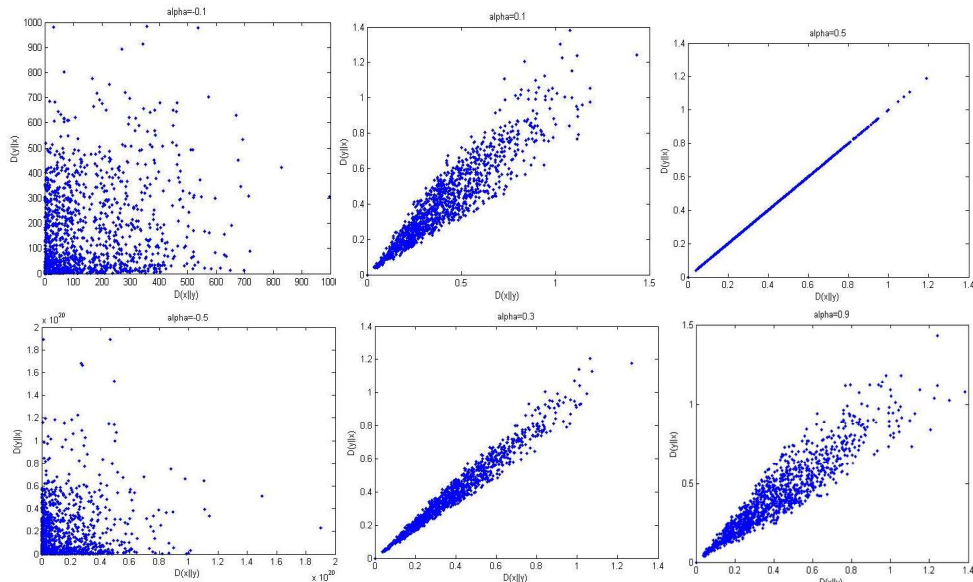


Źródło: opracowanie własne

Kolejne badanie przeprowadzimy dla sygnałów symulujących giełdowe szeregi czasowe, które wygenerowaliśmy jako  $x = \text{abs}(\text{cums}(\text{randn}(50,1000)))$  w zapisie Matlab.

Rysunek 4 przedstawia rozrzut punktów we współczesnych  $D(y||x)$ ,  $D(x||y)$  dla parametrów  $\alpha$ : -0.3, -0.1, 0.1, 0.5, 0.3, 0.9. Ponownie ujemne parametry  $\alpha$  wykazują brak zależności co do kierunku i wartości miar dywergencji  $\alpha$ , natomiast dla dodatnich  $\alpha$  mamy wyniki zgodne z oczekiwaniami. Widzimy zróżnicowanie wartości zarówno jeżeli chodzi o położenie wokół linii symetrii  $D(y||x) = D(x||y)$ , jak i zróżnicowanie względem przyjmowanych wielkości.

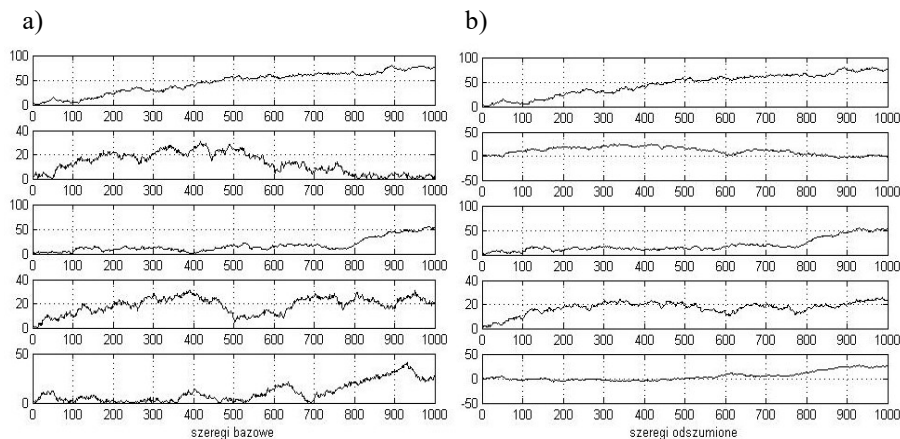
Rysunek 4. Wartości  $D(y||x)$  względem  $D(x||y)$  dla sygnałów losowych (szumy gaussowskie) dla parametrów  $\alpha$ : -0.1, 0.1, 0.5 w górnym wierszu oraz -0.3, 0.3, 0.9 w dolnym



Źródło: opracowanie własne

Te pary punktów, które leżą blisko punktu (0,0) oraz w bliskiej odległości od linii symetrii, możemy uznać za szumowe. Słuszność takiego podejścia prezentuje badanie wykonane na symulowanych finansowych szeregach czasowych, z których usunięto punkty uznane za szumowe. Jak widać, usunięcie tych komponentów pozwoliło na wygładzenie szeregów czasowych, przy zachowaniu głównych trendów. Oznacza to, że w istocie zostały usunięte szumy i krótkoterminowe fluktuacje.

Rysunek 5. Działanie systemu eliminacji szumów a) oryginalne bazowe szeregi czasowe, b) szeregi odszumione



Źródło: opracowanie własne

## PODSUMOWANIE

Przedstawiony system uczenia głębokiego do eliminacji szumów określa pewien ogólny schemat postępowania, którego poszczególne elementy, mogą być w szerokim stopniu modyfikowane. Dotyczy to zwłaszcza postaci dekompozycji wielowymiarowej. Wybór podejścia opartego na PCA/EVD, związany był, z jednej strony, z pożądanymi właściwościami matematycznymi tej metody, a z drugiej, z jej popularnością teoretyczną i praktyczną. Pozwoliło to poświęcić zasadniczą uwagę aspektowi identyfikacji komponentów szumowych, co wykonano w ramach problemu grupowania.

Koncepcja grupowania oparta na dywergencji alpha, znalazła potwierdzenie w przeprowadzonych eksperymentach. Metoda ta może być także rozważana jako samodzielne podejście analityczne, niezależne od prezentowanego systemu filtracji. Jednak wykorzystanie dywergencji alpha wydaje się zasadne do przypadków, w których występują szumy lub sygnały w jakimś stopniu o charakterze losowym. Dywergencja ta jest także stosunkowo obszerne opracowana teoretycznie. Niemniej, zastosowanie innych dywergencji, z których wiele jest mało zbadanych, otwiera szanse na uzyskanie nowych ciekawych wyników.

Ze względu na ograniczony charakter niniejszego opracowania pominieliśmy omawianie ostatniej warstwy naszego systemu jaką jest filtracja indywidualnych szeregów czasowych. Jest to jednak temat bardzo dobrze opisany i na chwilę obecną przyjmujemy, że stosujemy standardowe metody dla tego problemu. Nie wykluczamy jednak, wprowadzenia w przyszłości pewnych nowych rozwiązań także w tej warstwie.



## BIBLIOGRAFIA

- Agrawal R., Gehrke J., Gunopulos D., Raghavan P. (2005) Automatic Subspace Clustering of High Dimensional Data, *Data Mining and Knowledge Discovery*, 11, 5-33.
- Amari S. (2012) *Differential-Geometrical Methods in Statistics*. Springer Verlag.
- Amari S. and Nagaoka H. (2000) *Methods of Information Geometry*. Oxford University Press, New York.
- Amorim R. and Hennig C. (2015) Recovering the Number of Clusters in Data Sets with Noise Features using Feature Rescaling Factors. *Information Sciences*, 324, 126-145.
- Bregman L (1967) The Relaxation Method of Finding a Common Point of Convex Sets and its Application to the Solution of Problems in Convex Programming. *Comp. Math. Phys., USSR*, 7, 200-217.
- Celebi M., Kingravi H., Vela P. (2013) A Comparative Study of Efficient Initialization Methods for the k-Means Clustering Algorithm. *Expert Systems with Applications*, 40(1), 200-210.
- Cichocki A., Zdunek R., Phan A.-H., Amari S. (2009) *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis*. John Wiley.
- Comon P., Jutten Ch. (2010) *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press.
- Csiszar I. (1974) Information Measures: A Critical Survey. [In:] *Transactions of the 7th Prague Conference*, 83-86.
- Csiszar I. (2008) Axiomatic Characterizations of Information Measures. *Entropy*, 10, 261-273.
- Dhillon I. S., Tropp J. A. (2007) Matrix Nearness Problems with Bregman Divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4), 1120-1146.
- Everitt B. (2011) *Cluster Analysis*. Chichester, West Sussex, U.K, Wiley.
- Hamilton J. D. (1994) *Time Series Analysis*. Princeton, Princeton University Press.
- Hennig Ch., Meila M., Murtagh F. (2015) *Handbook of Cluster Analysis*. CRC Press.
- Hüllermeier E. (2010) Uncertainty in Clustering and Classification. [In:] Deshpande A., Hunter A. (eds) *Scalable Uncertainty Management. SUM 2010. Lecture Notes in Computer Science()*, 6379, 16-19. Springer, Berlin, Heidelberg.  
[https://doi.org/10.1007/978-3-642-15951-0\\_6](https://doi.org/10.1007/978-3-642-15951-0_6)
- Jolliffe I. T. (2002) *Principal Component Analysis*. Springer Verlag.
- Levin A., Weiss Y., Durand F. and Freeman W.T. (2009) Understanding and Evaluating Blind Deconvolution Algorithms. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, Miami, 20-25 June 2009, 1964-1971.
- Mandelbrot B. B. (1997) Multifractals and 1/f Noise. *Wild Self-Affinity in Physics (1963–1976)*, Springer-Verlag. <https://doi.org/10.1007/978-1-4612-2150-0>
- Maharaj E., D'Urso P., Caiado J., (2019) *Time Series Clustering and Classification*, CRC Press, Boca Raton.
- Meila M. (2003) Comparing Clusterings by the Variation of Information, *Learning Theory and Kernel Machines. Lecture Notes in Computer Science*, 2777, 173-187.
- Mitchell T. (1997) *Machine Learning*. McGraw-Hill, Boston.

- Oja E. (1992) Principal Components, Minor Components and Linear Neural Networks. *Neural Networks*, 5, 927-935.
- Oppenheim R., Schaffer A. (2014) *Discrete - Time Signal Processing*, Pearson India.
- Richter A., Sheuli P., Kępuska V., Silaghi M. (2022) *Signal Processing and Machine Learning with Applications*. Springer Cham.
- Shiryaev A. N. (1999) *Essentials of Stochastic Finance: Facts, Models, Theory*. Singapore, World Scientific.
- Therrien, C. W. (1992) *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall, New Jersey.
- Vaseghi S. V. (1997) *Advanced Signal Processing and Digital Noise Reduction*. John Wiley and Sons, Chichester.

### DEEP LEARNING SYSTEM FOR NOISE ELIMINATION USING ALPHA DIVERGENCE

**Abstract:** In this article, we present a system for noise eliminating from multivariate time series. The system contain the following layers: PCA/EVD decomposition, grouping based on the Alpha divergence function, reverse decomposition to PCA, and individual filtration. During the grouping stage, typical problems of this task will be discussed, motivating the search for new solutions. We will also present experiments confirming the validity of the developed concept.

**Keywords:** clustering, noise filtering, divergence functions, financial time series, deep learning

**JEL classification:** C02, C50

## CZY ZMIANA STANDINGU FINANSOWEGO SPÓŁEK PUBLICZNYCH WPŁYWA NA ZMIANY ODSETKA Kobiet W ORGANACH STATUTOWYCH?

**Dorota Witkowska**  <https://orcid.org/0000-0002-2810-6654>

**Krzysztof Kompa**  <https://orcid.org/0000-0001-9538-9589>

College of Business and Economics

University of Johannesburg, South Africa

e-mail: mariaw@uj.ac.za; kkompa@uj.ac.za

**Streszczenie:** Zwiększenie udziału kobiet w zarządzaniu jest "gorącym" tematem dyskusji. Celem pracy jest sprawdzenie czy zmiana standingu finansowego spółek przekłada się na zmianę frakcji kobiet w ich kierownictwie. Badaniem objęto 90 spółek nieprzerwanie notowanych na GPW w Warszawie w latach 2010-2019. Ocenę sytuacji finansowej spółek przeprowadzono za pomocą syntetycznych miar wektorowych, obliczonych dla danych finansowych raportowanych na koniec każdego roku. Badania nie wykazały występowania zależności potwierdzającej, że wraz ze zmianą sytuacji finansowej spółek ulega zmianie odsetek kobiet w kierownictwie.

**Słowa kluczowe:** nierówności implikowane płcią, kobiety w organach statutowych, spółki publiczne, kobiety a wyniki finansowe, wielowymiarowe metody porównawcze

**JEL classification:** L25, G38, M14

### WSTĘP

Z danych European Institute for Gender Equality (EIGE)<sup>1</sup> wynika, że w największych spółkach publicznych w krajach Unii Europejskiej UE-28, udział kobiet w kierownictwie spółek wynosił w drugiej połowie 2004 roku zaledwie 9% i wzrósł do 11,9% w 2010 roku, a 10 lat później już do 30%. Oznacza to, że o ile do 2010 roku roczne przyrosty odsetka kobiet w managementie wynosiły średnio 4,8%,

---

<sup>1</sup> [https://eige.europa.eu/gender-statistics/dgs/indicator/wmidm\\_bus\\_bus\\_wmid\\_comp\\_compbm/metadata](https://eige.europa.eu/gender-statistics/dgs/indicator/wmidm_bus_bus_wmid_comp_compbm/metadata)

o tyle w ostatnim dziesięcioleciu roczne przyrosty wyniosły 9,7% rocznie. Według tego samego źródła danych, kobiety na stanowiskach kierowniczych największych spółek publicznych w Polsce stanowiły w 2004 r. 9,1%, w 2010 r. 11,6, a w 2020 r. 22,8%. Oznacza to, że do 2010 r. odsetek kobiet wśród kadry kierowniczej przyrastał średnio o 4,1%, a w ostatnim dziesięcioleciu - o 7%. Zatem biorąc pod uwagę przyrosty od 2004 r., to wyniosły one średnio 5,9% rocznie dla Polski i 7,8% w UE-28. Zacznie bardziej dynamiczne przyrosty w ostatnim 10-leciu związane są z osiągnięciem przez kobiety najwyższych stanowisk – w przypadku najwyższego stanowiska (*president*) średnioroczne przyrosty wyniosły 7,1% w UE-28 i aż 21,4% w Polsce, z kolei wśród CEOów wyniosły one odpowiednio 14,5% i 9,8% rocznie w latach 2012 - 2020.

