Cena 15,00 zł
(VAT 8%)

Indeks 381306
e-ISSN 2543-8476
PL ISSN 0043-518X

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

MAJ / MAY
ROCZNIK / VOLUME 68

2023 | 5

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

POLSKIE TOWARZYSTWO STATYSTYCZNE
POLISH STATISTICAL ASSOCIATION

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

MAJ / MAY
ROCZNIK / VOLUME 68

2023 | 5 (744)

ADRES REDAKCJI / EDITORIAL OFFICE ADDRESS

Główny Urząd Statystyczny / Statistics Poland, al. Niepodległości 208, 00-925 Warszawa
tel./phone +48 22 608 32 25, e-mail: redakcja.ws@stat.gov.pl

Redakcja językowa: Wydział Czasopism Naukowych, Główny Urząd Statystyczny
Language editing: Scientific Journals Division, Statistics Poland

Redakcja techniczna, skład i łamanie, opracowanie materiałów graficznych, korekta: Zakład Wydawnictw Statystycznych – zespół pod kierunkiem Macieja Adamowicza
Technical editing, typesetting, preparation of graphic materials, proofreading: Statistical Publishing Establishment – team supervised by Maciej Adamowicz

Zakład Wydawnictw Statystycznych

Druk i oprawa / Printed and bound by:
Zakład Wydawnictw Statystycznych / Statistical Publishing Establishment
al. Niepodległości 208, 00-925 Warszawa, zws.stat.gov.pl

Wersja elektroniczna, stanowiąca wersję pierwotną czasopisma, jest dostępna na ws.stat.gov.pl
The primary version of the journal, issued in electronic form, is available at ws.stat.gov.pl

Indeks 381306

Informacje w sprawie sprzedaży czasopisma / Sales of the journal:
Zakład Wydawnictw Statystycznych / Statistical Publishing Establishment
tel./phone +48 22 608 32 10, +48 22 608 38 10
Prenumerata jest prowadzona przez / Subscription is available at RUCH S.A.
Zamówienia na prenumeratę można składać na stronie / Subscriptions can be ordered at
www.prenumerata.ruch.com.pl

Zam. 140/2023 – nakład 285 egz.

SPIS TREŚCI
CONTENTS

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), IV–V
Od redakcji / From the editorial team

OD REDAKCJI

W majowym numerze „Wiadomości Statystycznych. The Polish Statistician" znajdą Państwo trzy artykuły o charakterze empirycznym.

Praca dr Barbary Batóg i dr Katarzyny Wawrzyniak *Stabilność grupowania powiatów województwa zachodniopomorskiego pod względem sytuacji gospodarczej. Co zmieniła pandemia COVID-19?* dotyka ważnego problemu, jakim jest wpływ powszechnie zaistniałych niekorzystnych czynników, takich jak pandemia, na stabilność zróżnicowania rozwoju gospodarczego jednostek przestrzennych. Autorki przedstawiają wyniki badania stabilności grupowania powiatów woj. zachodniopomorskiego ze względu na ich sytuację gospodarczą w latach 2007–2021. Grupowanie przeprowadzają na podstawie wartości miary syntetycznej, skonstruowanej z wykorzystaniem zmiennych charakteryzujących sytuację przedsiębiorstw, sytuację na rynku pracy i sytuację finansową powiatów w poszczególnych latach, oraz metody trzech średnich. Obliczenia wykonują na danych zaczerpniętych z Banku Danych Lokalnych GUS. Z badania wynika, że grupowanie powiatów cechowało się bardzo dużą stabilnością, co oznacza, że pandemia COVID-19 zasadniczo nie zmieniła sytuacji gospodarczej tych obszarów.

Mgr Dawid Giemza w artykule *Ocena zmian komponentów wskaźnika NEET w krajach Unii Europejskiej z zastosowaniem testu* T^2 *Hotellinga* analizuje wielkość zmian komponentów wskaźnika NEET (młodzieży niepracującej, nieuczącej się i niedokształcającej się – ang. *neither in employment nor in education and training*) w podziale według cech społeczno-demograficznych i ekonomicznych. Badanie dotyczy młodzieży w krajach UE w latach 2019 i 2021, czyli roku poprzedzającym wybuch pandemii COVID-19 i pełnym roku jej trwania. Autor posługuje się wielowymiarowym wnioskowaniem statystycznym, opartym na testach porównań wielokrotnych. Obliczeń dokonuje na podstawie danych Eurostatu, z zastosowaniem testu T^2 Hotellinga dla pomiarów zależnych. Dowodzi, że porównanie kształtowania się komponentów pozwala na wychwycenie zmian, których nie uwidacznia wskaźnik w ujęciu ogólnym. Takie podejście umożliwia zatem bardziej szczegółową ocenę sytuacji osób zaliczanych do kategorii NEET.

Artykuł mgr. Adama Idczaka i dr. hab. Jerzego Korzeniewskiego, prof. UŁ, pt. *New algorithm for determining the number of features for the effective sentiment-classification of text documents* reprezentuje interdyscyplinarne podejście do statystyki i traktuje o analizie sentymentu, czyli wydźwięku emocjonalnego, dokumentów tekstowych. Autorzy przedstawiają nową technikę analizy tego rodzaju opartą na modelach unigramowych, możliwą do zastosowania w dowolnej metodzie klasyfikacji dokumentów ze względu na ich wydźwięk. Proponowana technika – której zaletami są intuicyjność i prostota obliczeniowa – polega na niezależnym od klasyfikatora doborze cech, co skutkuje zmniejszeniem rozmiaru ich przestrzeni. Omawiana technika opiera się na nowatorskim algorytmie ustalania liczby terminów wystarczających do efektywnej klasyfikacji, który bazuje na analizie korelacji pomiędzy pojedynczymi cechami dokumentów a ich wydźwiękiem. Autorzy weryfikują przydatność proponowanej techniki, wykorzystując naiwny klasyfikator Bayesa i regresję logistyczną. Analizują trzy zbiory dokumentów napisanych w języku polskim składające się z ponad tysiąca opinii klientów jednego z banków w Polsce. Uzyskują lepsze wyniki niż w przypadku podejścia standardowego: kilkunastokrotne zmniejszenie liczby terminów przy zastosowaniu proponowanej techniki na ogół poprawia jakość klasyfikacji.

Ponadto polecamy Państwa uwadze książki, które omawia Dorota Kierska w rubryce *Nowości wydawnicze w zbiorach Centralnej Biblioteki Statystycznej*, oraz najnowsze publikacje GUS zestawione przez Joannę Sadowy w rubryce *Wydawnictwa GUS*.

Życzymy miłej lektury.

FROM THE EDITORIAL TEAM

The May issue of *Wiadomości Statystyczne. The Polish Statistician* features three articles of an empirical character.

In their paper, Barbara Batóg, PhD, and Katarzyna Wawrzyniak, PhD, entitled *Stability of the grouping of powiats in Zachodniopomorskie Voivodship in relation to the economic situation. What has the COVID-19 pandemic changed?* address an important issue of the impact of commonly occurring unfavourable factors, such as the pandemic, on the stability of the diversity of spatial units' economic development. The authors present the results of their study examining the stability of the grouping of powiats of Zachodniopomorskie Voivodship in relation to their economic situation in the years 2007–2021. The grouping is based on the value of the synthetic measure, constructed with the use of variables characterising the situation of both enterprises and the labour market, as well the financial condition of the powiats in the particular years, and the three averages method. The data used for the calculations have been provided by the Local Data Bank of Statistics Poland. The research demonstrates that the grouping was very stable, suggesting that in general the COVID-19 pandemic did not change the economic situation of these areas.

Dawid Giemza, MSc, in the article *Assessment of the changes in the NEET rate structure in EU countries based on Hotelling's* T^2 *test* analyses the volume of changes in the NEET (*neither in employment nor in education and training*) rate by socio-demographic and economic characteristics of young people in EU countries in the years 2019 and 2021, i.e. in the year preceding the outbreak of the COVID-19 pandemic and a full year of its duration. The author applies multivariate statistical inference based on multiple comparison tests. Eurostat data serve as a basis for the calculations involving the use of Hotelling's T^2 test for dependent groups. The study indicates that the comparison of the NEET rate structure makes it possible to capture changes which the overall rate otherwise fails to show. Thus, such an approach allows a more detailed assessment of the situation of those who belong to the NEET category.

The article by Adam Idczak, MSc, and Jerzy Korzeniewski, PhD, DSc, professor at the University of Lodz, entitled *New algorithm for determining the number of features for the effective sentiment-classification of text documents* represents an interdisciplinary approach to statistics and relates to the sentiment analysis of text mining. The authors describe a new technique of this type of analysis based on unigram models, applicable to any type of a document sentiment-classification method. The proposed technique, whose advantages are intuitiveness and computational simplicity, involves feature selection independent of a classifier, which reduces the size of the feature space. The technique is based on a novel algorithm for the determination of the number of features sufficient for an effective classification. The algorithm consists in the analysis of correlations between single document features and document labels. The authors verify the usefulness of the proposed technique by means of the naive Bayes classifier and logistic regression. They analyse three sets of documents written in Polish, encompassing over 1,000 customer reviews of one of the banks in Poland. The authors obtain better results than in the case of the standard approach: an over 10-fold reduction of the number of terms through the proposed technique in most cases improves the quality of the classification.

We would also like to recommend the books reviewed by Dorota Kierska in the *New publications in the Central Statistical Library* column, as well as the latest Statistics Poland releases compiled by Joanna Sadowy in the *Publications of Statistics Poland* column.

We wish you pleasant reading.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), 1–19 https://doi.org/10.59139/ws.2023.05.1
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 29.11.2022, poprawiony/revised: 31.03.2023, zaakceptowany/accepted: 17.04.2023

Stabilność grupowania powiatów województwa zachodniopomorskiego pod względem sytuacji gospodarczej. Co zmieniła pandemia COVID-19?

Barbara Batóg[a], Katarzyna Wawrzyniak[b]

Streszczenie. W artykule przedstawiono wyniki badania stabilności grupowania powiatów woj. zachodniopomorskiego ze względu na sytuację gospodarczą w latach 2007–2021. Celem badania było uzyskanie odpowiedzi na pytanie, czy pandemia COVID-19 wpłynęła na sytuację gospodarczą powiatów. Dane do obliczeń zaczerpnięto z Banku Danych Lokalnych GUS. Dla każdej jednostki wyznaczono wartości miary syntetycznej w poszczególnych latach z wykorzystaniem wybranych zmiennych charakteryzujących sytuację przedsiębiorstw, sytuację na rynku pracy oraz sytuację finansową powiatów. Następnie pogrupowano powiaty za pomocą metody trzech średnich (osobno dla każdego roku) i na podstawie wybranych miar podobieństwa oceniono stabilność grupowania powiatów w analizowanym okresie rok do roku. Wykazano, że grupowanie powiatów charakteryzowało się bardzo dużą stabilnością, co oznacza, że pandemia zasadniczo nie zmieniła ich sytuacji gospodarczej.

Słowa kluczowe: grupowanie, badanie stabilności, sytuacja gospodarcza, powiaty, COVID-19
JEL: C38, R11

Stability of the grouping of powiats in Zachodniopomorskie Voivodship in relation to the economic situation. What has the COVID-19 pandemic changed?

The paper presents the results of a study examining the stability of the grouping of Polish powiats (communes) in Zachodniopomorskie Voivodship in relation to the economic situation in the years 2007–2021. The aim of the research was to determine whether the COVID-19 pandemic affected the economic situation of the powiats. The data used for the calculations were provided by the Local Data Bank of Statistics Poland. On the basis of selected variables characterising the situation of enterprises and the labour market, as well as the financial

[a] Uniwersytet Szczeciński, Wydział Ekonomii, Finansów i Zarządzania, Instytut Ekonomii i Finansów, Polska / University of Szczecin, Faculty of Economics, Finance and Management, Institute of Economics and Finance, Poland. ORCID: https://orcid.org/0000-0001-9236-7405. Autor korespondencyjny / Corresponding author: e-mail: barbara.batog@usz.edu.pl.
[b] Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, Wydział Ekonomiczny, Polska / West Pomeranian University of Technology in Szczecin, Faculty of Economics, Poland.
ORCID: https://orcid.org/0000-0003-4161-3877. E-mail: katarzyna.wawrzyniak@zut.edu.pl.

condition of powiats, the values of the synthetic measure were determined for each powiat in the particular years. Subsequently, the powiats were grouped according to the three averages method (separately for each year). Based on selected similarity measures, the stability of the grouping of powiats was assessed year-to-year in the analysed period. The research demonstrated that the grouping was very stable, which means that in general the pandemic did not affect the economic situation of the powiats.

Keywords: grouping, analysis of stability, economic situation, powiats, COVID-19

1. Wprowadzenie

Pandemia COVID-19 znacząco wpłynęła na funkcjonowanie gospodarki oraz jakość życia ludności na całym świecie. Skutki pandemii były przedmiotem licznych badań, koncentrujących się przede wszystkim na zjawiskach gospodarczych i demograficznych w ujęciu globalnym (świat, regiony świata, UE czy poszczególne kraje). Uzyskane wyniki świadczą przede wszystkim o negatywnych skutkach pandemii, chociaż istnieją także prace badawcze, w których wpływ pandemii został oceniony pozytywnie.

Wysokińska (2022) zauważyła, że skutki pandemii były szczególnie dotkliwe dla handlu. W UE w okresie od marca do kwietnia 2020 r. eksport ogółem spadł ze 176 mld do 125 mld euro, czyli o prawie 29%, a import ogółem – ze 148 mld do 125 mld euro – o 15%. W Polsce w 2020 r. w stosunku do 2019 r. odnotowano wzrost eksportu ogółem o 0,7% (do krajów UE – o 0,3%) i spadek importu ogółem o 3,8% (z krajów UE – o 6,1%). Jak podaje Witkowska (2022), w 2020 r. duży spadek zaobserwowano również w przypadku bezpośrednich inwestycji zagranicznych (BIZ) w skali globalnej i regionalnej. W całej Europie obniżka ta wyniosła 80%, a w krajach UE – 73%.

Negatywne skutki pandemii bardzo silnie odczuła również branża turystyczna. Roman i in. (2022) przytoczyli dane, z których wynika, że przed pandemią sektor podróży i turystyki tworzył 10,6% wszystkich miejsc pracy i 10,4% światowego PKB, a w wyniku kryzysu pandemicznego i znacznego ograniczenia mobilności turystów w 2020 r. zmniejszył swój udział w PKB do 5,5%. We wszystkich badanych krajach sektor turystyczny w 2020 r. poniósł straty, przy czym największe odnotowano m.in. w Hiszpanii, na Cyprze i na Węgrzech. Wpływ pandemii COVID-19 na sektor turystyczny na przykładzie m.in. Czarnogóry przedstawił Dziuba (2022). Autor wykazał, że w wyniku pandemii udział dochodów z turystyki w PKB w tym kraju spadł o 16%, a bezrobocie w grupie kobiet i młodzieży zatrudnionych w sektorach powiązanych z działalnością turystyczną wzrosło o 30%.

Wpływ pandemii COVID-19 na rynek pracy został opisany m.in. w opracowaniach Astorquizy Bustosa i in. (2022), Bieszk-Stolorz i Dmytrowa (2022) oraz Kryeziu i in. (2022). Astorquiza Bustos i in. wykazali, że w Kolumbii pandemia spo-

wodowała przede wszystkim utratę miejsc pracy, a tym samym wzrost bezrobocia oraz przyczyniła się do skrócenia czasu pracy i obniżki wynagrodzeń. Z badania przeprowadzonego przez Bieszk-Stolorz i Dmytrowa wynika, że wpływ pandemii na rynek pracy w UE był zróżnicowany, a wykryte różnice są konsekwencją różnego stopnia rozwoju rynków pracy w poszczególnych krajach. Kryeziu i in. stwierdzili, że w Kosowie w wyniku pandemii spadek dochodów odnotowało 92,2% małych i średnich firm bez względu na rodzaj działalności.

Negatywne skutki pandemii nie ominęły również rynków finansowych. Hamulczuk i Idzik (2022) zbadali testy koniunktury przeprowadzone przez GUS i Kantar Polska w latach 2004–2021 i zaobserwowali występowanie negatywnego związku pomiędzy siłą obostrzeń pandemicznych a koniunkturą w bankowości, szczególnie w pierwszej fazie kryzysu. Testy koniunktury pokazują, że w czasie kryzysu wywołanego pandemią COVID-19 koniunktura w sektorze bankowym pogorszyła się bardziej niż podczas kryzysu finansowego z lat 2008–2009. Z kolei Landmesser-Rusek (2022) wykazała, że kursy walut kształtowały się odmiennie w okresie przed pandemią oraz w jej pierwszej i drugiej fazie. Obszerne rozważania na temat skutków pandemii w Polsce (m.in. na rynku pracy i rynku finansowym) przedstawiono w książce pod redakcją Mastalerz-Kodzis i Zeug-Żebro (2023).

Zwiększona śmiertelność i konsekwencje zdrowotne są najbardziej oczywistymi negatywnymi skutkami pandemii COVID-19 (Wysokińska, 2022). Z badania opartego na danych pochodzących z ankiety *COVID Impact Survey*, przeprowadzonej w 2020 r. (w trakcie pierwszej fali pandemii) przez organizację Data Foundation w Stanach Zjednoczonych wśród osób dorosłych, wynika, że izolacja, zdalny tryb nauczania i pracy oraz mniejsza aktywność fizyczna przyczyniają się do pogarszania się stanu zdrowia psychicznego (Ptak-Chmielewska i in., 2022).

Niektórzy badacze zwrócili uwagę także na pozytywne aspekty pandemii, jak wzrost handlu elektronicznego. Wysokińska (2022) podaje, że udział sprzedaży detalicznej online w całkowitej sprzedaży detalicznej zwiększył się z 16% w 2019 r. do 19% w 2020 r. Przyczyny takiego stanu rzeczy omówiono m.in. w pracach Brougha i Martin (2021), Egera i in. (2021) oraz Machovej i in. (2022). Czernek-Marszałek i Piotrowski (2022), na podstawie ankiety skierowanej do przedsiębiorstw, wykazali, że pandemia przyspieszyła upowszechnienie rozwiązań cyfrowych i spowodowała zmianę sposobu ich stosowania.

W Polsce pierwsze przypadki zakażenia wirusem SARS-CoV-2 odnotowano w marcu 2020 r. i wtedy też wprowadzono pierwsze obostrzenia (zawieszenie lotów do Włoch i Chin; odwołanie imprez masowych; zamknięcie szkół, przedszkoli i placówek oświatowych; wprowadzenie pracy zdalnej przez część pracodawców), które z dnia na dzień stawały się coraz radykalniejsze. Pod koniec marca rząd podjął decyzję m.in. o zamknięciu zakładów fryzjerskich i kosmetycznych, ograniczeniu liczby

klientów w sklepach oraz zakazie wstępu do lasów i parków. Stopniowe znoszenie restrykcji rozpoczęto w czerwcu 2020 r., jednak już w październiku tego roku – wraz ze wzrostem liczby zachorowań (druga fala) – ponownie wprowadzono liczne obostrzenia, które zaczęto stopniowo luzować dopiero od stycznia 2021 r., gdy rozpoczęły się szczepienia przeciwko COVID-19. W celu ograniczenia gospodarczych skutków pandemii rząd zaproponował pakiet działań tworzących tarczę antykryzysową (Ustawa z dnia 16 kwietnia 2020 r. o szczególnych instrumentach wsparcia w związku z rozprzestrzenianiem się wirusa SARS-CoV-2).