Analizując dane EIGE zauważa się, że w 2010 roku tylko w 5 państwach EU reprezentacja kobiet w kierownictwie spółek publicznych przekroczyła 20%, a najwyższy odsetek obserwowano w Szwecji (26,4%). Natomiast w roku 2020 odsetek powyżej 20% dotyczył już 18 państw, z Francją (45,1%) na czele zestawienia. Wśród członków UE z najmniej liczną reprezentacją kobiet w menadżerencie znalazły się Luksemburg (2,4%), Węgry (4%) i Grecja (4,5%) w 2010 roku, a dziesięć lat później Estonia (8,8%), Malta (9,9%) i Węgry (9,9%). Widać zatem ogromne zmiany w dostępie kobiet do stanowisk kierowniczych, czego przyczyn – przynajmniej w przypadku kilku krajów – należy dopatrywać się w prawn-administracyjnej promocji równości, tj. wprowadzeniu kwot udziału kobiet w obsadzie gremiów kierowniczych.

Zmiany proporcji (obu płci) obserwowane w przypadku najwyższych stanowisk są nawet bardziej spektakularne, bowiem najwyższy odsetek kobiet wśród CEO, wyniósł 10% w 2012 roku (Malta, Rumunia i Słowenia) i 20% w 2020 r. (Słowacja). Z kolei najwięcej kobiet wśród *president* odnotowano w 2010 r. w Belgii (13,3%), a w 2020 r. w Polsce (36,8%). Przy czym, w przypadku obu grup stanowisk różnice są w wielu krajach bardzo duże biorąc pod uwagę stan początkowy, czasem wynoszący zero. Jednakże należy zauważyć, że nie we wszystkich państwach Unii Europejskiej nastąpiła poprawa sytuacji.

Warto także odnotować, że na najwyższym stanowisku (*president*) w 2020 r. aż w 9 krajach (tj. w  $\frac{1}{3}$  wszystkich członków UE) nie ma żadnej kobiety, a w 2010 r. – w 10, z czego w 4 krajach sytuacja się nie zmieniła. Z kolei analizując odsetek kobiet wśród CEOów, zauważa się brak kobiet w 16 krajach w 2012 r. i tylko w 4 państwach w 2020 r. Zaiściałej sytuacji nie można przy tym wiązać z pandemią, bowiem w drugim półroczu 2019 r. liczba krajów z zerowym odsetkiem kobiet wśród CEOów wyniosła 3 i 9 – dla najwyższych stanowisk. Zapewnienie równości kobiet i mężczyzn w dostępie do stanowisk kierowniczych w gospodarce stało się jednym z priorytetów Unii Europejskiej, czego wyrazem jest przyjęta przez Parlament Europejski dyrektywa o parytecie płci w kierownictwie spółek giełdowych<sup>2</sup>.

<sup>2</sup> Dyrektywa uchwalona 22.11.2022 r. określa tzw. parytety, które powinny zostać osiągnięte do 2026 r. W uzasadnieniu tej decyzji twierdzi się, że wprowadzenie parytetu płci

W ciągu ostatniego dziesięciolecia udział kobiet piastujących najwyższe stanowiska w zarządach i radach nadzorczych spółek notowanych na GPW w Warszawie zmienił się nieznacznie (odpowiednio wzrósł o 1,4% i 3,4%). Zasadniczo natomiast zwiększył się odsetek kobiet wśród wiceprezesów zarządów (o 47%) i zastępców przewodniczącego rady nadzorczej (o 16,7%). Odnotowano natomiast 10% spadek udziału kobiet – członkiń rad nadzorczych, przy 11% wzroście frakcji członkiń zarządów spółek. Jednocześnie nastąpił ponad trzykrotny wzrost udziału kobiet w zarządach największych spółek giełdowych. Natomiast w przypadku spółek średnich i małych odnotowano spadek odsetka kobiet w zarządach odpowiednio o 1,7% i 7%<sup>3</sup>.

Powstaje zatem potrzeba analizy czy obserwowane zmiany odsetka kobiet w organach statutowych spółek obecnych polskim rynku kapitałowym są powiązane z wynikami finansowymi tych spółek. W konsekwencji celem prezentowanej pracy jest odpowiedź na pytanie czy sytuacja finansowa spółek wpływa na odsetek kobiet w ich kierownictwie, czego nie rozważa się w szeroko rozumianej literaturze przedmiotu.

## PRZEGLĄD LITERATURY

Różnorodność w managementcie i potrzeba zwiększenia frakcji kobiet w gremiach zarządczych, są "gorącym" tematem w polityce, ekonomii i badaniach naukowych. Zwiększenie udziału kobiet w podejmowaniu decyzji na wszystkich szczeblach zarządzania ma silne uzasadnienia teoretyczne [Post i Byron 2015], wywodzące się z teorii agencji [Hillman i Dalziel 2003] i [Pfeffer i Salancik 1978], teorii tożsamości społecznej [Ashforth i Mael 1989], czy teorii kategoryzacji społecznej [Tajfel 1981]. A teoria wyższych eszelonów [Hambrick i Mason 1984] (por. też. [Hambrick 2007]), czy teoria masy krytycznej [Tremblay 2006] dobrze uzasadniają potrzebę szeroko rozumianej heterogeniczności rad nadzorczych. O ile w debacie publicznej różnorodności te nie są pojmowane jednakowo, to argumenty za dywersyfikacją, chociaż zazwyczaj w sposób nieświadomy, wywodzą się wprost z w/w teorii. Obejmują one szerokie spektrum poglądów, poczynając od twierdzeń, że obecność kobiet w managementcie wpływa na podniesienie efektywności finansowej przedsiębiorstw, ich reputacji i ratingu w ocenie społecznej odpowiedzialności, po spostrzeżenia, że różnorodność ma przeciwdziałać

---

w organach kierowniczych zaowocuje lepszym zarządzaniem, co przełoży się na większą rentowność biznesu i lepszą kondycję finansową spółek. Teza ta nie znajduje jednak jednoznacznego rozstrzygnięcia w literaturze przedmiotu, która wskazuje zarówno na pozytywne, jak i negatywne zależności pomiędzy obydwoma zjawiskami.

<sup>3</sup> Warto przy tym odnotować, że wg Notoria Serwis (stan na koniec 2020 r.) w 20 spółkach z kompozycji indeksu WIG20 jedynie 4 kobiety były na stanowiskach prezesa zarządu i 6 – na stanowiskach przewodniczącej rady nadzorczej, co zasadniczo odróżnia te dane od danych dystrybuowanych przez EIGE, podważając wiarygodność tych ostatnich.

monopolowi władzy jednakowo wykształconych, posiadających podobne - wypracowane w wyniku zbliżonych co do przebiegu karier zawodowych - poglądy menadżerów lub wręcz poszukiwanie sposobu na „przemycanie” żeńskiej części rodziny do kierownictwa spółek. Istnieją też zwolennicy „sprawiedliwości” społecznej, którzy kontestują ograniczenia w awansach kobiet (nazywane „szklanym sufitem” *glass ceiling*, por. [Cotter i in. 2001] oraz postulują konieczność zwalczaniu tego zjawiska, także w drodze przymusu prawnego i ustanowienia parytetu płci (kwot), co może skutkować pozornymi działaniami równościowymi.

Argumenty, którymi w literaturze przedmiotu uzasadnia się postulat zwiększonego udziału kobiet w organach kierowniczych dobrze pasują do teorii wymienionych na wstępie. Argumenty te dzieli się na dwie grupy: społeczne i ekonomiczne – por. badania [Campbell i Mingueza-Vera 2010], [Bohdanowicz 2010] czy [Post i Byron 2015]. Wśród argumentów społecznych wymienić należy przede wszystkim dezyderat równego traktowania, tj. zakaz dyskryminacji członków organów statutowych ze względu na płeć. Spośród argumentów ekonomicznych przemawiających za dywersyfikacją organów statutowych wymienia się poprawę wyników finansowych implikowaną zwiększeniem się frakcji kobiet w organach kierowniczych. Jednakże w literaturze pojawiają się również opinie, według których udział kobiet we władzach nie ma istotnego wpływu na wyniki przedsiębiorstw lub wykazuje się, że jest on wręcz negatywny. Wskazuje się również, że zwiększona liczba kobiet w kierownictwie spółek jest dobrze postrzegana przez inwestorów, chociaż zdarzają się sytuacje, gdy w obliczu słabnących, czy wręcz złych wyników przedsiębiorstw powołuje się kobiety na kierownicze stanowiska wyłącznie po to, aby obarczyć je winą za niepowodzenie organizacji – jest to zjawisko zwane szklanym klifem, *glass cliff*, opisane przez Ryana i Haslama [2005, 2007].

Warto przy tym odnotować, że z badaniami wpływu udziału kobiet w gremiach zarządczych na wyniki przedsiębiorstwa nie jest związana żadna uniwersalna metodyka badawcza. Dominują analizy ekonometryczne i korelacyjne dotyczące zależności zmian frakcji kobiet i pojedynczych wskaźników finansowych. Podejście proponowane w naszych badaniach [Kompa 2018, Kompa 2019; Kompa i Witkowska 2023] oraz [Witkowska, i in. 2019, s. 190-205] ma uzupełnić lukę metodyczną o wielowymiarowy pomiar efektywności przedsiębiorstw, w szczególności przez zastosowanie syntetycznych miar wektorowych, konstruowanych z wykorzystaniem danych z raportów rocznych.

Postawienie pytania, czy poprawa sytuacji firmy wpływa na zwiększenie udziału kobiet w kierownictwie w zasadzie nie jest poruszane w literaturze poza wspomnianym wcześniej zjawiskiem szklanego klifu, polegającym na mianowaniu kobiet na wysokie stanowiska w sytuacji wysokiego ryzyka lub kiedy sytuacja kierowanej instytucji lub przedsiębiorstwa jest zła [Ryan i Haslam 2005, 2007]. Można powiedzieć, że w takim przypadku niezależnie od rezultatu starań managerki, organizacja zawsze wygrywa, bowiem albo jej umiejętności wyprowadzą firmę z kłopotów, albo - jeśli się to nie uda - za niepowodzenie można będzie winić kobietę. W związku z powyższym nasze badania stanowią próbę wypełnienia luki

w literaturze, bowiem pozwolą wykazać czy sytuacja finansowa spółek publicznych w Polsce wpływa na zróżnicowanie organów statutowych pod względem płci.

## ORGANIZACJA BADANIA

Celem badania jest sprawdzenie czy zmiana standingu finansowego spółek przyczynia się do zmiany odsetka kobiet w ich kierownictwie. W analizach uwzględniono tylko te spółki, które 30.12.2019 r. tworzyły portfele indeksów: WIG20, mWIG40, sWIG80 i jednocześnie były notowane przez cały okres objęty badaniem, tak więc analizowane spółki tworzą próbę longitunalną. Dodatkowym warunkiem kwalifikacji spółek do analiz była dostępność pełnej informacji o składach ich zarządów i rad nadzorczych oraz o ich sprawozdaniach rocznych. Ostatecznie do badania wybrano 90 spółek – 73 spółki niefinansowe i 17 spółek finansowych, w tym 8 banków<sup>4</sup>, które stanowiły 97,8% spółek notowanych pod koniec grudnia 2010 roku, co jednak stanowi 64,54% spółek notowanych na tym rynku 10 lat później<sup>5</sup>.

W celu sprawdzenia czy zmiana sytuacji spółek skutkuje zmianą dywersyfikacji organów statutowych badano, czy występują zależności pomiędzy odsetkiem kobiet w organach statutowych, a standingiem finansowym firm, mierzonym syntetyczną miarą efektywności za pomocą współczynnika korelacji liniowej Pearsona i testu jego istotności. W analizach przyjęto założenie, że można uznać dane zjawisko jako przyczynę innego o ile przyczyna jest opóźniona w stosunku do pomiaru zjawiska. Innymi słowy przyjęto, że występuje określone następstwo zdarzeń polegające na wcześniejszej obserwacji sytuacji finansowej spółek w stosunku do pomiaru odsetka kobiet w managementcie. W szczególności można się spodziewać, że wykorzystując współczynnik korelacji liniowej Pearsona dodatnia statystycznie istotna korelacja obu zjawisk wskaże, że polepszanie się sytuacji spółek generuje zwiększoną obecność kobiet w managementcie. Z kolei ujemna korelacja może wskazywać na omawiane wcześniej zjawisko szklanego klifu, które oznacza, że wraz z pogarszaniem się sytuacji finansowej spółek powołuje się kobiety na kierownicze stanowiska.

Pierwszy etap badań polegał na wyznaczeniu frakcji kobiet w organach statutowych dla wszystkich spółek na dzień 30.06. w latach 2011-2019. W drugim etapie skonstruowano syntetyczne wektorowe mierniki efektywności finansowej dla wszystkich spółek (VSME) korzystając z wybranych wskaźników finansowych, wyznaczonych na podstawie raportów rocznych (stan na koniec roku) za lata 2010–2018. Mierniki obliczano oddzielnie dla spółek niefinansowych, finansowych i banków oraz wszystkich spółek, na zróżnicowanych zestawach wskaźników (wzory i wynik jak w pracy [Kompa, Witkowska 2022]). Trzeci etap polegał na wyznaczeniu

---

<sup>4</sup> Pierwotna baza danych została utworzona przez Błaszczyk [2021] na podstawie zasobów hurtowni danych Nortoria Serwis.

<sup>5</sup> Wprawdzie pod koniec 2010 r. na parkiecie notowano 400, a w 2019 – 449 spółek, to w analizowanym dziesięcioleciu wprowadzono do obrotu 186 nowych i wycofano 137 spółek.

współczynników korelacji liniowej Pearsona oraz testowaniu jego istotności. Analizie poddano zależności między zagregowaną oceną spółek i frakcją kobiet w organach kierowniczych dla opóźnionych wartości przyrostów mierników syntetycznych tj.  $dVSME_{it}$  w celu sprawdzenia czy wyniki finansowe w okresie  $t$  odwzorowywane są w zmianie odsetka kobiet  $dFK_{it}$  w okresach  $t+1$  i  $t+2$ . Przy czym w analizach uwzględniono przyrosty bezwzględne i względne:

$$d_1VSME_{it} = VSME_{it} - VSME_{i(t-1)}. \quad (1)$$

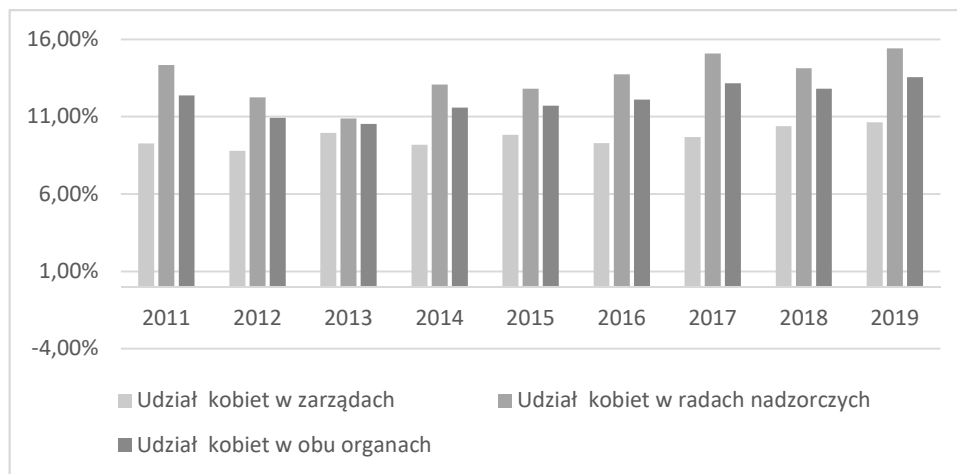
$$d_2VSME_{it} = \frac{VSME_{it} - VS_{i(t-1)}}{VSME_{i(t-1)}}. \quad (2)$$

gdzie:  $VSME_{it}$  – wartość wektorowego miernika efektywności finansowej obliczonego dla  $i$ -tej spółki w roku  $t$ , a  $d_1VSME_{it}$ ,  $d_2VSME_{it}$  – przyrosty mierników. Analogicznie wyznaczono przyrosty odsetka kobiet  $dFK_t$ .

## WYNIKI EMPIRYCZNE

W pierwszym etapie badania wyznaczono odsetek kobiet na poszczególnych stanowiskach w organach statutowych analizowanych spółek. Jak widać na rysunku 1, w latach 2011-2019 nastąpił znaczący wzrost odsetka kobiet zarówno w zarządach badanych spółek (15%), jak i w radach nadzorczych (8%). Natomiast obecność kobiet na najwyższych stanowiskach tj. prezesów zarządów i przewodniczących rad nadzorczych jest w dalszym ciągu rzadkością i waha się od 1 do 6 prezesek i od 5 do 10 przewodniczących w 90 badanych spółkach w kolejnych latach.

Rysunek 1. Obecność kobiet w organach kierowniczych badanych spółek notowanych na GPW w kolejnych latach



Źródło: obliczenia własne

Z uwagi na to, że w badaniach wykorzystano zarówno przyrosty względne, jak i bezwzględne frakcji kobiet, to oprócz badań przeprowadzonych na pełnym



zbiorze 90 spółek, utworzono dodatkowo 3 skrócone próby badawcze zawierające spółki, w których: (a) nie było zerowych udziałów kobiet w latach 2010-2019 (34 spółki), (b) nie było zerowych udziałów kobiet w latach 2012-2018 (44 spółki), (c) udział kobiet w całym badanym okresie wyniósł 20% i więcej (13 spółek).

Tabela 1. Współczynniki korelacji liniowej Pearsona dla wszystkich kategorii spółek

Okres pomiaru zmiennych								
$dFK_{it}$	$d_1VSMD_{it}$		$d_1VSMD_{it}$		$d_1VSMD_{it}$		$d_1VSMD_{it}$	
$t$	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$
	Wszystkie spółki		Spółki niefinansowe		Banki		Spółki finansowe reszta	
2012	-0,0315		0,0363		-0,0945		0,0214	
2013	0,0304	0,0523	0,0673	0,1317	-0,1323	0,1775	0,1215	-0,4276
2014	0,1568	-0,1066	0,1897	-0,1241	-0,5663	0,6040	-0,2479	0,4881
2015	<b>0,1772</b>	-0,0891	<b>0,2070</b>	-0,1559	-0,1156	-0,1644	0,1231	0,0389
2016	0,0759	-0,1566	-0,0365	-0,0511	-0,0866	0,3537	-0,2383	0,1159
2017	-0,0133	-0,0277	-0,0034	0,0492	0,2494	-0,4254	-0,1668	-0,0380
2018	0,1118	0,0171	-0,0003	0,0406	0,1023	0,1737	-0,4551	-0,3070
2019	0,0146	-0,1094	0,0357	-0,0648	<b>0,8100</b>	0,2658	-0,1765	0,2364
2012-19	0,0457	-0,0318	0,0438	-0,0051	-0,0811	0,2436	-0,0277	0,0213

Pogrubioną czcionką oznaczono odrzucenie hipotezy zerowej na poziomie istotności 0,05

Źródło: opracowanie własne

Tabela 2. Współczynniki korelacji liniowej Pearsona obliczone dla spółek bez zerowych frakcji kobiet w kierownictwie i spółek z przynajmniej ich 20% udziałem

Okres pomiaru zmiennych						
$dFK_{it}$	$d_1VSMD_{it}$		$d_2VSMD_{it}$		$d_1VSMD_{it}$	
$t$	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$
	Spółki bez zerowych frakcji kobiet				Spółki z 20% fracją kobiet	
2012	0,1239		-0,1882		<b>-0,4734</b>	
2013	-0,0778	<b>-0,3676</b>	-0,0501	0,0315	-0,0800	<b>0,4293</b>
2014	0,0444	-0,0314	-0,1402	-0,1235	-0,0307	0,1367
2015	0,0571	0,0456	-0,1062	0,2471	0,2737	-0,1563
2016	-0,1724	0,1612	<b>-0,3215</b>	-0,0793	0,1623	-0,2622
2017	-0,1970	0,2247	0,1452	0,1872	0,1017	0,1938
2018	0,1098	0,0297	0,0312	-0,0235	-0,0237	0,0066
2019	0,0666	-0,0645	-0,0101	-0,0635	-0,2752	<b>0,5939</b>
2012-2019	0,0208	0,0068	-0,0629	0,0341	-0,0188	0,0891

Pogrubioną czcionką oznaczono odrzucenie hipotezy zerowej na poziomie istotności 0,05

Źródło: opracowanie własne

Przyjęto, że powołanie na stanowisko kierownicze odbywa się 6 lub 18 miesięcy po zakończeniu roku, czyli – z uwagi na to, że pomiar frakcji kobiet dokonywany jest w czerwcu – ocena sytuacji finansowej spółki dokonywana jest w roku poprzednim (pomiar grudniowy) lub dwa lata wcześniej. Wyniki analiz przedstawiono w tabelach 1-3.

Tabela 3. Współczynniki korelacji liniowej Pearsona obliczone dla spółek bez zerowych frakcji kobiet w kierownictwie w latach 2012-2018

$dFK_{it}$	Okres pomiaru zmiennych			
	$d_1VSMD_{it}$		$d_2VSMD_{it}$	
	$t-1$	$t-2$	$t-1$	$t-2$
2014	0,0386		0,0340	
2015	0,1011	-0,0927	-0,0770	-0,1441
2016	-0,0902	0,0786	<b>-0,2602</b>	-0,1012
2017	-0,0768	0,0648	0,1493	0,0866
2018	0,0876	0,0025	-0,1195	0,0151
2019	0,1608	-0,0659	0,0185	-0,1858
2014-2019	0,0341	0,0249	0,0362	-0,0562

Pogrubioną czcionką oznaczono odrzucenie hipotezy zerowej na poziomie istotności 0,05

Źródło: opracowanie własne

Z prezentowanych w tabelach wyników widać, że nie ma istotnych korelacji między wynikami finansowymi spółek, a zmianą liczby kobiet w organach statutowych – na 182 obliczone współczynniki, statystycznie istotne okazało się jedynie 8 z nich (4,4%). W połowie były to wartości dodatnie, podobnie jak połowa z nich dotyczyła odsetka kobiet obserwowanych w 6-miesięcznym okresie po ocenie standingu spółek. W przypadku analiz istotności tych relacji dla banków i pozostałych finansowych spółek stwierdzono występowanie istotnej statystycznie zależności dla znacznie wyższych wartości współczynników korelacji Pearsona z powodu małej liczby spółek analizowanych w kolejnych latach.

## PODSUMOWANIE

Udział kobiet we władzach przedsiębiorstw i instytucji jest niezadawalający, w kontekście przesłanek polityk równościowych, w większości krajów. Dąży się zatem do znacznego zwiększenia roli kobiet w gremiach decyzyjnych. Wskazuje się przy tym różne przyczyny, dla których istnieje potrzeba zapewnienia różnorodności w kierownictwie. Część z nich ma charakter ekonomiczny i przywołuje wyniki badań naukowych, inne robią wrażenie deklaracji politycznych. Niemniej jednak sam problem jest ważny i dostrzegany w wielu krajach, wpisując się w ogólną tematykę nierówności na rynku pracy i *corporate governance*. Można przy tym wyróżnić kraje, w których sytuację można uznać za zadowalającą, albo przynajmniej

stanowiącą dobry punkt wyjścia do dalszych działań oraz kraje, gdzie sytuacja jest zdecydowanie zła.

Analiza zatrudnienia kobiet w spółkach notowanych na Gieldzie Papierów Wartościowych w Warszawie wskazuje, że frakcja kobiet w tych podmiotach systematycznie rośnie [Kompa i in. 2015]. Zasadne jest zatem pytanie, czy podobna tendencja jest widoczna w odniesieniu do obsady stanowisk kierowniczych tych spółek? Innymi słowy, czy wzrostowi zatrudnienia w spółkach publicznych towarzyszy proporcjonalny wzrost frakcji kobiet w zarządach i radach nadzorczych?

Z przeprowadzonych badań wynika, że w Polsce odsetek kobiet w organach statutowych systematycznie wzrasta, chociaż w 2019 roku stanowiły one jedynie 10,6% członków zarządów i 15,4% członków rad nadzorczych analizowanych spółek. Jednocześnie udział kobiet na najwyższych stanowiskach w zarządach i radach nadzorczych spółek notowanych na GPW w Warszawie jest nieznaczący i w badanym okresie nie zauważa się istotnych zmian.

Próba sprawdzenia czy zmiana sytuacji finansowej spółek powoduje istotną zmianę frakcji kobiet w managementie pokazała, że taka zależność na polskim kapitałowym nie występuje, ponieważ odrzucenie hipotezy zerowej było możliwe tylko w 5,5% badanych przypadków, w tym odnotowano po 4 przypadki korelacji dodatniej i ujemnej.

Nasze badanie wpisuje się w światowe trendy badawcze, uzupełniając je o nowatorskie podejście do pomiaru standingu przedsiębiorstw i sprawdzenie czy wyniki finansowe przekładają się na zwiększenie obecności kobiet w kierownictwie. Jesteśmy przy tym świadomi, że nasze wyniki mogą być zakłócone niedostateczną długością próby badawczej i znikomym udziałem frakcji kobiet w managementie.

## BIBLIOGRAFIA

- Ashforth B. E., Mael F. (1989) Social Identity Theory and the Organization. *Academy of Management Review*, 14, 20-39.
- Błaszczak A. (2021) Obecność kobiet w organach statutowych spółek publicznych w Polsce, *Zeszyty Naukowe Polskiego Towarzystwa Ekonomicznego w Zielonej Górze*, 15, 117-134.
- Bohdanowicz L. (2010) Kobiety w radach nadzorczych spółek publicznych – pokonywanie barier. *Przegląd Corporate Governance*, 3(23), 23-33.
- Campbell K., Minguez Vera A. (2009) Female Board Appointments and Firm Valuation: Short and Long-Term Effects. *Journal of Management & Governance*, 14(1), 37-59.
- Cotter D. A., Hermsen J. M., Ovadia S., Vanneman R. (2001) The Glass Ceiling Effect. *Social Forces*, 80(2), 655-81. DOI: 10.1353/soft2001.0091.
- Hambrick D. C. (2007) Upper Echelons Theory: An Update. *Academy of Management Review*, 32, 334-343.
- Hambrick D. C., Mason P. A. (1984) Upper Echelons: The Organization as a Reflection of its Top Managers. *Academy of Management Review*, 9, 193-206.