Pomimo wprowadzenia tarczy antykryzysowej sytuacja gospodarcza w Polsce w 2020 r. pogorszyła się w stosunku do 2019 r., co potwierdzają dane GUS dostępne w Banku Danych Makroekonomicznych[1], dotyczące takich kategorii, jak:

- PKB (ceny stałe) – spadek o 2%;
- nakłady inwestycyjne (ceny stałe) – spadek o 5%;
- produkcja sprzedana przemysłu ogółem (ceny stałe) – spadek o 1,9%;
- pracujący w gospodarce narodowej (stan na koniec roku) – spadek z 16 120 600 do 14 789 100, czyli o 8,3%;
- stopa bezrobocia rejestrowanego ogółem (stan na koniec roku) – wzrost o 1,1 p.proc.

Jednak już rok później, gdy ze względu na szczepienia znacznie poluzowano ograniczenia, sytuacja gospodarcza się poprawiła, co przełożyło się na pozytywną zmianę wartości wskaźników makroekonomicznych w 2021 r. w stosunku do 2020 r.:

- PKB (ceny stałe) – wzrost o 6,8%;
- nakłady inwestycyjne (ceny stałe) – wzrost o 1%;
- produkcja sprzedana przemysłu ogółem (ceny stałe) – wzrost o 14,4%;
- pracujący w gospodarce narodowej (stan na koniec roku) – wzrost z 14 789 100 do 15 002 600, czyli o 1%;
- stopa bezrobocia rejestrowanego ogółem (stan na koniec roku) – spadek o 0,9 p.proc.

Podobne zmiany wyżej wymienionych wskaźników w latach 2019–2021 zaobserwowano w woj. zachodniopomorskim.

Przegląd publikacji oraz analiza danych GUS skłoniły autorki do przeprowadzenia badania mającego na celu uzyskanie odpowiedzi na pytanie, czy pandemia COVID-19 zmieniła sytuację gospodarczą na niższym poziomie agregacji, czyli w powiatach woj. zachodniopomorskiego.

[1] https://bdm.stat.gov.pl.

2. Metoda badania

Badanie dotyczyło sytuacji gospodarczej powiatów woj. zachodniopomorskiego w latach 2007–2021, którą scharakteryzowano za pomocą dziewięciu wybranych zmiennych (zestawienie). Dane do obliczeń zaczerpnięto z Banku Danych Lokalnych (BDL) GUS[2]. Zmienne dobrano w taki sposób, aby przedstawiały sytuację powiatów w różnych obszarach gospodarki. Ponadto w każdym roku sprawdzono ich zróżnicowanie oraz korelację pomiędzy nimi. Okazało się, że prawie we wszystkich latach wybrane zmienne charakteryzowały się wystarczająco wysokim zróżnicowaniem (oprócz przeciętnego wynagrodzenia brutto) oraz odpowiednio niskim skorelowaniem: na 540 obliczonych współczynników korelacji pomiędzy wszystkimi zmiennymi i w poszczególnych latach tylko 25 było w przedziale (0,8; 0,9), a 3 były w przedziale (–0,9; –0,8) w różnych latach i dla różnych par zmiennych (Młodak, 2006). Charakter zmiennych został określony na podstawie ich wpływu na sytuację gospodarczą.

Zestawienie zmiennych wykorzystanych w badaniu

Nazwa	Charakter
Sytuacja przedsiębiorstw	
Przeciętne miesięczne wynagrodzenie brutto w zł	stymulanta
Liczba podmiotów gospodarki narodowej na 10 tys. mieszkańców w wieku produkcyjnym	stymulanta
Liczba nowo zarejestrowanych podmiotów gospodarki narodowej na 1000 mieszkańców	stymulanta
Sytuacja finansowa powiatów	
Dochody budżetów powiatów z tytułu podatku dochodowego od osób fizycznych na mieszkańca w zł	stymulanta
Dochody budżetów powiatów z tytułu podatku dochodowego od osób prawnych na mieszkańca w zł	stymulanta
Wydatki majątkowe inwestycyjne z budżetu ogółem na mieszkańca w zł	stymulanta
Sytuacja na rynku pracy	
Stopa bezrobocia rejestrowanego w %	destymulanta
Liczba pracujących na 1000 ludności ogółem	stymulanta
Liczba ofert pracy na koniec roku na 1000 bezrobotnych	stymulanta

Źródło: opracowanie własne na podstawie danych z BDL.

Badanie składało się z trzech etapów. Na pierwszym pogrupowano powiaty – osobno dla każdego roku badanego okresu – na podstawie wartości miary syntetycznej obliczonej z wykorzystaniem wybranych zmiennych. W literaturze można znaleźć wiele propozycji zastosowania różnych postaci miar syntetycznych (np. Gatnar i Walesiak, 2004; Kukuła, 2000; Łuniewska i Tarczyński, 2006; Markowska, 2016; Młodak, 2006; Pawełek, 2008; Sokołowski i Markowska, 2017; Strahl, 2006). W ba-

[2] https://bdl.stat.gov.pl/bdl.

daniu wykorzystano normalizację zmiennych za pomocą unitaryzacji zerowanej, opisanej wzorem (1), a miarę syntetyczną wyznaczono jako średnią ze znormalizowanych wartości zmiennych, obliczaną ze wzoru (2). Minima i maksima zostały wyznaczone osobno dla każdego roku, ponieważ podział powiatów na grupy również przeprowadzono dla każdego roku. W związku z tym wzory (1) i (2) dotyczą tego samego roku.

Normalizacja ma zatem postać:

$$z_{ij} = \begin{cases} \dfrac{x_{ij} - min_j}{max_j - min_j} & \text{dla} \quad \text{stymulant,} \\ \dfrac{x_{ij} - max_j}{min_j - max_j} & \text{dla} \quad \text{destymulant,} \end{cases} \tag{1}$$

a miernik kompleksowy:

$$ms_i = \frac{1}{K}\sum_{j=1}^{K} z_{ij}, \tag{2}$$

gdzie:

x_{ij} – wartość j-ej zmiennej dla i-tego obiektu przed normalizacją,
z_{ij} – wartość j-ej zmiennej dla i-tego obiektu po normalizacji,
ms_i – wartość miary syntetycznej dla i-tego obiektu.
i – numer obiektu, $i = 1, \dots, n$, a n– liczba obiektów,
j – numer zmiennej, $j = 1, \dots, K$, a K– liczba zmiennych.

Do wydzielenia grup powiatów różniących się wartością analizowanych zmiennych w badanych latach zastosowano metodę trzech średnich, która polega na tym, że punktami podziału są: średnia dla całej zbiorowości (średnia ogólna), średnia z wartości poniżej średniej ogólnej oraz średnia z wartości powyżej średniej ogólnej. W ten sposób uzyskuje się cztery grupy obiektów.

Na drugim etapie wyznaczono miary podobieństwa grupowania powiatów w poszczególnych latach i na podstawie wartości tych miar oceniono stabilność grupowania, a tym samym zidentyfikowano te okresy i jednostki, w których stabilność została zakłócona. Wykorzystano dwie miary:

- miarę zgodności procentowej (Wędrowska, 2012):

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i \cdot 100\%, \tag{3}$$

gdzie:
n – liczba obiektów,
i – numer obiektu,
$c_i = 1$, gdy obiekt i jest w tej samej grupie w dwóch kolejnych latach,
$c_i = 0$, gdy obiekt i nie jest w tej samej grupie w dwóch kolejnych latach.

Miara zgodności procentowej grupowania daje odpowiedź na pytanie, jaka część obiektów (tu: powiatów) nie zmieniła przynależności do grupy w grupowaniach w kolejnych latach;

- indeks Randa (Rand, 1971):

$$CR = \frac{1}{\frac{n(n-1)}{2}} \sum_{r,s;r<s} c_{rs}, \tag{4}$$

gdzie:
n – liczba obiektów,
r, s – numery obiektów,
$c_{sr} = 1$, gdy obiekty s i r są w tej samej grupie w dwóch kolejnych latach lub nie są w tej samej grupie w dwóch kolejnych latach,
$c_{sr} = 0$, gdy obiekty s i r są w tej samej grupie w jednym roku i nie są w tej samej grupie w następnym roku.