- Hillman A. J., Dalziel T. (2003) Boards of Directors and Firm Performance: Integrating Agency and Resource Dependence Perspectives. *The Academy of Management Review*, 28(3), 383-396.
- Kompa K. (2018) Parytet płci w kierownictwie spółek giełdowych a zmiany w sytuacji finansowej firm w latach 2010-2013. *Przedsiębiorczość i Zarządzanie*, 19(3/II).
- Kompa K. (2019) Zmiany w kierownictwie spółek giełdowych a zmiany ich sytuacji finansowej. [w:] Śliwiński A. (red.) *Zarządzanie w warunkach ryzyka*. Oficyna Wydawnicza SGH, Warszawa.
- Kompa K., Witkowska D. (2022) Wektorowa syntetyczna miara efektywności firmy dla polskich spółek publicznych, *MIBE*, 23(2), 47-58.
- Kompa K., Witkowska D. (2023) Impact of Women's Representation in Top Management on Performance of WSE-Listed Companies. [w:] Dietl M., Zarzecki D. (red.) *Understanding the Polish Capital Market from Emerging to Developed*. Routledge, Taylor & Francis Group, London and New York, 280-298.
- Kompa K., Witkowska D., Jarosz B. (2015) Kobiety w kierownictwie spółek giełdowych. [w:] Tarczyński W. (red.) *Rynek kapitałowy. Skuteczne Inwestowanie, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego*, 257-268.
- Pfeffer J., Salancik G. R. (1978) *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper Row.
- Post C., Byron K. (2015). Women on Boards and Firm Financial Performance: A Meta-Analysis. *Academy of Management Journal*, 58(5), 1546-1571.
- Ryan M.K., Haslam S.A. (2005) The Glass Cliff: Evidence that Women are Over-Represented in Precarious Leadership Positions. *British Journal of Management*, 16(2), 81-90. <https://doi.org/10.1111/j.1467-8551.2005.00433.x>.
- Ryan M.K., Haslam S.A. (2007) The Glass Cliff: Exploring the Dynamics Surrounding the Appointment of Women to Precarious Leadership Positions. *The Academy of Management Review*, 32(2), 549-572.
- Tajfel H. (1981) *Human Groups and Social Categories*. Cambridge: Cambridge University Press.
- Tremblay M. (2006) The Substantive Representation of Women and PR: Some Reflections on the Role of Surrogate Representation and Critical Mass. *Politics & Gender*, 2(4), 502-511.
- Witkowska D., Kompa K., Matuszewska-Janica A. (2019) *Sytuacja kobiet na rynku pracy: wybrane aspekty*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

#### **DO CHANGES OF THE FINANCIAL STANDING OF PUBLIC COMPANIES INFLUENCE CHANGES IN THE FRACTION OF WOMEN IN STATUTORY BODIES?**

**Abstract:** Increasing the proportion of women in management is a "hot" topic of discussions. The purpose of the study is to find out if a change in the financial standing of companies translates into a change in the fraction of women in their management. The study covered 90 companies continuously listed on the Warsaw Stock Exchange in 2010-2019. The financial standing of

the companies was assessed using synthetic vector measures calculated for financial data reported at the end of each year. The research did not show the existence of a relationship confirming that the percentage of women in management changes with the change in the financial situation of the companies.

**Keywords:** gender inequality, women in statutory bodies, public companies, women and financial performance, multivariate comparative methods

**JEL classification:** L25, G38, M14

## STATYSTYCZNA WERYFIKACJA „DLC” TEKKEN 7

**Robert Woźniak**  <https://orcid.org/0009-0009-6020-5247>

Wydział Zastosowań Informatyki i Matematyki  
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie  
e-mail: robo.wozniak@gmail.com

**Streszczenie:** Głównym celem pracy jest weryfikacja czy postacie dodatkowo płatne mają przewagę nad postaciami startowymi w grach akcji. Cel pracy zrealizowano przeprowadzając statystyczną analizę zwycięstw w grze TEKKEN 7 dla wszystkich 50 postaci (wojowników) występujących w grze. Do wyodrębnienia jednorodnych grup postaci wykorzystano metody *k*-średnich, grupowania hierarchicznego oraz model based clustering. Dodatkowo za pomocą metody TOPSIS został zbudowany ranking postaci.

**Słowa kluczowe:** TEKKEN 7, wielowymiarowa analiza statystyczna, metoda *k*-średnich, klastrowanie hierarchiczne, model based clustering, TOPSIS

**JEL classification:** C38, M210

### SŁOWNIK POJĘĆ

1. „Fighting games”- bijatyka, rodzaj gier wideo skupiającej się na walce dwóch wojowników, którzy dysponują różnymi stylami walki.
2. „Data frame” – właściwości ruchu, każdy ruch ma indywidualne parametry, gdy uderzy, jest zablokowany, czy uniknięty w jakikolwiek inny sposób.
3. „Fighter” – wojownik, postać, charakter – dostępny w grze, istnieje podział na: „startowy” oraz „DLC”.
4. „DLC” – postać dodatkowa, dostępna odpłatnie. Ma na celu zwiększyć zainteresowanie grą, przedłużyć jej cykl życia oraz zwiększyć zysk z sprzedaży gry. Nie można jej (postaci) uruchomić bez zakupu gry.

<https://doi.org/10.22630/MIBE.2023.24.4.21>

5. Patch – aktualizacja usprawniająca działanie gry.
6. Kanibalizm – zjawisko, które występuje podczas wprowadzania gier do sprzedaży bardzo bliskim odstępem czasowym. Zainteresowany gracz, mimo, że chce zagrać w obie gry, musi wybrać tylko jedną ponieważ fizycznie nie jest w stanie grać w dwie gry.
7. Cykl życia gry – popularność gry na rynku, mierzona w godzinach spędzania graczy w grze.
8. „TIER LIST” – ranking postaci uszeregowanych od najlepszej do najgorszej. Postacie są umieszczone w grupach oznaczony często od liter „S” do „E”. Istnieje możliwość zmiany oznaczeń wedle twórcy.
9. Sklepy cyfrowe – usługi elektroniczne oferujące dystrybucje gier i rzeczy z nimi związanych oraz monitorowanie cyklu życia gry, śledzenie błędów oraz miejsce zamieszczania opinii graczy.
10. „VS”- versus, w badaniu będzie oznaczać liczbę wygranych meczy do wybranego przeciwnika.

### **Uwaga**

Wymienione pojęcia nie są opracowaniem mechaniki gry. Są wprowadzeniem do artykułu w celu łatwiejszego zrozumienia przeprowadzonego badania, nie jest to także żaden przewodnik do gry ani poradnik do mechaniki.

### **WSTĘP**

Od początku powstania „Fighting Games” gracze sondują, które postacie są najlepsze i czym wyróżniają się na tle innych. Użytkownicy gier sprawdzają wybrane cechy każdej z postaci takie jak: liczba wygranych turniejów przez daną postać, statystyki wygranych, łatwość obsługi postaci („Move set, Data frame”). Gracze dyskutują na temat każdej postaci a konsekwencją tych dyskusji są listy rankingowe - „TIERLIST”. Rankingi te służą do wstępnego rozeznania potencjalnego przeciwnika – postaci z którą gracz chce się zmierzyć. W tabelach 1. i 2. zaprezentowano przykładowe „TIER LIST”.

Tabela 1. Przykładowa TIER LIST TEKKEN 3

TIER	Postaci
S:	(True)Ogre
A:	Jin, Xiaoyu, Heihachi, Law
B:	Lei, Nina, Julia, Yoshimits, Anna, Mokujin
C:	Eddy, Paul, King, Hwoarang, Gun Jack
D:	Bryan, Kuma/Panda
E:	Dr. B, Gon

Źródło: opracowanie własne na podstawie danych zaczerpniętych z strony internetowej: [wiki.supercombo.gg](http://wiki.supercombo.gg) (dostęp z dnia 04.05.2022)

Tabela 2. Przykładowa TIER LIST TEKKEN 6

TIER	Postaci
S:	Nina Williams, Lee Chaolan, Alisa Bosconovitch, Armor King, Yoshimitsu, Devil Jin, Kazuya Mishima, Bryan Fury, Jin Kazama, Lili Rochefort
A:	Raven, King, Wang Jinrei, Zafina, Ling Xiaoyu, Jack 6, Paul Phoenix, Hwoarang, Sergei Dragunov, Heihachi Mishima, Marshall Law, Asuka Kazama
B:	Steve Fox, Miguel Caballero Rojo, Devil Kazuya, Baek Doo San, Eddy Gordo, Ganryu, Lei Wulong, Anna Williams, Leo Kliesen
C:	Lars Alexandersson, Julia Chang, Marshall Law, Feng Wei, Panda, Wang Jinrei, Roger Jr

Źródło: opracowanie własne na podstawie danych zaczerpniętych z strony internetowej: [wiki.supercombo.gg](http://wiki.supercombo.gg) (dostęp z dnia 04.05.2022)

Wraz z postępującym rozwojem technologicznym, powstawaniem baz danych, łatwiejszym dostępem do Internetu, bankowości elektronicznej, lepszych podzespołów komputerowych pojawiły się sklepy cyfrowe – platforma zarówno dla producentów jak i graczy. W tym miejscu producenci mogą dystrybuować swoje gry, monitorować cykle życia tych produktów oraz analizować reakcje społeczności graczy. Gracze z kolei mają ułatwiony dostęp do interesujących ich produktów, mogą oceniać gry oraz dyskutować na ich temat.

Po powstaniu sklepów cyfrowych przyszła moda na patch oraz promocję „DLC” w gatunku bijatyk. Bardzo często deweloperzy rozszerzają grę o dodatkowe postacie (DLC) ale są one sprzedawane po premierze gry. Wprowadzenie dodatkowych postaci spowodowało uaktualnienie TIER LIST. Gracze obserwujący



zmiany w rankingach analizowali dlaczego wprowadzenie jednej dodatkowej postaci zmienia układ rankingu oraz zaburza ustalony (poprzedni) porządek.

Głównym celem pracy jest weryfikacja, czy postacie dodatkowo płatne mają przewagę nad postaciami startowymi. Innymi słowy celem badania jest odpowiedź na następujące pytania:

1. Czy „DLC” jest dobrym zabiegiem marketingowym?
2. Czy „DLC” lepiej wypadają na tle postaci standardowych?
3. Czy można określić jakiegokolwiek powiązanie pomiędzy „DLC” ?

## BADANIA LITERATUROWE

W wielu artykułach przeprowadzono analizy statystyczne w zakresie różnych gier komputerowych. W publikacji pod tytułem “Positive and Negative effects of Game Analytics in the Game Design” [Powell 2016] zbadano mocne i słabe strony budowy gier metodą “Grounded Theory”. W pracy pokazano, że za pomocą tej techniki można analizować zainteresowanie graczy wybranymi elementami gier, takimi jak dodatkowe darmowe dodatki („DLC”) czy nowymi darmowymi (dodatkowymi) poziomami gier.

Celem badania “Strategic Timing of Entry: Evidence from Video Game” [Engelstätter, Ward 2013] była statystyczna weryfikacja strategii wejścia produktu na rynek. Stwierdzono, że odpowiedni termin wprowadzenia gry na rynek minimalizuje efekt “kanibalizmu gier wideo”. Podobne informacje można odnaleźć w pracy “Statistical analysis of Game Behaviour” [Brodzki, Fleischer, Sarnik 2010]. Wnioski zaprezentowane potwierdzają występowanie “kanibalizmu” gier. Dodatkowo w tym artykule stwierdzono bardzo niską korelację pomiędzy recenzjami zawodowych recenzentów na stronach internetowych a cyklem życia gry. Przedmiotem badania zaprezentowanym w artykule “Positioning Game Review as a Crucial Element of Game” [Tong 2021] był wpływ recenzji użytkowników na sprzedaż gry. W konkluzji zaznaczono, że dłuższy okres życia gry nie wpływa na sprzedaż. Deweloperzy powinni skupić się na mechanizmach ułatwiających rozgrywkę nowym graczom, niż skupiać się na obecnych użytkownikach.

## METODYKA BADAŃ

### **Metoda k-średnich:**

Metoda k-średnich to technika klastrowania, która grupuje dane na podstawie ich podobieństwa. Głównym celem jest podzielenie zbioru danych na k klastrów, gdzie każdy klaster zawiera punkty danych podobne do siebie nawzajem. Dokładniejszy opis można znaleźć np. [Kassambara 2017].

**Metoda grupowania hierarchicznego:**

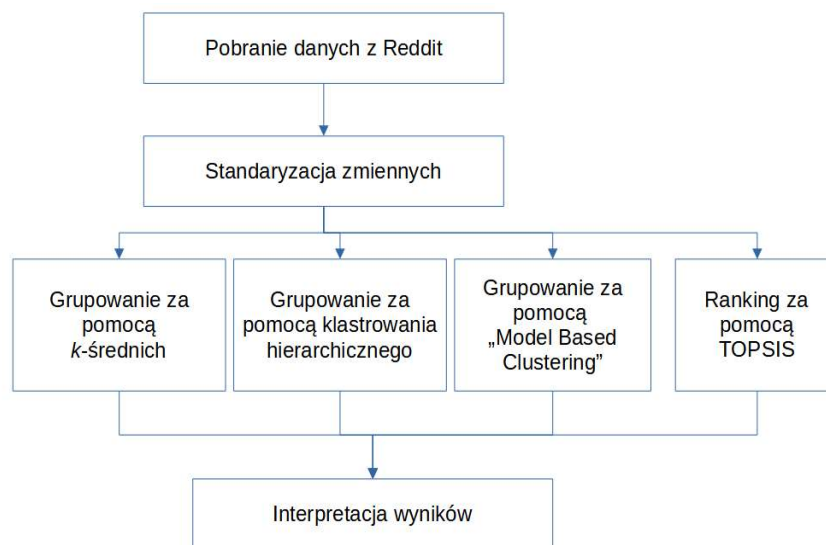
Do podziału analizowanego zbioru danych na grupy jest wykorzystywana metoda deglomeracyjna. Wynikiem hierarchicznego grupowania jest zestaw zagnieżdżonych klastrów, zwykle prezentowanych jako dendrogram - wielopoziomowe drzewo, w której klastry na jednym poziomie są połączone, tworząc większe klastry z jednego poziomu na drugi. Pozwala określić poziom, na jakim drzewo ma zostać przycięte, aby wygenerować odpowiednią liczbę klastrów. Dokładniejszy opis można znaleźć np. [Johnson, Wichern 2008].

**Model based clustering:**

Model based clustering to technika klastrowania, która opiera się na założeniu, że dane pochodzą z pewnego modelu statystycznego. W przeciwieństwie do tradycyjnych metod, takich jak  $k$ -średnich, w których klastry są identyfikowane na podstawie odległości między punktami danych, model-based clustering zakłada, że dane w klastrze podlegają pewnemu rozkładowi statystycznemu. Dokładniejszy opis można znaleźć np. [Bouveyron, Celeux, Murphy, Raftery 2019].

Szczegółowy schemat badania zaprezentowano na rysunku 1.

Rysunek 1. Poszczególne etapy prowadzenia badania.



Źródło: opracowanie własne

## MATERIAŁ EMPIRYCZNY

Do badania wykorzystano dane ze strony Reddit [Reddit 2020], zawierające informacje typu „VS”. Przykładowy fragment danych zaprezentowano w tabeli 3.

Tabela 3. Dane wykorzystane do analizy empirycznej (fragment badania)

	VS vs. Akuma	VS vs. Alisa	VS vs. Anna	VS vs. Armor King
Akuma		47	48	46
Alisa	53		50	52
Anna dlc	52	50		53

Źródło: opracowanie własne na podstawie danych pochodzących ze strony: [www.reddit.com](http://www.reddit.com)

W grze TEKKEN 7 dostępnych jest 50 postaci w tym 11 „DLC”. Zgromadzone informacje z Reddita zawierają wyniki rywalizacji pomiędzy dwoma różnymi wojownikami. Przykładowo postać „Alisa” ma 50 zwycięstw przeciwko postaci „Anna”. Analogicznie się postępuje w przypadku innych wojowników. Puste pola zostały zastąpione najczęstszą występującą wartością w danej kolumnie, z powodu, że bezsensowne jest badanie zwycięstw tej samej postaci do niej samej.

## WYNIKI BADAŃ EMPIRYCZNYCH

W tabeli 4 i tabeli 5 zaprezentowano odpowiednio wyniki uzyskane z metod klastrowania hierarchicznego oraz metody  $k$ -średnich. Za pomocą obu metod uzyskano tą samą liczbę klastrów, o bardzo podobnej strukturze. Innymi słowy, wyniki uzyskane dla obu metod są bardzo zbliżone – występują jedyne nieliczne przesunięcia pomiędzy grupami. W obydwu przypadkach można zauważyć odrębne grupy złożone z postaci dodatkowych (grupa 1. i 2.), grupę w której występuje silna dominacja „DLC” nad startowymi (grupa 3.) oraz pozostałych grup złożonych z podstawowych wojowników.

Tabela 4. Wyniki klastrowania hierarchicznego

Klaster	Nazwa Postaci
1	Leroy dlc, Ganryu dlc, Zafina dlc
2	Anna dlc, Kuma, Eddy, Jack-7, Eliza dlc, Panda, Gigas, Geese dlc, Marduk dlc, Lucky Chloe, Negan dlc, Claudio, Bob, Alisa
3	Noctis dlc, Paul Phoenix, Akuma, Armor King dlc, Katarina, Lili, Julia dlc, Nina, Lars
4	Feng, Heichachi, Shaheen, Hwoarong, Miguel, King, Bryan, Dargunov, Josie, Leo, Asuka, Lee, Master Raven, Lei dlc

Klaster	Nazwa Postaci
5	Xiaoyu Ling, Ssteve Fox, Yoshimitsu, Kazumi, Kazuya, Devil Jin ,Jin

Źródło: opracowanie własne

Tabela 5. Wyniki klastrowania metodą  $k$ -średniej

Klaster	Postać
1	Leroy dlc, Ganryu dlc , Zafina dlc
2	Anna dlc, Kuma, Eddy, Jack-7, Eliza dlc, Panda, Gigas, Geese dlc, Marduk dlc
3	Noctis dlc, Paul Phoenix, Akuma, Armor King dlc, Katarina, Lili, Julia dlc, Nina, Lars, Feng, Heichachi, Shaheen, Hwoarong, Miguel, Josie, Lee, Master Raven, Lei dlc
4	Jin, Yoshimitsu, Steve Fox
5	Xiaoyu Ling, Kazumi, Kazuya, Devil Jin, Asuka, King, Bryan, Law, Leo, Dragunov

Źródło: opracowanie własne

W tabeli 6 zaprezentowano wyniki grupowania za pomocą model based clustering. Można zauważyć w niej mniej klastrów – dwa pierwsze mają wyraźną przewagę „DLC” a pozostałych głównie klastrach występują głównie postacie standardowe.

Tabela 6. Wyniki klastrowania model based clustering

Klaster	Postać
1	Fakhumram dlc, Kunimitsu dlc, Negan dlc, Alisa, Claudio, Marduk dlc, Leroy dlc, Gigas, Ganryu dlc, Anna dlc, Lucky Chloe, Zafina dlc, Panda, Kuma, Geese dlc
2	Eddy, Eliza dlc, Jack 7, Feng, Bob, Lars, Nina, Lili, Julia dlc, Katarina, Nina Paul Pheonix, Armor King dlc, Noctis dlc
3	Master Raven, Lei dlc, Heichachi Shaheen, Miguel, Bryan, King, Hworang, Josie, Dragunov
4	XiaYu Ling, Steve Fox, Asuka, Yoshimitsu, Devil Jin, Law, Kazumi, Kazuya, Jin, Akuma

Źródło: opracowanie własne

Ostatnią częścią badania była budowa rankingu za pomocą metody TOPSIS. Fragment rankingu zaprezentowano w tabeli 7. Pierwsze cztery miejsca zajmują postacie „DLC”. W częściowym zaprezentowanym rankingu można znaleźć 7 z 11 dostępnych płatnych postaci.

Tabela 7. Wyniki TOPSIS (wyniki częściowe)

Ranking	Nazwa postaci	Wynik	Ranking	Nazwa Postaci	Wynik
1	Leroy dlc	0.85745	8	Kuma	0.60897
2	Ganryu dlc	0.77017	9	Marduk dlc	0.58975
3	Zafina dlc	0.71553	10	Lucky Chloe	0.58838
4	Anna dlc	0.66565	11	Negan dlc	0.58505
5	Claudio	0.65896	12	Panda	0.57936
6	Alisa	0.65681	13	Gigas	0.57113
7	Geese dlc	0.61706	14	Eliza dlc	0.56353

Źródło: opracowanie własne

## PODSUMOWANIE I WNIOSKI

W pracy zaprezentowano wyniki grupowania postaci gry TEKKEN 7 za pomocą trzech metod  $k$ -średnich, grupowania hierarchicznego oraz model based clustering. Za pomocą dwóch pierwszych uzyskano trzy skupienia zdominowane przez „DLC”. Dodatkowo można zauważyć wyraźny podział pomiędzy „DLC” a postaciami standardowymi. Wyniki uzyskane za pomocą model based clustering potwierdzają poprzedni podział. W rankingu zbudowanym metodą TOPSIS można zauważyć, że postacie „DLC” pod względem liczby zwycięstw są lepsze od postaci standardowych.

Po przeprowadzonych analizach można sformułować następujące odpowiedzi na pytania badawcze:

1. „DLC” jest dobrym zabiegiem marketingowym, odnoszą dużą liczbę zwycięstw, są pożądanymi postaciami dla graczy zatem generują dodatkowe zyski dla producenta.
2. Na podstawie przeprowadzonych analiz można zauważyć wyraźną przewagę „DLC” nad wojownikami standardowymi
3. „DLC” znajdują się w osobnych klastrach zatem można tylko przypuszczać o związku pomiędzy nimi. Jednak badanie nie jest wyjaśnieniem całkowitych

różnic pomiędzy indywidualnymi wojownikami, niezbędne jest także wykorzystanie innych metod, na przykład metod sieci, aby zanalizować przewagę na poziomie jednostkowym.

Podsumowując: za pomocą wykorzystanych metod uzyskano odpowiedzi na postawione pytania. Stwierdzono, że istnieje różnica pomiędzy „DLC” a postaciami standardowymi. Płatne postacie są lepsze pod względem liczby zwycięstw, będą więc chętniej wybierane przez graczy. W konsekwencji deweloper będzie rozwijał tą dodatkową opcję. Taki zabieg marketingowy potencjalnie przyniesie większe zyski.

Prezentowane badanie jest załącznikiem do dalszych badań w dziedzinie „Fighting games” oraz wyłonienia działań dewelopera jak i zachowań graczy.

## BIBLIOGRAFIA

- Bouveyron C., Celeux G., Murphy T., Raftery A. (2019) *Model-Based Clustering and Classification for Data Science: With Applications in R (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press. doi:10.1017/9781108644181.
- Engelstätter B., Ward M. (2013) *Strategic Timing of Entry: Evidence from Video Games*. SSRN Electronic Journal, <https://doi.org/10.2139/ssrn.2379471>.
- Fleischer G., Sarnik K., Brodzki E. (2010) *Statistical Analysis of Gamer Behavior*. Worcester Polytechnic Institute.
- Johnson R. A., Wichern D. W. (2008) *Applied Multivariate Statistical Analysis*. New Jersey, Prentice Hall.
- Kassambara A. (2017) *Practical Guide to Cluster Analysis in R, Unsupervised Machine Learning*. STDH.
- Powell R. (2016) *Positive and Negative Effects of Game Analytics in the Game Design process: A Grounded Theory Study*. <https://www.semanticscholar.org/paper/Positive-and-Negative-effects-of-Game-Analytics-in-Powell/8fb33e472e2e03ce4a3f0aece75f222d68310183> [dostęp: 02.02.2023].
- Tong X. (2021) *Positioning Game Review as a Crucial Element of Game User Feedback in the Ongoing Development of Independent Video Games*. *Computers in Human Behavior Reports*, 3, 100077. <https://doi.org/10.1016/j.chbr.2021.100077>.

**MULTIDIMENSIONAL STATISTICAL ANALYSIS TEKKEN 7**

**Abstract:** The main goal of the study is to verify whether premium characters have an advantage over the starter characters. The objective was achieved by conducting a statistical analysis of victories in the TEKKEN 7 game for all 50 characters (warriors) present in the game. Methods such as *k*-means clustering, hierarchical clustering, and Model-Based Clustering were employed to identify homogeneous groups of characters. Additionally, a ranking based on the applied methods was constructed using the TOPSIS method.

**Keywords:** *k*-means, hierarchical clustering, model based clustering, TOPSIS

**JEL classification:** C38, M210

## WYKORZYSTANIE ANALIZY DYSKRYMINACYJNEJ DO OCENY RYZYKA BANKRUCTWA PRZEDSIĘBIORSTW BRANŻY MOTORYZACYJNEJ W OKRESIE PANDEMII COVID-19

Katarzyna Boratyńska<sup>1</sup>  <https://orcid.org/0000-0001-8803-3234>

Instytut Ekonomii i Finansów  
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie  
e-mail: katarzyna\_boratynska@sggw.edu.pl

**Katsiaryna Yafremava**  
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

**Streszczenie:** W artykule rozpoznano skuteczność wykorzystania siedmiu modeli dyskryminacyjnych do oceny zagrożenia bankructwem przedsiębiorstw sektora motoryzacyjnego w okresie pandemii COVID-19. Zastosowano metody takie jak: analizę dyskryminacyjną, analizę opisową oraz analizę porównawczą. Na podstawie uzyskanych wyników modeli stwierdzono, że w czasie pandemii COVID-19 liczba badanych przedsiębiorstw zagrożonych bankructwem oraz znajdujących się według modelu w tzw. „szarej strefie” znacząco wzrosła. Tylko dwa modele nie wykazały zagrożenia bankructwem badanych spółek w analizowanym okresie.

**Słowa kluczowe:** analiza dyskryminacyjna, modele predykcji bankructwa przedsiębiorstw, finanse przedsiębiorstw, branża motoryzacyjna, pandemia COVID-19

**JEL classification:** G01, G17, G33

### WSTĘP

Sektor motoryzacyjny, będący jednym z kluczowych obszarów gospodarek na całym świecie, jest poddawany ciągłej presji konkurencji oraz tempa zmian warunków rynkowych. Przedstawiciele przedsiębiorstw działających w tym sektorze muszą podejmować decyzje gospodarcze w celu osiągnięcia przewagi

---

<sup>1</sup> autorka korespondencyjna



konkurencyjnej i zwiększenia efektywności operacyjnej. W czasie pandemii COVID-19 narzędzia analizy danych, w tym modele dyskryminacyjne, stały się ważnym instrumentem wspierającym te decyzje.

Analiza dyskryminacyjna jest metodą statystyczną, która może być stosowana do oceny ryzyka bankructwa przedsiębiorstw. Opiera się ona na badaniu zestawu wskaźników finansowych, takich jak np. wskaźniki płynności, rentowności i zadłużenia. Na podstawie wskaźników finansowych można opracować model służący do prognozowania bankructwa przedsiębiorstw.

Analiza dyskryminacyjna może być szczególnie przydatna w branży motoryzacyjnej. Zastosowanie tej metody może pomóc przedsiębiorstwom funkcjonować w trudnych warunkach rynkowych przez wczesne wykrycie zagrożeń i podjęcie odpowiednich działań prewencyjnych.

W artykule dokonano także identyfikacji ograniczeń związanych z wykorzystaniem modeli dyskryminacyjnych. Mogą one obejmować trudności związane z właściwą interpretacją wyników modeli, czy dostosowaniem modeli do zmieniających się warunków rynkowych. K. Jajuga [2019, s. 142] zwraca uwagę, że „...modele ekonomiczne są mało odporne na zmiany w gospodarce, a zmiany te są coraz bardziej dynamiczne...”. (...). „...Brak odporności modelu, który dobrze sprawdzał się w przeszłości, na zmiany warunków w gospodarce, powoduje jego nieprzydatność...” [Jajuga 2019, s. 142].