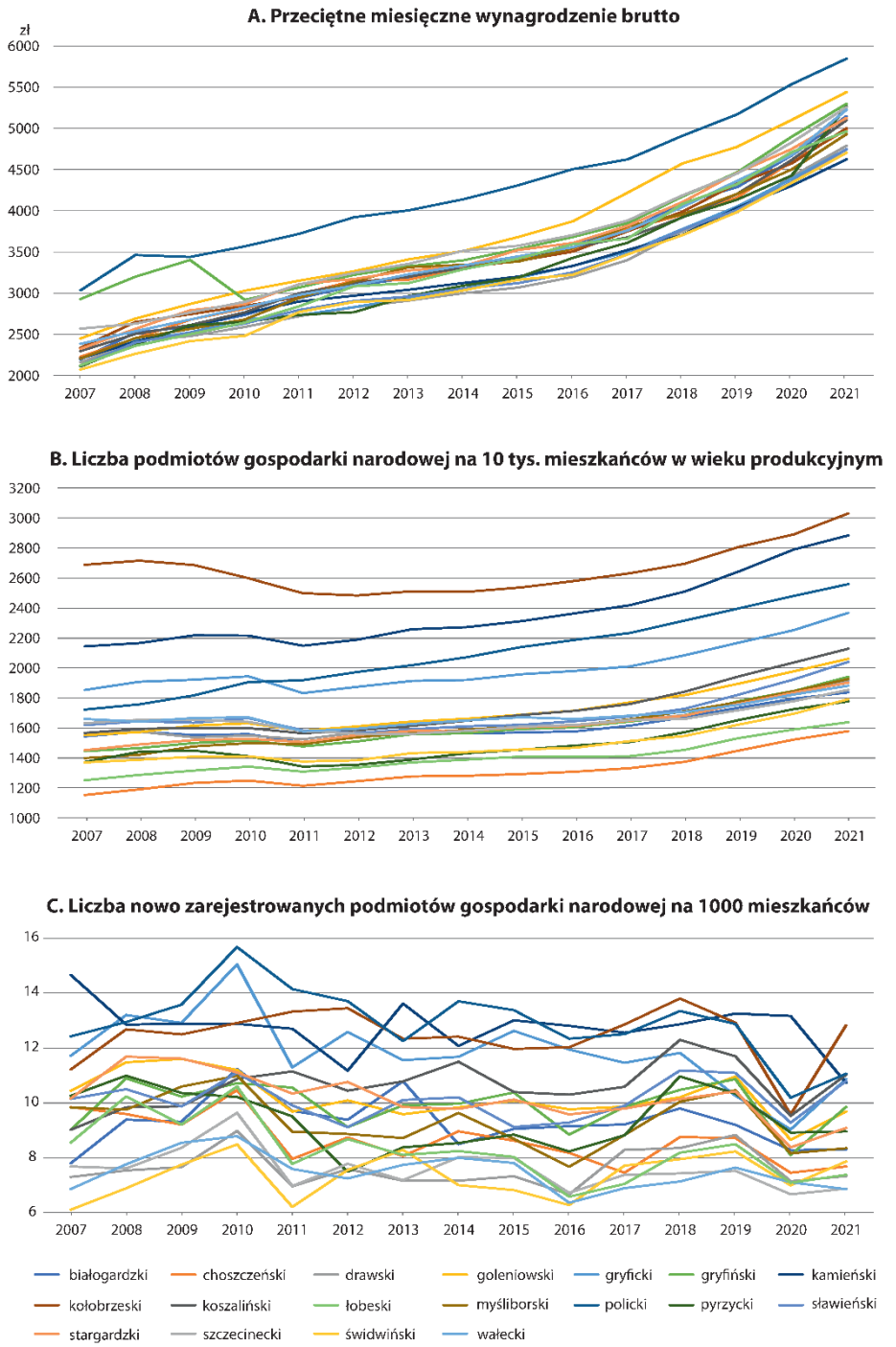
Indeks Randa różni się od miary zgodności procentowej tym, że dotyczy pary obiektów (tu: powiatów), więc uwzględnia to, czy pary powiatów znajdują się w tej samej grupie w dwóch grupowaniach w kolejnych latach, czy też jeden z nich zmienił swoją przynależność do grupy.

Na ostatnim etapie wskazano obszary (zmienne), które najbardziej przyczyniły się do zakłócenia stabilności grupowania powiatów ze względu na ich sytuację gospodarczą.

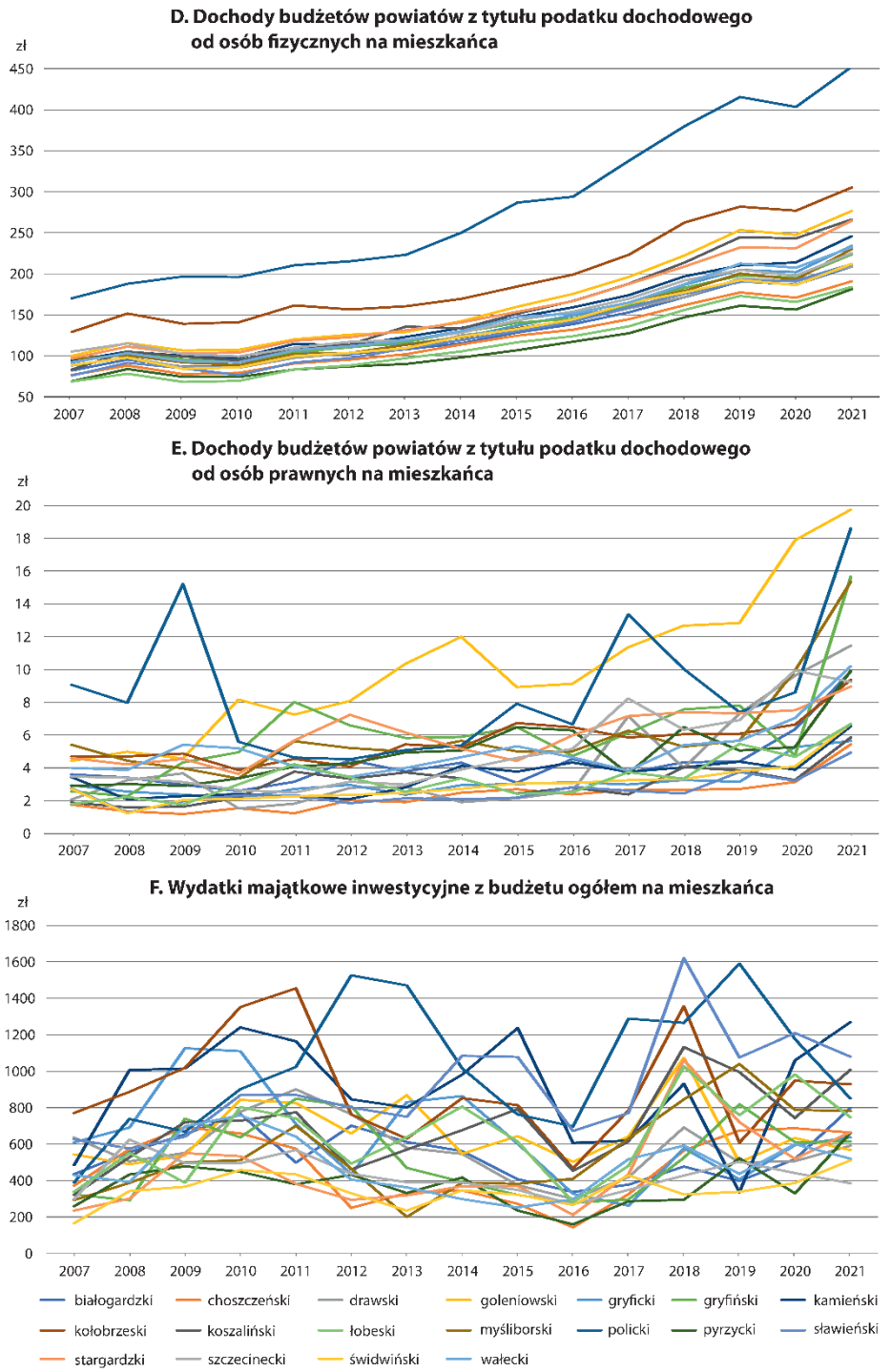
3. Wyniki badania

Badaną zbiorowość stanowiło 21 powiatów woj. zachodniopomorskiego, w tym trzy miasta na prawach powiatu: Szczecin, Koszalin i Świnoujście. Wstępne obliczenia wykazały, że dla tych trzech miast wartość każdej z analizowanych zmiennych, a co za tym idzie wartość miary syntetycznej, znacznie odbiegała od wartości w pozostałych powiatach. Te jednostki w każdym roku tworzyły odrębną grupę, charakteryzującą się najlepszą sytuacją gospodarczą, co utrudniało identyfikację prawidłowości w zakresie dynamiki zmiennych w pozostałych powiatach. W związku z tym zdecydowano się na pominięcie miast na prawach powiatów w dalszej analizie.

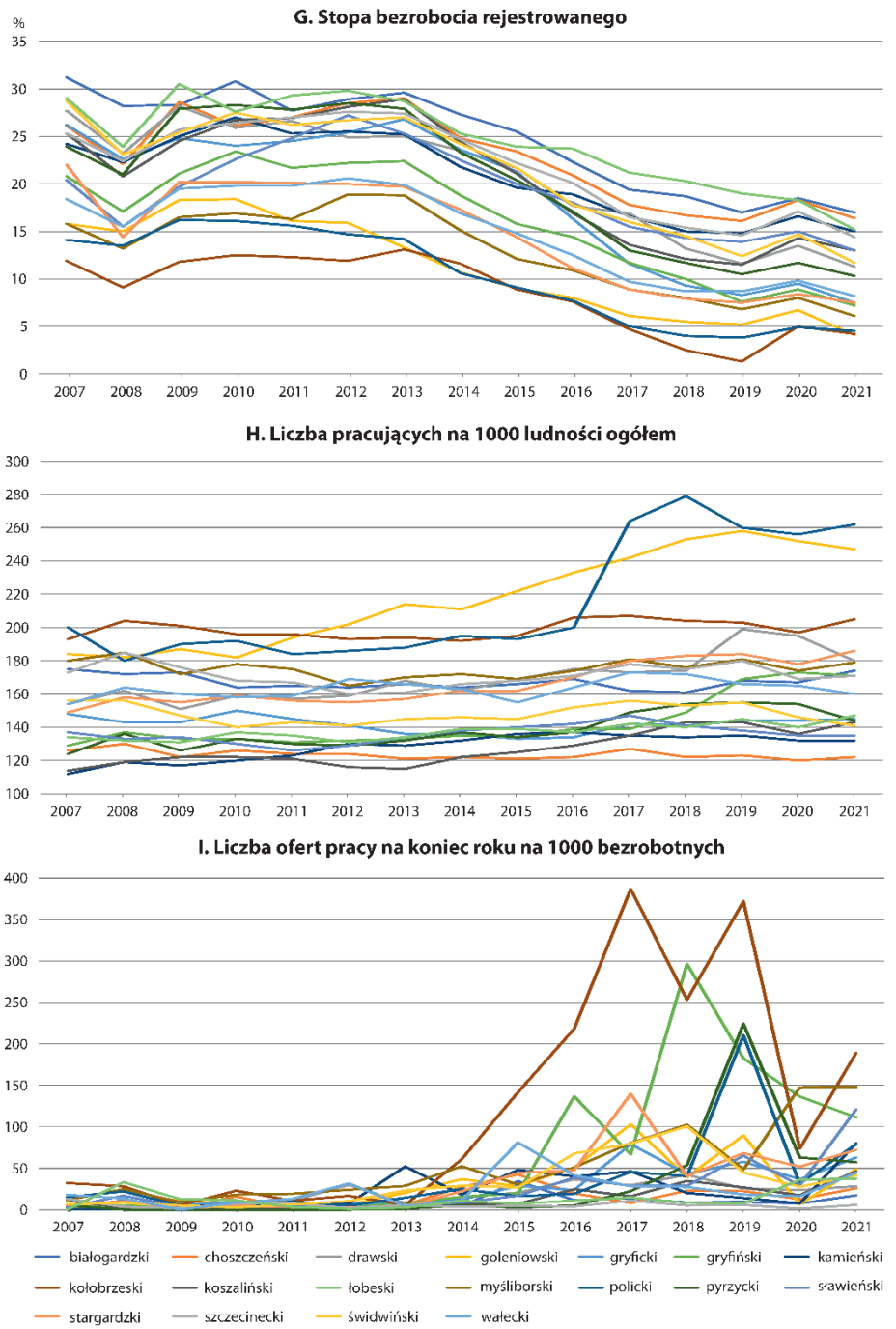
Wykr. 1. Zmienne obrazujące sytuację gospodarczą powiatów