Branża motoryzacyjna w Polsce jest jednym z najszybciej rozwijających się sektorów gospodarki, w związku z tym powinna podlegać stałemu monitoringowi w zakresie zagrożenia bankructwem podmiotów, które w niej funkcjonują. Od lat dziewięćdziesiątych XX wieku nastąpił istotny wzrost inwestycji, produkcji i zatrudnienia w tym sektorze. Jednym z głównych czynników przyciągających inwestycje do branży motoryzacyjnej w Polsce jest wysoko wykwalifikowana siła robocza oraz stosunkowo niskie koszty pracy w porównaniu do innych krajów europejskich. Polska ma bogate dziedzictwo przemysłu mechanicznego i tradycję wytwarzania części samochodowych, co przyciąga producentów poszukujących dostawców o wysokich standardach jakościowych.

Podczas pandemii COVID-19 w branży motoryzacyjnej nastąpiło zerwanie globalnych łańcuchów dostaw, co spowodowało dostosowanie się badanych podmiotów do dynamicznie zmieniających się warunków gospodarowania. Wykorzystanie metod takich jak analiza dyskryminacyjna może znacząco wspierać monitorowanie zagrożeń i ocenę sytuacji finansowej podmiotów funkcjonujących w branży motoryzacyjnej.

## ISTOTA ANALIZY Dyskryminacyjnej

Analiza dyskryminacyjna stwarza możliwości konstrukcji formuły matematycznej, identyfikującej przynależność obiektów do jednej z kilku wyróżnionych grup, przy czym najprostszym przypadkiem jest podział na dwie grupy. Konstrukcja reguły przynależności opiera się na wielowymiarowej

informacji o każdej badanej jednostce (obiekcie) i jej przynależności do określonej grupy [Kisielińska, Borkowski i in. 2021, s. 61-62].

Analiza dyskryminacyjna jest metodą statystyczną służącą do badania, w jaki sposób grupy różnią się od siebie na podstawie jednej lub więcej zmiennych. Celem analizy dyskryminacyjnej jest wyodrębnienie cech, które najlepiej rozróżniają grupy i pozwalają na skuteczne klasyfikowanie obserwacji do odpowiedniej kategorii na podstawie wartości tych cech. Analiza dyskryminacyjna zakłada, że grupy są wzajemnie wykluczające się, a zmienne mają rozkład normalny w każdej grupie.

W analizie dyskryminacyjnej kluczową rolę odgrywa dobór zmiennych, które będą służyły do rozróżnienia badanych grup. W tym celu można wykorzystać różne techniki selekcji zmiennych, takie jak analiza wariancji, analiza korelacji czy metody oparte na regresji. Analiza dyskryminacyjna może być również stosowana do badania wielowymiarowej struktury danych. Prekursorem badań z wykorzystaniem analizy dyskryminacyjnej w obszarze ekonomii i finansów jest E. I. Altman [1968]. Według [Wysocki, Kozera 2012, s. 169] rolą analizy dyskryminacyjnej i opartych na niej systemów wczesnego ostrzegania jest dokonanie kompleksowej oceny kondycji ekonomiczno-finansowej przedsiębiorstwa oraz ujawnienie elementów wskazujących na wzrastające ryzyko upadłości.

Analiza dyskryminacyjna może być stosowana do przewidywania ryzyka bankructwa przedsiębiorstw za pomocą modeli dyskryminacyjnych. Analizując dane finansowe można zidentyfikować kluczowe czynniki, które odróżniają firmy zagrożone bankructwem od tych, znajdujących się w dobrej kondycji finansowej. Interpretacja wyniku dokonywana jest na podstawie porównania obliczonej wartości  $Z$  funkcji dyskryminacyjnej z wartością graniczną, której poziom ustala twórca modelu [Zielińska-Sitkiewicz, 2012, s. 289–290].

## CEL BADAŃ, MATERIAŁ I METODY BADAWCZE

Celem badań było rozpoznanie skuteczności wykorzystania modeli dyskryminacyjnych do oceny zagrożenia bankructwem przedsiębiorstw sektora motoryzacyjnego w okresie pandemii COVID-19.

W artykule zastosowano następujące metody badawcze: analizę dyskryminacyjną, wybrane modele predykcji bankructwa przedsiębiorstw stworzone w Polsce (szczegółowo przedstawione w dalszej części opracowania), analizę opisową oraz analizę porównawczą. W badaniach skupiono się wyłącznie na modelach dyskryminacyjnych stworzonych dla uwarunkowań gospodarczych w Polsce. Przenoszenie modeli między państwami nie jest wskazane w związku z odmiennymi uwarunkowaniami, dla których są tworzone modele zagraniczne, jak również wynika z odmiennej specyfiki systemów rachunkowości. Okres badań objął lata 2018–2021 (w celach porównawczych dwa lata przed i dwa lata

w trakcie pandemii COVID-19). W badaniach wykorzystano dane ze sprawozdań finansowych analizowanych podmiotów, zamieszczone w bazie EMIS.

### Modele predykcji bankructwa przedsiębiorstw

Czerwińska-Kayzer i Florek [2012, s. 267] podkreślają, że intensywne badania nad systemami wczesnego ostrzegania, które rozpoczęły się w latach dziewięćdziesiątych XX w. doprowadziły do sytuacji, że literatura przedmiotu zawiera wiele różnorodnych modeli, o różnej przydatności w ocenie sytuacji finansowej przedsiębiorstw. W badaniach zastosowano, kierując się upowszechnieniem w literaturze przedmiotu, siedem modeli dyskryminacyjnych. Ocenę zagrożenia bankructwem badanych podmiotów dokonano z wykorzystaniem modeli predykcji bankructwa przedsiębiorstw stworzonych w Polsce przez: Dorotę Hadasik; Dariusza Wierzbę; Artura Hołdę; Elżbietę Mączyńską, Macieja Zawadzkiego (model F INE PAN); Mirosława Hamrola, Bartłomieja Czajkę i Macieja Piechockiego („model poznański”); Błażeja Prusaka (P1); Elżbietę Mączyńską, Macieja Zawadzkiego (model G INE PAN). Modele te omówiono kolejno w tej części artykułu, począwszy do modelu Doroty Hadasik.

#### Model Doroty Hadasik

Model Doroty Hadasik został opracowany na podstawie analizy sprawozdań finansowych 61 przedsiębiorstw, w tym 39 nie bankrutujących i 22 bankrutujących. Model ten wykazał wyjściową skuteczność na poziomie 95,08% według informacji od autorki modelu, natomiast na podstawie badań M. Hamrola i J. Chodakowskiego uzyskał ogólną skuteczność prognoz na poziomie 57,6% [Hamrol, Chodakowski, 2008, s. 29].

Model D. Hadasik opisany jest wzorem [Hamrol, Chodakowski, 2008, s. 21]:

$$Z_{HA} = 2,36261 + 0,365425X_1 - 0,765526X_2 - 2,40435X_3 + 1,29079X_4 + 0,00230258X_5 - 0,0127826X_6 \quad (1)$$

gdzie:

$X_1$  = aktywa bieżące/zobowiązania bieżące,

$X_2$  = (aktywa bieżące – zapasy)/zobowiązania bieżące,

$X_3$  = zobowiązania ogółem/aktywa ogółem,

$X_4$  = (aktywa bieżące – zobowiązania krótkoterminowe)/pasywa ogółem,

$X_5$  = należności/przychody ze sprzedaży,

$X_6$  = zapasy/przychody ze sprzedaży.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_{HA}$ :

$Z_{HA} = -0,374345$  wartość krytyczna modelu

$Z_{HA} \leq -0,374345$  przedsiębiorstwa zagrożone upadłością

$Z_{HA} > -0,374345$  przedsiębiorstwa o dobrej kondycji finansowej

### Model Dariusza Wierzby

Model Dariusza Wierzby został opracowany na podstawie analizy sprawozdań finansowych 48 przedsiębiorstw, w tym 24 spółek niezagrożonych bankrutem i 24 zagrożonych bankrutem. Ten model uzyskał 92% skuteczności w próbie uczącej i 75,9% ogólnej skuteczności prognoz w badaniu M. Hamrola i J. Chodakowskiego [Hamrol, Chodakowski, 2008, s. 29].

Model Dariusza Wierzby przyjmuje postać [Wierzba, 2000, s. 79-105]:

$$Z_W = 3,26X_1 + 2,16X_2 + 0,3X_3 + 0,69X_4 \quad (2)$$

gdzie:

$X_1$  = (zysk z działalności operacyjnej – amortyzacja)/aktywa ogółem,

$X_2$  = (zysk z działalności operacyjnej – amortyzacja)/  
przychody ze sprzedaży produktów,

$X_3$  = aktywa obrotowe/zobowiązania ogółem,

$X_4$  = kapitał obrotowy/aktywa ogółem.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_W$ :

$Z_W = 0$  wartość krytyczna modelu

$Z_W < 0$  przedsiębiorstwa zagrożone upadłością

$Z_W > 0$  przedsiębiorstwa o dobrej kondycji finansowej

### Model Artura Hołdy

Model Artura Hołdy został opracowany na podstawie analizy sprawozdań finansowych 80 przedsiębiorstw z lat 1993-1996, w tym jako bankruci zakwalifikowano 40, z kolei 40 jako przedsiębiorstwa niezagrożone upadłością [Hołda, 2001, s. 306].

Model Artura Hołdy można opisać równaniem [Kisielińska, Waszkowski, 2010, s. 19]:

$$Z_{HO} = 0,605 + 0,681X_1 - 0,0196X_2 + 0,157X_3 + 0,00969X_4 + 0,000672X_5 \quad (3)$$

gdzie:

$X_1$  = aktywa obrotowe/zobowiązania krótkoterminowe,

$X_2$  = (zobowiązania ogółem/aktywa ogółem)  $\times$  100,

$X_3$  = przychody ogółem/aktywa ogółem,

$X_4$  = (wynik finansowy netto/aktywa ogółem)  $\times$  100,

$X_5$  = (zobowiązania krótkoterminowe  $\times$  360)/  
/(łącznie koszty działalności operacyjnej –  
pozostałe koszty działalności operacyjnej).

Interpretacja wartości funkcji dyskryminacyjnej  $Z_{HO}$ :

$Z_{HO} = 0$  wartość krytyczna modelu

$Z_{HO} \leq 0$  przedsiębiorstwa zagrożone upadłością

$-0,3 \leq Z_{HO} \leq 0,1$  „szara strefa”, czyli wysokie prawdopodobieństwo błędnej klasyfikacji

$Z_{HO} > 0$  przedsiębiorstwa o dobrej kondycji finansowej

### Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)

Model Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) został opracowany na podstawie analizy sprawozdań finansowych 80 przedsiębiorstw. Według badań P. Antonowicza [2007] model ten charakteryzował się sprawnością na poziomie 94,20%.

Model F INE PAN określony jest wzorem [Mączyńska, Zawadzki 2006, s. 21-24]:

$$Z_6 = 9,478X_1 + 3,613X_2 + 3,246X_3 + 0,455X_4 + 0,802X_5 - 2,478 \quad (4)$$

gdzie:

$X_1$  = wynik z działalności operacyjnej/aktywa ogółem,

$X_2$  = kapitał własny/aktywa ogółem,

$X_3$  = (wynik finansowy netto + amortyzacja)/zobowiązania ogółem,

$X_4$  = aktywa obrotowe/zobowiązania krótkoterminowe,

$X_5$  = przychody ze sprzedaży/aktywa ogółem.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_6$ :

$Z_6 = 0$  wartość krytyczna modelu

$Z_6 \leq 0$  przedsiębiorstwa zagrożone upadłością

$Z_6 > 0$  przedsiębiorstwa o dobrej kondycji finansowej

### Model Mirosława Hamrola, Bartłomieja Czajki i Macieja Piechockiego (nazwany modelem poznańskim - nazwę tę zastosowano w dalszej części artykułu)

Model poznański został opracowany na podstawie analizy sprawozdań finansowych 100 przedsiębiorstw z lat 1999-2002, w tym 50 spółek niezagrożonych bankructwem i 50 zagrożonych bankructwem. Model ten wykazał wyjściową skuteczność na poziomie 96% według informacji od autorów modelu natomiast na podstawie badań M. Hamrola i J. Chodakowskiego uzyskał ogólną skuteczność prognoz na poziomie 54,8% [Hamrol, Chodakowski, 2008, s. 29].

Model poznański przyjmuje postać [Hamrol, M., Czajka, B., Piechocki, M., 2004, s. 35-39]:

$$Z_{PO} = 3,562W_7 + 1,588W_{16} + 4,288W_5 + 6,719W_{13} - 2,368 \quad (5)$$

gdzie:

$W_7$  = wynik finansowy netto/majątek całkowity,

$W_{16}$  = (majątek obrotowy – zapasy)/zobowiązania krótkoterminowe,

$W_5$  = kapitał stały/majątek całkowity,

$W_{13}$  = wynik finansowy ze sprzedaży/przychody ze sprzedaży.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_{PO}$ :

$Z_{PO} = 0$  wartość krytyczna modelu

$Z_{PO} \leq 0$  przedsiębiorstwa zagrożone upadłością

$Z_{PO} > 0$  przedsiębiorstwa o dobrej kondycji finansowej

Poszczególne zmienne  $W$  ( $W_7$ ,  $W_{16}$ ,  $W_5$ ,  $W_{13}$ ) w modelu poznańskim oznaczają wskaźniki finansowe, które według badań jego twórców z przyjętej

grupy 31 wskaźników finansowych weszły jako zmienne do finalnej funkcji modelu poznańskiego [Hamrol, M., Czajka, B., Piechocki, M., 2004, s. 35-39].

#### Model Błażeja Prusaka (P1)

Zbiór danych, na którym B. Prusak prowadził badanie, został podzielony na próbę uczącą oraz próbę testową. Pierwsza składała się z 40 firm znajdujących się w dobrej kondycji finansowej oraz 40 zagrożonych upadłością. Jednostki były zestawione parami według branż. Próba testowa składała się z 39 firm, które zbankrutowały i 39 przedsiębiorstw niezagrażonych bankructwem.

Model Błażeja Prusaka (P1) można opisać, równaniem [Prusak 2005, s. 151]:

$$Z_{PR1} = 6,5245X_1 + 0,148X_2 + 0,4061X_3 + 2,1754X_4 - 1,5685 \quad (6)$$

gdzie:

$X_1$  = wynik z działalności operacyjnej/suma bilansowa,

$X_2$  = koszty działalności operacyjnej/zobowiązania krótkoterminowe,

$X_3$  = aktywa obrotowe/zobowiązania krótkoterminowe,

$X_4$  = wynik z działalności operacyjnej/przychody netto ze sprzedaży.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_{PR1}$ :

$Z_{PR1} = -0,13$  wartość krytyczna modelu

$Z_{PR1} \leq 0$  przedsiębiorstwa zagrożone upadłością

$-0,13 < Z_{PR1} \leq 0,65$  „szara strefa”, czyli wysokie prawdopodobieństwo błędnej klasyfikacji

$Z_{PR1} > 0,65$  przedsiębiorstwa o dobrej kondycji finansowej

#### Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)

Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) został opracowany na podstawie analizy sprawozdań finansowych 80 przedsiębiorstw. Według badań P. Antonowicza [Antonowicz 2007] ten model charakteryzował się sprawnością na poziomie 94,82%.

Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) określony jest wzorem [Mączyńska, Zawadzki 2006, s. 21-24]:

$$Z_7 = 9,498X_1 + 3,566X_2 + 2,903X_3 + 0,452X_4 - 1,498 \quad (7)$$

gdzie:

$X_1$  = wynik z działalności operacyjnej/aktywa ogółem,

$X_2$  = kapitał własny/aktywa ogółem,

$X_3$  = (wynik finansowy netto + amortyzacja)/zobowiązania ogółem,

$X_4$  = aktywa obrotowe/zobowiązania krótkoterminowe.

Interpretacja wartości funkcji dyskryminacyjnej  $Z_7$ :

$Z_7 = 0$  wartość krytyczna modelu

$Z_7 \leq 0$  przedsiębiorstwa zagrożone upadłością

$Z_7 > 0$  przedsiębiorstwa o dobrej kondycji finansowej

### Próba badawcza

Do badań zostało wybranych dziesięć następujących, nienotowanych na giełdzie przedsiębiorstw z branży motoryzacyjnej prowadzących działalność w Polsce: Volkswagen Group Polska Sp. z o.o., Hyundai Motor Poland Sp. z o.o., Kia Polska Sp. z o.o., BMW Vertriebs GmbH Sp. z o.o., Ford Polska Sp. z o.o., Toyota Central Europe Sp. z o.o., Volvo Polska Sp. z o.o., Porsche Inter Auto Polska Sp. z o.o., Mercedes-Benz Polska Sp. z o.o., Renault Polska Sp. z o.o..

Wybór obiektów badawczych został dokonany w sposób uwzględniający spełnienie określonych kryteriów, tj.:

- dostępność sprawozdań finansowych za lata 2018-2021,
- przynależność każdego przedsiębiorstwa do branży motoryzacyjnej,
- spółki branży motoryzacyjnej nienotowane na giełdzie,
- forma organizacyjno-prawna przedsiębiorstwa to spółka z o.o.

Zastosowanie danych finansowych z bazy EMIS pozwoliło na przeprowadzenie analizy porównawczej pomiędzy różnymi przedsiębiorstwami z branży motoryzacyjnej.

### WYNIKI BADAŃ

Na podstawie przeprowadzonych badań można wywnioskować, że spośród 10 badanych przedsiębiorstw spółki Volkswagen Group Polska Sp. z o.o. oraz Hyundai Motor Poland Sp. z o.o. wyróżniały się dobrą kondycją finansową. W analizowanym okresie obejmującym lata 2018-2021 żaden z siedmiu modeli dyskryminacyjnych nie wykazał zagrożenia bankructwem w przypadku tych dwóch spółek.

Wartości funkcji dyskryminacyjnych dla spółki Kia Polska Sp. z o.o. wskazują na stabilną sytuację finansową w analizowanym okresie (tabela 1). Jedynie w przypadku modelu poznańskiego wartość funkcji dyskryminacyjnej spadła nieznacznie poniżej wartości krytycznej w 2018 roku. Model Błażeja Prusaka (P1) wskazał, że spółka Kia Polska Sp. z o.o. znajduje się w „szarej strefie” i jest zagrożona upadłością w 2020 roku. Według wyników wartości funkcji dyskryminacyjnych reprezentowanych przez modele Doroty Hadasik; Dariusza Wierzby; Artura Hołdy; Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) oraz Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) nie stwierdzono zagrożenia upadłością (por. tabela 1).

W przypadku spółki BMW Vertriebs GmbH Sp. z o.o. wartości funkcji dyskryminacyjnych świadczą o zróżnicowanej sytuacji finansowej w roku 2020. Według modelu Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) i modelu poznańskiego, przedsiębiorstwo to było zagrożone bankructwem w 2020 roku. Model Błażeja Prusaka (P1) również wskazał, że spółka BMW Vertriebs GmbH Sp. z o.o. znajdowała się w „szarej strefie” i była zagrożona upadłością

w 2020 roku. W przypadku pozostałych modeli funkcje dyskryminacyjne nie wskazywały na ryzyko bankructwa, jednak w roku 2020 przyjmowały znacznie niższe wartości w porównaniu do roku 2018 (tabela 1). Może to świadczyć o wpływie pandemii COVID-19 na działalność tych podmiotów.

Tabela 1. Wyniki zastosowanych modeli dyskryminacyjnych dla spółek Kia Polska Sp. z o.o. oraz BMW Vertriebs Gmbh Sp. z o.o. w latach 2018–2021

Oznaczenie modelu	Nazwa modelu	Kia Polska Sp. z o.o.			
		2018	2019	2020	2021
$Z_{HA}$	Model Doroty Hadasik	0,99	1,04	0,94	1,18
$Z_W$	Model Dariusza Wierzby	1,27	1,29	1,26	1,43
$Z_{HO}$	Model Artura Hołdy	0,87	1,00	0,65	1,19
$Z_6$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	4,13	4,60	2,59	4,35
$Z_{PO}$	Model poznański	-0,002	0,35	0,44	1,27
$Z_{PR1}$	Model Błażeja Prusaka (P1)	0,93	1,04	0,31	1,04
$Z_7$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	1,51	1,72	1,27	2,21
Oznaczenie modelu	Nazwa modelu	BMW Vertriebs Gmbh Sp. z o.o.			
		2018	2019	2020	2021
$Z_{HA}$	Model Doroty Hadasik	0,08	0,84	0,62	-0,22
$Z_W$	Model Dariusza Wierzby	1,18	1,34	0,89	1,35
$Z_{HO}$	Model Artura Hołdy	1,39	1,40	0,82	1,09
$Z_6$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego(model F INE PAN)	3,51	4,12	1,84	3,54
$Z_{PO}$	Model poznański	1,88	1,20	-0,42	1,37
$Z_{PR1}$	Model Błażeja Prusaka (P1)	1,33	1,83	0,56	1,44
$Z_7$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	1,32	1,62	-0,45	1,30

Objaśnienia do tabeli: ciemnoszarym kolorem oznaczono spółki zagrożone bankructwem według modelu; jasnoszarym kolorem oznaczono spółki znajdujące się w „szarej strefie” według modelu

Źródło: Badania własne na podstawie danych ze sprawozdań finansowych wybranych spółek, dostępnych w bazie danych Emerging Markets Information Service – Polska (EMIS) za lata 2018-2021

Dla spółki Ford Polska Sp. z o.o. wyniki analizy wartości większości funkcji dyskryminacyjnych wskazują na brak stabilnej sytuacji finansowej (tabela 2). W analizowanym okresie model poznański oraz model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) przyjmują ujemne wartości funkcji dyskryminacyjnej, co świadczy o potencjalnym zagrożeniu bankructwem. Z modelu Prusaka (P1) wynika również, że przedsiębiorstwo Ford Polska Sp. z o.o. znajdowało się w „szarej strefie” i groziło mu ryzyko upadłości w latach 2019-2021. Według modelu Artura Hołdy, przedsiębiorstwo było w „szarej strefie”



w roku 2020 i 2021. Model Doroty Hadasik również pokazał, że spółka była zagrożona bankrutwem w 2019 roku. Zgodnie z wynikami wartości funkcji dyskryminacyjnych reprezentowanych przez modele Dariusza Wierzby, Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) nie wystąpiło zagrożenie upadłością w analizowanym okresie.

Tabela 2. Wyniki zastosowanych modeli dyskryminacyjnych dla spółek Ford Polska Sp. z o.o. oraz Toyota Central Europe Sp. z o.o. w latach 2018–2021

Oznaczenie modelu	Nazwa modelu	Ford Polska Sp. z o.o.			
		2018	2019	2020	2021
$Z_{HA}$	Model Doroty Hadasik	-0,366	-0,39	0,46	0,47
$Z_W$	Model Dariusza Wierzby	1,18	1,17	1,09	1,04
$Z_{HO}$	Model Artura Hołdy	0,63	0,49	-0,06	0,07
$Z_6$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	3,52	3,12	1,39	1,77
$Z_{PO}$	Model poznański	-0,24	-0,37	-1,17	-1,18
$Z_{PR1}$	Model Błażeja Prusaka (P1)	0,82	0,55	-0,06	0,08
$Z_7$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	-0,08	-0,20	-0,40	-0,33
Oznaczenie modelu	Nazwa modelu	Toyota Central Europe Sp. z o.o.			
		2018	2019	2020	2021
$Z_{HA}$	Model Doroty Hadasik	0,35	0,39	0,35	0,37
$Z_W$	Model Dariusza Wierzby	1,03	1,11	1,07	1,10
$Z_{HO}$	Model Artura Hołdy	-0,15	-0,12	-0,10	-0,05
$Z_6$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	0,16	0,33	0,43	0,52
$Z_{PO}$	Model poznański	-1,00	-0,96	-0,95	-0,70
$Z_{PR1}$	Model Błażeja Prusaka (P1)	-0,68	-0,54	-0,64	-0,62
$Z_7$	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	-0,87	-0,58	-0,76	-0,61

Objaśnienia do tabeli: ciemnoszarym kolorem oznaczono spółki zagrożone bankrutwem według modelu; jasnoszarym kolorem oznaczono spółki znajdujące się w „szarej strefie” według modelu

Źródło: Badania własne na podstawie danych ze sprawozdań finansowych wybranych spółek, dostępnych w bazie danych Emerging Markets Information Service – Polska (EMIS) za lata 2018-2021

W przypadku spółki Toyota Central Europe Sp. z o.o. wyniki analizy wartości większości funkcji dyskryminacyjnych świadczą o niestabilnej sytuacji finansowej (tabela 2). Model poznański, model Błażeja Prusaka (P1) oraz model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) wykazują ujemne wartości przez cały analizowany okres. Z modelu Artura Hołdy wynika, że firma Kia Polska Sp. z o.o. znajduje się w „szarej strefie” i była zagrożona upadłością przez cały badany okres. Na podstawie wartości funkcji

dyskryminacyjnych reprezentowanych modelami Doroty Hadasik, Dariusza Wierzby, Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) można stwierdzić, że w analizowanym okresie nie wystąpiło zagrożenie bankrutwem.

W odniesieniu do spółki Volvo Polska Sp. z o.o. (por. tabela 3), wyniki modelu poznańskiego, jak i modelu Błażeja Prusaka (P1) wykazują ujemne wartości w analizowanym okresie, co wskazywało na potencjalne zagrożenie bankrutwem. Na podstawie analizy wyników modelu Artura Hołdy, można stwierdzić, że przedsiębiorstwo Volvo Polska Sp. z o.o. znajdowało się w "szarej strefie" w latach 2018, 2019 i 2021, natomiast w roku 2020 istniało zagrożenie upadłością. Zgodnie z wynikami modeli: Doroty Hadasik; Dariusza Wierzby; Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN oraz model G INE PAN), nie zaobserwowano zagrożenia bankrutwem.

Tabela 3. Wyniki zastosowanych modeli dyskryminacyjnych dla spółek Volvo Polska - Sp. z o.o. oraz Porsche Inter Auto Polska Sp. z o.o. w latach 2018–2021

Oznaczenie modelu	Nazwa modelu	Volvo Polska - Sp. z o.o.			
		2018	2019	2020	2021
Z <sub>HA</sub>	Model Doroty Hadasik	0,91	0,74	0,52	0,58
Z <sub>W</sub>	Model Dariusza Wierzby	0,45	0,49	0,52	0,52
Z <sub>HO</sub>	Model Artura Hołdy	-0,06	-0,13	-0,32	-0,22
Z <sub>6</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	0,61	0,99	1,04	1,28
Z <sub>PO</sub>	Model poznański	-0,40	-0,37	-0,21	-0,32
Z <sub>PR1</sub>	Model Błażeja Prusaka (P1)	-0,41	-0,40	-0,38	-0,39
Z <sub>7</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	0,12	0,22	0,36	0,32
Oznaczenie modelu	Nazwa modelu	Porsche Inter Auto Polska Sp. z o.o.			
		2018	2019	2020	2021
Z <sub>HA</sub>	Model Doroty Hadasik	0,99	0,89	1,01	0,93
Z <sub>W</sub>	Model Dariusza Wierzby	0,84	0,85	0,75	0,86
Z <sub>HO</sub>	Model Artura Hołdy	0,62	0,36	0,68	0,99
Z <sub>6</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	1,15	0,62	1,26	2,11
Z <sub>PO</sub>	Model poznański	-0,89	-1,34	-1,20	0,12
Z <sub>PR1</sub>	Model Błażeja Prusaka (P1)	0,18	-0,22	0,29	0,90
Z <sub>7</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	0,17	-0,13	-0,02	0,57

Objaśnienia do tabeli: ciemnoszarym kolorem oznaczono spółki zagrożone bankrutwem według modelu; jasnoszarym kolorem oznaczono spółki znajdujące się w „szarej strefie” według modelu

Źródło: Badania własne na podstawie danych ze sprawozdań finansowych wybranych spółek, dostępnych w bazie danych Emerging Markets Information Service – Polska (EMIS) za lata 2018-2021

Dla spółki Porsche Inter Auto Polska Sp. z o.o. wyniki modelu poznańskiego sygnalizują o potencjalnym zagrożeniu upadłością w latach 2018, 2019 i 2020 (tabela 3). Model Błażeja Prusaka (P1) wskazywał na zagrożenie bankrutwem w 2019 roku, z kolei w 2018 i 2020 roku spółka utrzymywała się w „szarej strefie”. Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) również sygnalizował ryzyko upadłości w 2019 i 2020 roku. Na podstawie wyników analiz przeprowadzonych przez modele Doroty Hadasik; Dariusza Wierzby; Artura Hołdy oraz Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN), nie stwierdzono zagrożenia bankrutwem.