Wykr. 1. Zmienne obrazujące sytuację gospodarczą powiatów (cd.)



Wykr. 1. Zmienne obrazujące sytuację gospodarczą powiatów (dok.)



Źródło: opracowanie własne na podstawie danych z BDL.

Na wykr. 1 przedstawiono, jak kształtowały się wartości analizowanych zmiennych w 18 powiatach woj. zachodniopomorskiego w latach 2007–2021.

W całym badanym okresie przeciętne miesięczne wynagrodzenie brutto wzrosło we wszystkich powiatach (wykr. 1A). Najwyższy poziom osiągnęło w pow. polickim, a najniższy – w pow. świdwińskim.

Od 2011 r. systematycznie zwiększała się też liczba podmiotów gospodarki narodowej przypadających na 10 tys. mieszkańców w wieku produkcyjnym (wykr. 1B). Najwyższą wartością tej zmiennej charakteryzował się pow. kołobrzeski, a najniższą – pow. choszczeński.

Na wykr. 1C ukazano szeregi czasowe liczby nowo zarejestrowanych podmiotów gospodarki narodowej przypadających na 1000 mieszkańców. Można zauważyć wyraźny spadek rozpatrywanej wielkości w 2020 r. w porównaniu z 2019 r. w 17 powiatach (jedynie w pow. kamieńskim poziom tej zmiennej w porównywanych latach był stały). Oznacza to, że pandemia COVID-19 miała negatywny wpływ na powstawanie nowych podmiotów. Największy spadek wartości badanej cechy odnotowano w powiatach kołobrzeskim i policki, a najmniejszy – w pow. wałeckim. Z kolei w 2021 r. w porównaniu z 2020 r. w większości jednostek zaobserwowano wzrost poziomu tej zmiennej – z wyjątkiem powiatów kamieńskiego, stargardzkiego i wałeckiego (spadek) oraz białogardzkiego i myśliborskiego (bez zmian).

Dochody budżetów powiatów z tytułu podatku dochodowego od osób fizycznych przypadające na mieszkańca (wykr. 1D) w latach przed pandemią systematycznie wzrastały we wszystkich jednostkach i również we wszystkich odnotowano spadek wartości tej zmiennej w 2020 r. w porównaniu z 2019 r. (wpływ pandemii) i ponowny wzrost w 2021 r. w porównaniu z 2020 r.

W przypadku dochodów budżetów powiatów z tytułu podatku dochodowego od osób prawnych przypadających na mieszkańca (wykr. 1E) w analizowanym okresie nie zaobserwowano jednakowej prawidłowości dla wszystkich powiatów. Można jednak zauważyć, że w 2020 r. w stosunku do 2019 r. w wielu powiatach (np. goleniowskim, kołobrzeskim i policki) wartość tej zmiennej wzrosła, chociaż są też powiaty (gryfiński, łobeski, szczecinecki i świdwiński), w których spadła. W 2021 r. wzrost dochodów w stosunku do poprzedniego roku odnotowano już we wszystkich powiatach.

Na wykr. 1F zaprezentowano, jak zmieniała się wielkość wydatków majątkowych inwestycyjnych z budżetu ogółem przypadających na mieszkańca. Poziom tej zmiennej w poszczególnych latach był bardzo zróżnicowany w badanym okresie. Można zauważyć, że zarówno w 2020 r., jak i w 2021 r. w kilku jednostkach (np. policki i szczecineckim) pandemia spowodowała znaczne ograniczenie wydatków majątkowych inwestycyjnych w porównaniu z 2019 r. W kilku jednostkach wydatki te zmalały w 2020 r. w porównaniu z 2019 r. i zwiększyły się w 2021 r. w porównaniu z 2020 r. (np. w pow. świdwińskim). Były też powiaty (np. łobeski), w których po-

ziom tej zmiennej wzrósł w 2020 r. w porównaniu z 2019 r., a w 2021 r. był zbliżony do 2020 r.

W przypadku stopy bezrobocia rejestrowanego (wykr. 1G) we wszystkich powiatach widoczny jest spadek w 2008 r. w porównaniu z 2007 r., w latach 2009–2013 – wzrost rok do roku, w latach 2014–2019 – spadek rok do roku, a następnie w 2020 r. wzrost w stosunku do 2019 r. (wyjątek stanowi pow. łobeski, w którym odnotowano spadek z 19,0% do 18,3%) i w 2021 r. ponowny spadek w stosunku do roku poprzedniego.

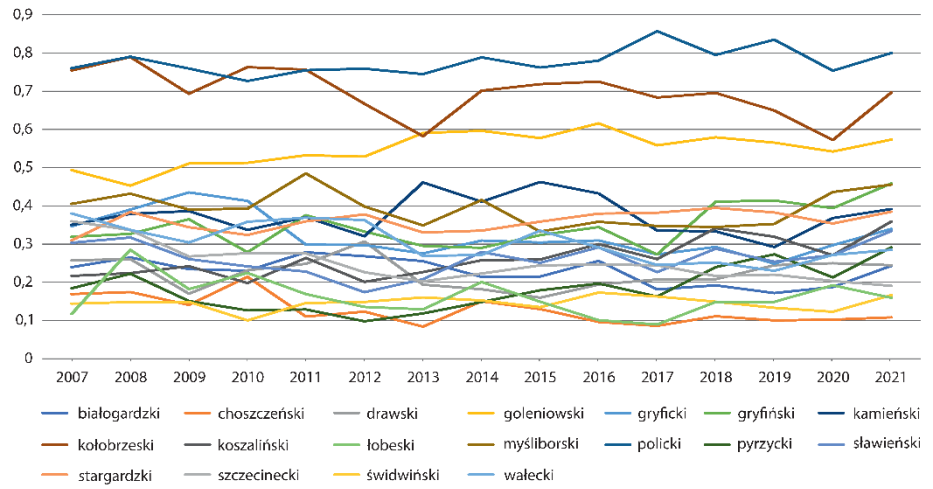
Z wykr. 1H wynika, że w całym analizowanym okresie liczba pracujących na 1000 ludności ogółem charakteryzowała się podobną dynamiką w 17 powiatach. Jedynie w pow. policki odnotowano znaczny wzrost tej zmiennej w latach 2017 i 2018 w stosunku do roku poprzedniego. Uwzględniając wpływ pandemii, należy zauważyć, że liczba pracujących na 1000 ludności ogółem zmniejszyła się w 2020 r. w stosunku do 2019 r. w większości powiatów (z wyjątkiem pow. pyrzyckiego, w którym nieznacznie się zwiększyła), natomiast w 2021 r. w stosunku do 2020 r. dalszy spadek wartości tej zmiennej zaobserwowano w sześciu powiatach (drawskim, goleniowskim, gryfińskim, pyrzyckim, świdwińskim i wałeckim); pozostałe jednostki cechowały się nieznacznym wzrostem.

Liczba ofert pracy na koniec roku na 1000 bezrobotnych (wykr. 1I) była bardzo zróżnicowana w poszczególnych latach w badanych jednostkach. Można zauważyć, że w 2020 r. w stosunku do 2019 r. poziom tej zmiennej znacznie spadł we wszystkich powiatach, ale już w 2021 r. wzrósł w stosunku do 2020 r., z wyjątkiem powiatów gryfińskiego (spadek w porównaniu z 2020 r.) i myśliborskiego (bez znacznych zmian).

Z powyższej analizy wynika, że tylko w przypadku dwóch zmiennych: przeciętnego miesięcznego wynagrodzenia brutto i liczby podmiotów gospodarki narodowej na 10 tys. mieszkańców w wieku produkcyjnym negatywny wpływ pandemii COVID-19 nie był zauważalny w żadnym powiecie, natomiast wyraźnie zaznaczył się we wszystkich powiatach w przypadku pozostałych siedmiu zmiennych.

W celu określenia wpływu pandemii na sytuację gospodarczą w badanych jednostkach wyznaczono miarę syntetyczną z wykorzystaniem analizowanych zmiennych (wykr. 2). Z wykr. 2 wynika, że w całym analizowanym okresie najlepszą sytuacją gospodarczą pod względem rozpatrywanych zmiennych (wartości miary syntetycznej bliskie 1) charakteryzowały się powiaty policki, kołobrzeski i goleniowski, a najgorszą (wartości miary syntetycznej bliskie 0) – powiaty łobeski, świdwiński i choszczeński. Warto również zauważyć, że w 2020 r. poziom miary syntetycznej w stosunku do 2019 r. był niższy prawie we wszystkich powiatach, co można zinterpretować jako pogorszenie sytuacji gospodarczej opisywanej za pomocą zmiennych wybranych do badania.