Tabela 4. Wyniki zastosowanych modeli dyskryminacyjnych dla spółek Mercedes-Benz Polska Sp. z o.o. oraz Renault Polska Sp. z o.o. w latach 2018–2021

Oznaczenie modelu	Nazwa modelu	Mercedes-Benz Polska Sp. z o.o.			
		2018	2019	2020	2021
Z <sub>HA</sub>	Model Doroty Hadasik	0,48	0,53	0,57	0,65
Z <sub>W</sub>	Model Dariusza Wierzby	0,69	0,89	1,01	1,25
Z <sub>HO</sub>	Model Artura Hołdy	0,22	0,21	0,30	0,55
Z <sub>6</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	1,66	1,75	1,63	2,98
Z <sub>PO</sub>	Model poznański	-0,34	-0,53	-0,35	0,24
Z <sub>PR1</sub>	Model Błażeja Prusaka (P1)	0,22	0,0001	0,01	0,55
Z <sub>7</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	0,47	0,20	0,33	1,17
Oznaczenie modelu	Nazwa modelu	Renault Polska Sp. z o.o.			
		2018	2019	2020	2021
Z <sub>HA</sub>	Model Doroty Hadasik	-0,26	-0,89	-0,40	-0,41
Z <sub>W</sub>	Model Dariusza Wierzby	1,18	1,16	0,96	1,10
Z <sub>HO</sub>	Model Artura Hołdy	1,05	1,19	0,74	0,45
Z <sub>6</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN)	4,33	4,77	2,50	2,52
Z <sub>PO</sub>	Model poznański	0,43	0,89	0,33	-0,20
Z <sub>PR1</sub>	Model Błażeja Prusaka (P1)	1,60	1,63	0,45	0,32
Z <sub>7</sub>	Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN)	0,43	0,64	-0,14	-0,22

Objaśnienia do tabeli: ciemnoszarym kolorem oznaczono spółki zagrożone bankrutwem według modelu; jasnoszarym kolorem oznaczono spółki znajdujące się w „szarej strefie” według modelu

Źródło: Badania własne na podstawie danych ze sprawozdań finansowych wybranych spółek, dostępnych w bazie danych Emerging Markets Information Service – Polska (EMIS) za lata 2018-2021

W przypadku spółki Mercedes-Benz Polska Sp. z o.o. (tabela 4), wartości funkcji dyskryminacyjnych wskazują na zróżnicowaną sytuację finansową. Model poznański sygnalizował o potencjalnym zagrożeniu bankrutwem w latach 2018-

2020. Z modelu Błażeja Prusaka (P1) również wynika, że przedsiębiorstwo znajdowało się w „szarej strefie” w analizowanym okresie. Na podstawie wyników modeli: Doroty Hadasik; Dariusza Wierzby; Artura Hołdy; Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) oraz Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) nie zidentyfikowano ryzyka bankructwa.

Dla przedsiębiorstwa Renault Polska Sp. Z o.o. wartości funkcji dyskryminacyjnych wskazują na zróżnicowaną sytuację finansową (tabela 4). Model Doroty Hadasik sygnalizował o potencjalnym zagrożeniu upadłością w latach 2019- 2021. Model poznański wskazywał na zagrożenie upadłością w 2021 roku. Z kolei w latach 2018 i 2020 spółka pozostawała w „szarej strefie” według Modelu Błażeja Prusaka (P1). Z modelu Elżbiety Mączyńskiej, Macieja Zawadzkiego (model G INE PAN) wynika, że firma Renault Polska Sp. z o.o. była zagrożona bankructwem w 2020 i 2021 roku. Zgodnie z wynikami wartości funkcji dyskryminacyjnych reprezentowanych przez modele Dariusza Wierzby; Artura Hołdy; Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) nie stwierdzono ryzyka bankructwa w analizowanym okresie.

W wyniku analizy danych można zaobserwować, że wraz z początkiem pandemii COVID-19 wzrosła liczba podmiotów zagrożonych bankructwem lub należących do kategorii "szarej strefy" według większości badanych modeli, co wskazuje, że sytuacja finansowa niektórych badanych przedsiębiorstw w latach 2020-2021 uległa pogorszeniu. Jednakże Model Dariusza Wierzby oraz Model Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) nie wykazały ryzyka bankructwa badanych spółek w analizowanym okresie.

## PODSUMOWANIE

W artykule zweryfikowano skuteczność siedmiu polskich modeli analizy dyskryminacyjnej do oceny ryzyka bankructwa dla dziesięciu spółek branży motoryzacyjnej. Ze względu na nielosowy dobór próby wyniki badań dotyczą badanych podmiotów i nie są uogólniane na całą populację generalną przedsiębiorstw.

Na podstawie danych ze sprawozdań finansowych została wyliczona wartość funkcji dyskryminacyjnych wybranych modeli, dzięki czemu można było ocenić ich skuteczność w przewidywaniu bankructwa. W przypadku modelu Dariusza Wierzby oraz modelu Elżbiety Mączyńskiej, Macieja Zawadzkiego (model F INE PAN) nie wystąpił ani jeden przypadek zagrożenia upadłością badanych przedsiębiorstw w czasie pandemii COVID-19. Natomiast pozostałe pięć modeli wysyłało sygnały ostrzegawcze, że przedsiębiorstwa znajdowały się w „szarej strefie” i były zagrożone bankructwem w latach 2018-2021. W okresie pandemii COVID-19 wzrosła liczba ujemnych wartości analizowanych modeli dyskryminacyjnych.

Można stwierdzić, że wykorzystane modele analizy dyskryminacyjnej są częściowo skuteczne w prognozowaniu zagrożenia upadłością dla badanych przedsiębiorstw. W toku przeprowadzonych badań stwierdzono, że dokonując oceny ryzyka bankructwa podmiotów, istnieje konieczność uwzględnienia wielu modeli z uwagi na brak ich homogeniczności. Zastosowanie podejścia „wielomodelowego” wynikało z przyjętej perspektywy badawczej, która wskazuje na konieczność uwzględnienia różnorodności w analizie ryzyka bankructwa. Podejście oparte na wykorzystaniu wielu modeli umożliwia bardziej wszechstronną ocenę ryzyka upadłości i prognozowania przyszłych tendencji. Zwrócono także uwagę na dość dobre dopasowanie modeli (mimo ujęcia *ex post*) do dynamicznie zmieniających się uwarunkowań wynikających z pandemii COVID-19.

Dotychczas tworzone modele branżowe dotyczyły m.in. sektora lotniczego [Siciński 2018]. Siciński [2021] zaprezentował również badania nad systemami wczesnego ostrzeżenia przedsiębiorstw przed ryzykiem upadłości w branży transportowej. Należy podkreślić istnienie luki badawczej w obszarze zarówno tworzenia, jak i zastosowania modeli dedykowanych dla branży motoryzacyjnej, co otwiera drogę dla prowadzenia dalszych badań w zakresie podjętej problematyki badawczej. Warto podkreślić, że jakość modelu dyskryminacyjnego stworzonego dla określonej próby badawczej jest wyższa od modelu uniwersalnego, czyli takiego, który stworzony został dla podmiotów należących do wielu branż. Ze względu na ograniczoną objętość artykułu w dalszych badaniach warto skupić się także na omówieniu modeli zagranicznych dotyczących predykcji bankructwa przedsiębiorstw. Kolejnym, istotnym obszarem badawczym jest tworzenie modeli odpornych na dynamiczne zamiany gospodarcze.

## BIBLIOGRAFIA

- Altman E. I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23(4), 589-609.
- Antonowicz P. (2007) *Metody oceny i prognozowania kondycji ekonomiczno-finansowej przedsiębiorstw*. Wydawnictwo Ośrodek Doradztwa i Doskonalenia Kadr, Gdańsk.
- Czerwińska-Kayzer D., Florek J. (2012) Wykorzystanie wybranych modeli analizy dyskryminacyjnej w ocenie sytuacji finansowej przedsiębiorstw produkujących pasze. *Zeszyty Naukowe SGGW: Ekonomia i Organizacja Gospodarki Żywnościowej*, 96, 265-276.
- Emerging Markets Information Service – Polska: EMIS (2023) Sprawozdania finansowe spółek: Volkswagen Group Polska Sp. z o.o. za lata 2018-2021, Toyota Central Europe Sp. z o.o. za lata 2018-2021, Mercedes-Benz Polska Sp. z o.o. za lata 2018-2021, BMW Vertriebs GmbH Sp. z o.o. za lata 2018-2021, Renault Polska Sp. z o.o. za lata 2018-2021, Volvo Polska - Sp. z o.o. za lata 2018-2021, Porsche Inter Auto Polska Sp. z o.o. za lata 2018-2021, Kia Polska Sp. z o.o. za lata 2018-2021, Ford Polska Sp. z o.o. za lata 2018-2021, Hyundai Motor Poland Sp. z o.o. za lata 2018-2021. Pobrano z: <https://www.emis.com/php/search/searchv2?prod%5B0%5D=PL&keyword=>. [dostęp: 01.05.2023].

- Hamrol M., Chodakowski J. (2008) Prognozowanie zagrożenia finansowego przedsiębiorstwa. Wartość predykcja polskich modeli analizy dyskryminacyjnej. *Badania Operacyjne i Decyzje*, 3, 17-32.
- Hamrol M., Czajka B., Piechocki M. (2004) Upadłość przedsiębiorstwa – model analizy dyskryminacyjnej. *Przegląd Organizacji*, 6, 35-39.
- Hołda A. (2001) Prognozowanie jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej. *Rachunkowość*, 5, 306-331.
- Jajuga K. (2019) Nauki ekonomiczne – dylematy klasyfikacji dyscyplin. Tendencje zmian. [w:] Gorynia M. (red. nauk.) *Ewolucja nauk ekonomicznych. Jedność a różnorodność. Relacje do innych nauk. Problemy klasyfikacyjne*. Polska Akademia Nauk, Warszawa, 140-150.
- Kisielińska J., Borkowski B., Czech K., Górska A., Koszela G, Krawiec M., Landmesser-Rusek J., Ochnio L., Pietrych Ł., Pietrzykowski R., Wasilewska E., Zielińska-Sitkiewicz M. (2021) *Wielowymiarowa analiza danych w ekonomice rolnictwa*. Wydawnictwo SGGW, Warszawa.
- Kisielińska J., Waszkowski A. (2010) Polskie modele do prognozowania bankructwa przedsiębiorstw i ich weryfikacja. *Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego: Ekonomika i Organizacja Gospodarki Żywnościowej*, 82, 17-31.
- Mączyńska E., Zawadzki M. (2006) Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw. *Ekonomista*, 2, 1-24, <http://www.pte.pl/pliki/2/12/Ekonomista%2025%2002%2006ostfragment.pdf> [dostęp: 2.09.2023].
- Prusak B. (2005) *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstwa*. Centrum Doradztwa i Informacji Difin, Warszawa.
- Siciński J. (2018) Wybrane systemy wczesnego ostrzegania w predykcji upadłości europejskich linii lotniczych w latach 2009-2017. *Zarządzanie i Finanse*, 16(1), 263-277.
- Siciński J. (2021) System wczesnego ostrzegania przedsiębiorstw przed ryzykiem upadłości na przykładzie branży transportowej. *Centrum Myśli Strategicznych*. Sopot.
- Wierzbą D. (2000) *Wczesne wykrywanie przedsiębiorstw zagrożonych upadłością na podstawie analizy wskaźników finansowych – teoria i badania empiryczne*. Wydawnictwo Wyższej Szkoły Ekonomiczno-Informatycznej w Warszawie, Warszawa.
- Wysocki F., Kozera A. (2012) Wykorzystanie analizy dyskryminacyjnej w ocenie ryzyka upadłości przedsiębiorstw przemysłu mięsnego. *Journal of Agribusiness and Rural Development*, nr 4(26), 167-182.
- Zielińska-Sitkiewicz M. (2012) Zastosowanie metod wielowymiarowej analizy dyskryminacyjnej do oceny kondycji firm deweloperskich. *Metody Ilościowe w Badaniach Ekonomicznych*, 13(3), 288-301.

#### USING DISCRIMINANT ANALYSIS TO EVALUATE THE RISK OF CORPORATE BANKRUPTCY IN THE MOTORIZATION SECTOR DURING THE COVID-19 PANDEMIC

**Abstract:** This article presents the effectiveness of seven discriminant models in assessing the risk of corporate bankruptcy in the motorization

sector during the COVID-19 pandemic. The study uses the following methods: discriminant analysis, descriptive analysis, and comparative analysis. This study indicates that in the research sample, the number of enterprises at risk of bankruptcy and in the so-called "gray zone", according to the model, has increased significantly during the COVID-19 pandemic. Only two models did not present a risk of bankruptcy for the surveyed companies during the analyzed period.

**Keywords:** discriminant analysis, corporate bankruptcy prediction models, corporate finance, motorization sector, the COVID-19 pandemic

**JEL classification:** G01, G17, G33