Wykr. 2. Wartości miary syntetycznej dla powiatów



Źródło: opracowanie własne na podstawie danych z BDL.

Na podstawie wartości miary syntetycznej dokonano grupowania powiatów (tabl. 2). Poszukiwano odpowiedzi na pytanie, czy w latach pandemii zmieniła się przynależność powiatów do grup. Analiza wyników przedstawionych w tablicy pozwala zauważyć, że w okresie 2007–2021 (nawet w latach pandemii) przynależności do grupy nie zmieniły następujące powiaty: choszczeński (grupa IV), kołobrzeski (grupa I), myśliborski (grupa II), policki (grupa I) i świdwiński (grupa IV). Pozostałe powiaty przechodziły tylko do grupy bezpośrednio wyższej albo niższej. W żadnej jednostce nie dokonała się diametralna zmiana, tzn. przejście z grupy I do IV lub odwrotnie. Tylko w dwóch powiatach (pyrzyckim i szczecineckim) sytuacja gospodarcza w pierwszym roku pandemii pogorszyła się na tyle, że jednostki spadły z grupy III do grupy IV.

Te niewielkie zmiany w przynależności powiatów do grup w poszczególnych latach znajdują potwierdzenie w wynikach zgodności grupowania rok do roku w latach 2008–2021. Jak można zauważyć w tabl. 2, niewiele powiatów zmieniało przynależność do grupy w dwóch kolejnych latach.

Tabl. 1. Przynależność powiatów do grup wydzielonych ze względu na wartość miary syntetycznej

Powiaty	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Białogardzki	III	III	III	III	III	III	III	IV	III	III	IV	IV	IV	IV	IV
Choszczeński	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV
Drawski	III	IV	IV	III	III	III	IV	IV	IV	IV	III	IV	III	III	IV
Goleniowski	I	II	I	I	I	I	I	I	I	I	I	I	I	I	I
Gryficki	II	II	II	II	III	III	III	III	III	III	III	III	III	III	III
Gryfiński	III	III	II	III	II	II	III	III	III	II	III	II	II	II	II
Kamieński	II	II	II	II	II	II	II	II	II	II	II	II	III	II	II
Kołobrzeski	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
Koszaliński	IV	IV	III	IV	III	III	III	III	III	III	III	II	III	III	III
Łobeski	IV	III	IV	III	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV
Myśliborski	II	II	II	II	II	II	II	II	II	II	II	II	II	II	II
Policki	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
Pyrzycki	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	III	III	IV	III
Sławieński	III	III	III	III	III	IV	III	III	III	III	III	III	III	III	III
Stargardzki	III	II	II	III	II	II	II	II	II	II	II	II	II	II	II
Szczecinecki	II	III	III	III	III	III	III	IV	III	III	III	III	III	IV	IV
Świdwiński	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV
Wałecki	II	III	III	II	II	II	III	III	II	III	III	III	III	III	III

Uwaga. Grupy o sytuacji gospodarczej: I – najlepszej, II – dobrej, III – słabej, IV – najgorszej.

Źródło: opracowanie własne na podstawie danych z BDL.

Tabl. 2. Zgodność przynależności powiatów do grup rok do roku

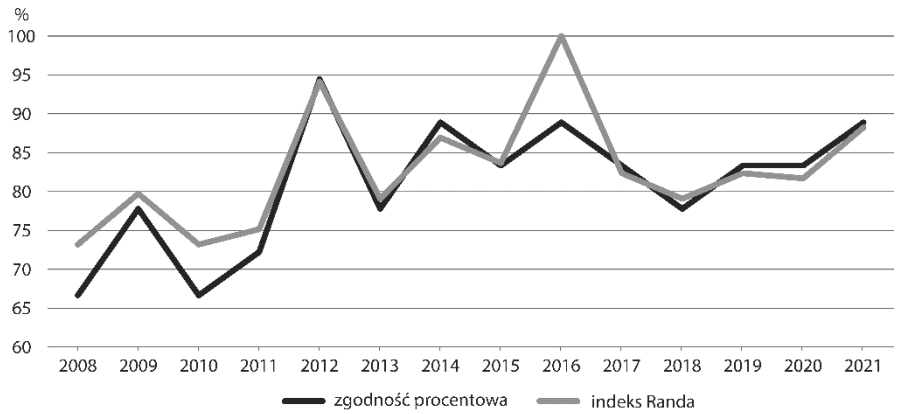
Powiaty	2008/ 2007	2009/ 2008	2010/ 2009	2011/ 2010	2012/ 2011	2013/ 2012	2014/ 2013	2015/ 2014	2016/ 2015	2017/ 2016	2018/ 2017	2019/ 2018	2020/ 2019	2021/ 2020
Białogardzki	1	1	1	1	1	1	0	0	1	0	1	1	1	1
Choszczeński	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Drawski	0	1	0	1	1	0	1	1	1	0	0	0	1	0
Goleniowski	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Gryficki	1	1	1	0	1	1	1	1	1	1	1	1	1	1
Gryfiński	1	0	0	0	1	0	1	1	0	0	0	1	1	1
Kamieński	1	1	1	1	1	1	1	1	1	1	1	0	0	1
Kołobrzeski	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Koszaliński	1	0	0	0	1	1	1	1	1	1	0	0	1	1
Łobeski	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Myśliborski	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Policki	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pyrzycki	1	1	1	1	1	1	1	1	1	1	0	1	0	0
Sławieński	1	1	1	1	0	0	1	1	1	1	1	1	1	1
Stargardzki	0	1	0	0	1	1	1	1	1	1	1	1	1	1
Szczecinecki	0	1	1	1	1	1	0	0	1	1	1	1	0	1
Świdwiński	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Wałecki	0	1	0	1	1	0	1	0	0	1	1	1	1	1

Uwaga. Przynależność powiatu w dwóch kolejnych latach do: 1 – tej samej grupy, 0 – różnych grup.

Źródło: opracowanie własne na podstawie tabl. 1.

Powyższe wnioski można uznać za uzasadnione również na podstawie uzyskanych wartości miary zgodności procentowej i indeksu Randa w latach 2008–2021, które zaprezentowano na wykr. 3.

Wykr. 3. Miara zgodności procentowej i indeks Randa rok do roku



Źródło: opracowanie własne na podstawie tabl. 2.

Wartości miary zgodności procentowej i wartości indeksu Randa kształtowały się w analizowanych latach powyżej 65%. Indeks Randa osiągnął najwyższe wartości w latach 2012 i 2016. Oznacza to, że grupowania w kolejnych latach charakteryzowały się bardzo dużą stabilnością, jeżeli chodzi o pozostawanie zarówno powiatu, jak i par powiatów w tej samej grupie. W 2020 r. zgodność procentowa pozostała na tym samym poziomie, a indeks Randa nieznacznie zmalał w stosunku do 2019 r., co prowadzi do wniosku, że pojawiły się niewielkie zmiany w przynależności powiatów do grup. W 2021 r. wartości obu miar wyraźnie wzrosły w stosunku do 2020 r., ale przynależność powiatów do grup w 2020 r. była prawie taka sama jak w 2021 r. Można zatem stwierdzić, że w 2020 r. pandemia miała pewien wpływ na grupowanie powiatów, natomiast w 2021 r. zaszły niewielkie zmiany w grupowaniu powiatów w stosunku do 2020 r.

4. Podsumowanie

W artykule przedstawiono wyniki badania stabilności grupowania powiatów woj. zachodniopomorskiego ze względu na sytuację gospodarczą w latach 2007–2021, którego celem było uzyskanie odpowiedzi na pytanie, czy pandemia COVID-19 zmieniła sytuację gospodarczą tych jednostek. Wykorzystano dziewięć zmiennych charakteryzujących sytuację przedsiębiorstw, sytuację na rynku pracy i sytuację finansową powiatów. Z badania wynika, że w latach pandemii dynamika poszczegól-

nych zmiennych była w większości powiatów bardzo podobna. Wartości siedmiu analizowanych zmiennych wyraźnie zmalały w 2020 r. w stosunku do 2019 r. Największe spadki w większości powiatów odnotowano w przypadku liczby ofert pracy na koniec roku na 1000 bezrobotnych oraz liczby nowo zarejestrowanych podmiotów gospodarki narodowej na 1000 mieszkańców. Jedynie dwie zmienne: przeciętne miesięczne wynagrodzenie brutto oraz liczba podmiotów gospodarki narodowej na 10 tys. mieszkańców w wieku produkcyjnym pozostały odporne na skutki pandemii.

Oceniając wpływ pandemii COVID-19 na sytuację gospodarczą powiatów przez pryzmat wszystkich analizowanych zmiennych, należy zauważyć, że w 2020 r. w stosunku do 2019 r. tylko w dwóch powiatach (pyrzyckim i szczecineckim) sytuacja gospodarcza pogorszyła się na tyle, że zmieniły one przynależność do grupy (z III na IV). W jednym powiecie (kamieńskim) sytuacja gospodarcza się poprawiła, dzięki czemu awansował z grupy III do II. Pozostałe powiaty – nawet jeżeli ich sytuacja gospodarcza pogorszyła się w 2020 r. – pozostały w tej samej grupie co w 2019 r.

Uzyskane wyniki świadczą o bardzo dużej stabilności grupowania powiatów w badanym okresie. Oznacza to, że pandemia zasadniczo nie zmieniła ich sytuacji gospodarczej. Jednym z powodów takiego stanu rzeczy jest to, że w większości powiatów woj. zachodniopomorskiego analizowane zmienne charakteryzowały się podobną dynamiką w latach pandemii i dlatego nie uzyskano znacznych różnic w grupowaniu powiatów w latach 2020 i 2021.

Bibliografia

Astorquiza Bustos, B. A., Quintero Peña, J. W., Ramos-Barrera, M. G. (2022). Differential effects of the lockdown on labour market outcomes. Evidence for an emerging economy. *Economics and Sociology*, *15*(1), 125–140. https://doi.org/10.14254/2071–789X.2022/15–1/8.

Bieszk-Stolorz, B., Dmytrów, K. (2022). Assessment of the Similarity of the Situation in the EU Labour Markets and Their Changes in the Face of the COVID-19 Pandemic. *Sustainability*, *14*(6), 1–20. https://doi.org/10.3390/su14063646.

Brough, A. R., Martin, K. D. (2021). Consumer Privacy During (and After) the COVID-19 Pandemic. *Journal of Public Policy & Marketing*, *40*(1), 108–110. https://doi.org/10.1177/0743915620929999.

Czernek-Marszałek, K., Piotrowski, P. (2022). Cyfryzacja w przedsiębiorstwach turystycznych w warunkach COVID-19. Pozytywne i negatywne konsekwencje. *Przegląd Organizacji*, (4), 3–12. https://doi.org/10.33141/po.2022.04.01.

Dziuba, R. (2022). Zrównoważony rozwój turystyki w dobie pandemii COVID-19 – przykład państw Bałkanów Zachodnich kandydujących do Unii Europejskiej. W: Z. Wysokińska, J. Witkowska (red.), *Unia Europejska w procesie zrównoważonego rozwoju w warunkach pandemii COVID-19* (s. 129–161). Wydawnictwo Uniwersytetu Łódzkiego. https://doi.org/10.18778/8220-933-4.4.

Eger, L., Komárková, L., Egerová, D., Mičík, M. (2021). The effect of COVID-19 on consumer shopping behaviour: Generational cohort perspective. *Journal of Retailing and Consumer Services*, *61*, 1–11. https://doi.org/10.1016/j.jretconser.2021.102542.

Gatnar, E., Walesiak, M. (red.). (2004). *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Hamulczuk, M., Idzik, M. (2022). Ocena aktywności ekonomicznej w polskim sektorze bankowym w okresie pandemii COVID-19 na podstawie testów koniunktury. *Wiadomości Statystyczne. The Polish Statistician*, *67*(9), 1–23. https://doi.org/10.5604/01.3001.0015.9861.

Kryeziu, L., Bağış, M., Kurutkan, M. N., Krasniqi, B. A., Haziri, A. (2022). COVID-19 impact and firm reactions towards crisis: Evidence from a transition economy. *Journal of Entrepreneurship, Management and Innovation*, *18*(1), 169–196. https://doi.org/10.7341/20221816.

Kukuła, K. (2000). *Metoda unitaryzacji zerowanej*. Wydawnictwo Naukowe PWN.

Landmesser-Rusek, J. (2022). Relationship between the COVID-19 pandemic and currency exchange rates studied by means of the Dynamic Time Warping Method. *Wiadomości Statystyczne. The Polish Statistician*, *67*(5), 1–23. https://doi.org/10.5604/01.3001.0015.8535.

Łuniewska, M., Tarczyński, W. (2006). *Metody wielowymiarowej analizy porównawczej na rynku kapitałowym*. Wydawnictwo Naukowe PWN.

Machová, R., Korcsmáros, E., Marča, R., Esseová, M. (2022). An International Analysis of Consumers' Consciousness During the COVID-19 Pandemic in Slovakia and Hungary. *Folia Oeconomica Stetinensia*, *22*(1), 130–151. https://doi.org/10.2478/foli–2022-0007.

Markowska, M. (2016). Regiony polskie w klasyfikacji pod względem poziomu inteligentnego rozwoju i wrażliwości na kryzys ekonomiczny. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, (433), 138–153. http://dx.doi.org/10.15611/pn.2016.433.14.

Mastalerz-Kodzis, A., Zeug-Żebro, K. (red.). (2023). *Gospodarka, społeczeństwo i rynki finansowe w dobie wyzwań współczesnego świata*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Młodak, A. (2006). *Analiza taksonomiczna w statystyce regionalnej*. Difin.

Pawełek, B. (2008). *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*. Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Ptak-Chmielewska, A., Baszniak, K., Kurpanik, J. (2022). Wpływ pandemii COVID-19 na stan zdrowia psychicznego społeczeństwa. *Wiadomości Statystyczne. The Polish Statistician*, *67*(9), 24–52. https://doi.org/10.5604/01.3001.0015.9862.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356.

Roman, M., Roman, M., Grzegorzewska, E., Pietrzak, P., Roman, K. (2022). Influence of the COVID-19 Pandemic on Tourism in European Countries: Cluster Analysis Findings. *Sustainability*, *14*(3), 1–17. https://doi.org/10.3390/su14031602.

Sokołowski, A., Markowska, M. (2017). Iteracyjna metoda liniowego porządkowania obiektów wielocechowych. *Przegląd Statystyczny. Statistical Review*, *64*(2), 153–162. https://doi.org/10.5604/01.3001.0014.0788.

Strahl, D. (red.). (2006). *Metody oceny rozwoju regionalnego*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Ustawa z dnia 16 kwietnia 2020 r. o szczególnych instrumentach wsparcia w związku z rozprzestrzenianiem się wirusa SARS-CoV-2 (Dz.U. 2020 poz. 695).

Wędrowska, E. (2012). *Miary entropii i dywergencji w analizie struktur*. Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego w Olsztynie.

Witkowska, J. (2022). Bezpośrednie inwestycje zagraniczne a równoważenie rozwoju Unii Europejskiej: aspekty ekologiczne. W: Z. Wysokińska, J. Witkowska (red.), *Unia Europejska w procesie zrównoważonego rozwoju w warunkach pandemii COVID-19* (s. 66–71). Wydawnictwo Uniwersytetu Łódzkiego.

Wysokińska, Z. (2022). Unia Europejska w handlu międzynarodowym na drodze do zrównoważonej gospodarki cyfrowej w warunkach pandemii COVID-19. W: Z. Wysokińska, J. Witkowska (red.), *Unia Europejska w procesie zrównoważonego rozwoju w warunkach pandemii COVID-19* (s. 13–64). Wydawnictwo Uniwersytetu Łódzkiego. https://doi.org/10.18778/8220–933–4.1.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), 20–39 https://doi.org/10.59139/ws.2023.05.2
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 17.07.2022, poprawiony/revised: 31.03.2023, zaakceptowany/accepted: 26.04.2023

Ocena zmian komponentów wskaźnika NEET w krajach Unii Europejskiej z zastosowaniem testu T^2 Hotellinga

Dawid Giemza[a]

Streszczenie. Decydenci i badacze zajmujący się rynkiem pracy i edukacją dużo uwagi poświęcają wskaźnikowi młodzieży niepracującej, nieuczącej się i niedokształcającej się (ang. *neither in employment nor in education and training* – NEET) jako jednemu z mierników komplementarnych wobec stopy bezrobocia. Zazwyczaj ocenia się zmianę poziomu tego wskaźnika ogółem, bez uwzględniania zmian jego komponentów. Celem badania omówionego w artykule jest określenie wielkości zmian komponentów wskaźnika NEET w podziale według cech społeczno-demograficznych i ekonomicznych młodzieży w krajach Unii Europejskiej z zastosowaniem testu T^2 Hotellinga dla pomiarów zależnych. W badaniu posłużono się wielowymiarowym wnioskowaniem statystycznym; wykorzystano testy porównań wielokrotnych. Analizą objęto lata 2019 i 2021, czyli rok poprzedzający wybuch pandemii COVID-19 oraz pełny rok jej trwania. Obliczenia przeprowadzono na podstawie danych Eurostatu.

Z badania wynika, że porównanie zmian komponentów wskaźnika młodzieży NEET pozwala na bardziej szczegółową ocenę sytuacji osób młodych zaliczanych do tej kategorii niż porównanie jedynie poziomu wskaźnika NEET ogółem. Z tego powodu przy podejmowaniu decyzji w zakresie działań politycznych, przede wszystkim dotyczących polityki rynku pracy, należałoby uwzględnić podejście wielowymiarowe.

Słowa kluczowe: młodzież NEET, dekompozycja wskaźnika NEET, rynek pracy, test T^2 Hotellinga, wielowymiarowe wnioskowanie statystyczne, testy porównań wielokrotnych

JEL: C12, C88, J64

Assessment of the changes in the NEET rate structure in EU countries based on Hotelling's T^2 test

Abstract. Decision-makers and researchers dealing with labour market and education are largely focused on the NEET rate of young people as an indicator complementary to the unemployment rate. Typically in studies, it is the change in the overall NEET rate of young people that is evaluated, rather than changes in its individual components. The aim of the study presented in this paper is to detect the volume of changes in the components of the NEET rate determined on the basis of socio-demographic and economic features of young people in EU countries, by means of Hotelling's T^2 test for dependent measurements. The study employed multivariate statistical inference, and more specifically, the multiple comparison tests. The analysis covered the years 2019 and 2021, i.e. the year preceding the outbreak of the COVID-19

[a] Uniwersytet Ekonomiczny w Katowicach, Wydział Gospodarki Przestrzennej i Transformacji Regionów, Polska / University of Economics in Katowice, Faculty of Spatial Economics and Regions Transition, Poland. ORCID: https://orcid.org/0000-0001-7051-526X. E-mail: dawid.giemza@edu.uekat.pl.

pandemic and the full year of its duration. The calculations were performed on the data drawn from the Eurostat database.

The study demonstrates that the comparison of the NEET youth rate components allows a more detailed assessment of the change in the labout market situation of young people belonging to this category than comparing the overall NEET rates only. For this reason, it would be advisable to apply a multidimensional approach when making decisions concerning labour market policies.

Keywords: NEET youth, NEET rate decomposition, labour market, Hotelling's T^2 test, multivariate statistical inference, multiple comparison tests

1. Wprowadzenie

Skutki trudnej sytuacji na rynku pracy bardzo mocno odczuwa młodzież. W 2008 r., wraz z wybuchem kryzysu gospodarczego, zdano sobie sprawę, że tradycyjne miary opisujące sytuację osób młodych na rynku pracy, które koncentrują się na: aktywności zawodowej (współczynnik aktywności zawodowej), zatrudnieniu (współczynnik zatrudnienia), bezrobociu (stopa bezrobocia) i bierności zawodowej (współczynnik dezaktywizacji zawodowej), nie odzwierciedlają w pełni trendów panujących na elastycznych i turbulentnych współczesnych rynkach pracy. Znalazło to potwierdzenie w europejskich badaniach porównawczych (Holte, 2018). Z tego powodu na znaczeniu zyskał wskaźnik przedstawiający odsetek osób młodych jednocześnie niepracujących, nieuczących się i niedokształcających się, nazywanych młodzieżą NEET (ang. *neither in employment nor in education and training*), obliczany w stosunku do ogólnej populacji osób młodych[1]. Stało się tak pomimo krytyki, jakiej ten miernik został poddany w Wielkiej Brytanii (np. Elder, 2015; Furlong, 2006; Serracant, 2014; Tamesberger i Bacher, 2014) po wprowadzeniu go do rządowego raportu *Bridging the gap: New opportunities for 16–18 year-olds not in education, employment or training* (Social Exclusion Unit [Office of the Deputy Prime Minister, London], 1999). Od tamtego czasu jest on powszechnie wykorzystywany do międzynarodowych porównań skali zjawiska NEET, które obecnie zalicza się do najważniejszych problemów społecznych dotyczących młodzieży w Unii Europejskiej.

W badaniach zjawiska NEET wśród młodzieży zazwyczaj ocenia się jedynie zmianę wartości wskaźnika ogółem, bez uwzględniania czynników społeczno-demograficznych i ekonomicznych. Natomiast w badaniu omawianym w niniejszym artykule wzięto pod uwagę oba te aspekty. Celem badania jest określenie wielkości zmian

[1] W literaturze przedmiotu występują rozbieżności w definiowaniu młodzieży NEET ze względu na kryterium wieku osób młodych. Bacher i in. (2017) przyjmują dla nich przedział od 15 do 24 lat, Turek i in. (2014) – od 15 do 29 lat, a Quintano i in. (2018) – od 15 do 34 lat. W badaniu omawianym w niniejszym artykule przyjęto, że są to osoby w wieku od 15 do 29 lat; pojęcia *osoby młode* i *młodzież* są używane zamiennie, co jest często spotykane w literaturze naukowej (Kryńska i in., 1999).

komponentów wskaźnika NEET w podziale według cech społeczno-demograficznych i ekonomicznych w krajach UE z wykorzystaniem testu T^2 Hotellinga dla pomiarów zależnych. Postawiono hipotezę, że podejście wielowymiarowe pozwala na bardziej szczegółową ocenę zmiany sytuacji młodzieży NEET na rynku pracy niż zastosowanie samego wskaźnika NEET ogółem.

2. Komponenty wskaźnika NEET

W literaturze przedmiotu można znaleźć co najmniej kilka propozycji dotyczących sposobu wyznaczania wskaźnika NEET. Zgodnie z metodologią Eurostatu, uwzględniającą definicje Międzynarodowej Organizacji Pracy (MOP; ang. International Labour Organization – ILO), wskaźnik NEET stanowi stosunek liczby młodzieży NEET do ogólnej liczby młodzieży w tych samych grupach wieku i płci, z wyłączeniem osób, które w badaniu statystycznym nie udzieliły odpowiedzi na pytanie o udział w regularnej edukacji i szkoleniach (Balcerowicz-Szkutnik i Wąsowicz, 2017). Wskaźnik ten zatem określa, jaki odsetek całej populacji młodzieży stanowią osoby, które jednocześnie nie pracują ani w ciągu czterech tygodni poprzedzających badanie nie uczestniczyły w systemie nauki i szkoleń.

Wskaźnik NEET jest obliczany na podstawie danych z europejskiego badania siły roboczej (ang. *Labour Force Survey* – LFS), zgodnie z ustalonymi zasadami jakości i niezawodności statystycznej (European Commission, 2011). Sposób obliczania tego miernika można zapisać za pomocą następującego wzoru (Zagórowska i Goleński, 2016):

$$\text{Wskaźnik NEET} = \frac{LM\ NEET}{OLM}, \tag{1}$$

gdzie:
LM NEET – liczba młodzieży NEET,
OLM – ogólna liczba młodzieży.

Opracowano także uproszczone sposoby wyznaczania wskaźnika NEET. Jedna z takich formuł została wyrażona jako stosunek liczby młodych bezrobotnych niebędących studentami (uczniami), powiększonej o liczbę młodych biernych zawodowo niebędących studentami (uczniami), do liczebności całej populacji młodzieży (Elder, 2015). Ma ona jednak wadę, polegającą na tym, że osoby, które nie udzieliły odpowiedzi na pytania zawarte w badaniu statystycznym, nie są odliczane od ogólnej liczby młodzieży (Balcerowicz-Szkutnik i Wąsowicz, 2016).

W innej propozycji uproszczonego pomiaru wskaźnika NEET wykorzystuje się populację osób bezrobotnych powiększoną o osoby nieaktywne zawodowo niebędą-

ce studentami ani uczniami (Elder, 2015). W konstrukcji tej formuły pomija się jednak to, że niektóre osoby bezrobotne mają także status studenta lub ucznia. Nie każdy młody bezrobotny należy bowiem do zbioru NEET, tak jak nie każda osoba NEET jest bezrobotna (oczywiście pewna grupa osób należy do obu tych zbiorów jednocześnie). Z tego powodu liczba bezrobotnych będących studentami lub uczniami powinna zostać wykluczona z obliczeń.

Opierając się na formule podanej przez Eldera (2015), wzorowi (1) można nadać bardziej szczegółową postać – przedstawić wskaźnik NEET jako sumę dwóch ilorazów, z których jeden wyraża udział młodzieży bezrobotnej NEET w ogólnej liczbie młodzieży (odsetek bezrobotnych NEET), a drugi – udział młodzieży biernej zawodowo NEET w ogólnej liczbie młodzieży (odsetek biernych zawodowo NEET)[2]:

$$\text{Wskaźnik NEET} = \frac{LMB - LMBEiS}{OLM} + \frac{LMBZ - LMBZEiS}{OLM} = \\ = \frac{LMB\ NEET}{OLM} + \frac{LMBZ\ NEET}{OLM}, \tag{2}$$

gdzie:

LMB – liczba młodzieży bezrobotnej,

LMBEiS – liczba młodzieży bezrobotnej uczącej się i/lub uczestniczącej w szkoleniach,

LMBZ – liczba młodzieży biernej zawodowo,

LMBZEiS – liczba młodzieży biernej zawodowo uczącej się i/lub uczestniczącej w szkoleniach,

$\frac{LMB\ NEET}{OLM}$ – odsetek bezrobotnych NEET,

$\frac{LMBZ\ NEET}{OLM}$ – odsetek biernych zawodowo NEET.

Jak podkreślono w opracowaniu Bachera i in. (2017), uwzględnienie części zasobu młodzieży biernej (nieaktywnej) zawodowo we wskaźniku NEET jest niewątpliwą zaletą tej miary. Pozwala bowiem dostrzec nowe grupy nieobecnych na rynku pracy (m.in. osoby z niepełnosprawnością, matki opiekujące się dziećmi czy kobiety zajmujące się prowadzeniem domu), a w konsekwencji – zwrócić uwagę decydentów w strukturach administracji publicznej, szczególnie rządowej, na tych młodych ludzi (Eurofound, 2016). Co więcej, wskaźnik NEET w porównaniu ze stopą bezrobocia daje bardziej realistyczny obraz wielkości niewykorzystanych zasobów pracy. To niezwykle istotne, ponieważ współcześnie bardzo dużą wagę przykłada się do zwiększania aktywności zawodowej i społecznej młodzieży. Należy bowiem pamiętać, że

[2] W raportach MOP (ILO, 2013) kategoria biernych zawodowo NEET jest nazywana *neither in the labour force nor in education or training* (NLFET).

na rynek pracy wchodzą coraz mniej liczne roczniki, a w wielu krajach UE postępuje proces starzenia się ludności.

Odmienny sposób wyznaczania wskaźnika NEET zaproponowano w opracowaniu Ripamontiego i Barberisa (2021), a mianowicie:

$$\text{Wskaźnik NEET} = \frac{[Y - (YE + YT)]}{Y}, \tag{3}$$

gdzie:

Y – liczba młodzieży ogółem,

YE – liczba młodzieży pracującej,

YT – liczba młodzieży niepracującej, ale uczącej się i/lub uczestniczącej w szkoleniach.

Wskaźnik NEET poddano dekompozycji, uwzględniając zarówno aktywność zawodową w zakresie poszukiwania pracy, jak i płeć młodych ludzi. Otrzymano następujący wzór:

$$\begin{aligned}\text{Wskaźnik NEET} &= \frac{LMB\ NEET}{OLM} + \frac{LMBZ\ NEET}{OLM} \\ &= \frac{LBM\ NEET + LBK\ NEET}{OLMM + OLMK} + \frac{LBZM\ NEET + LBZK\ NEET}{OLMM + OLMK},\end{aligned} \tag{4}$$

gdzie:

LBM NEET – liczba bezrobotnych mężczyzn NEET,

LBK NEET – liczba bezrobotnych kobiet NEET,

LBZM NEET – liczba biernych zawodowo mężczyzn NEET,

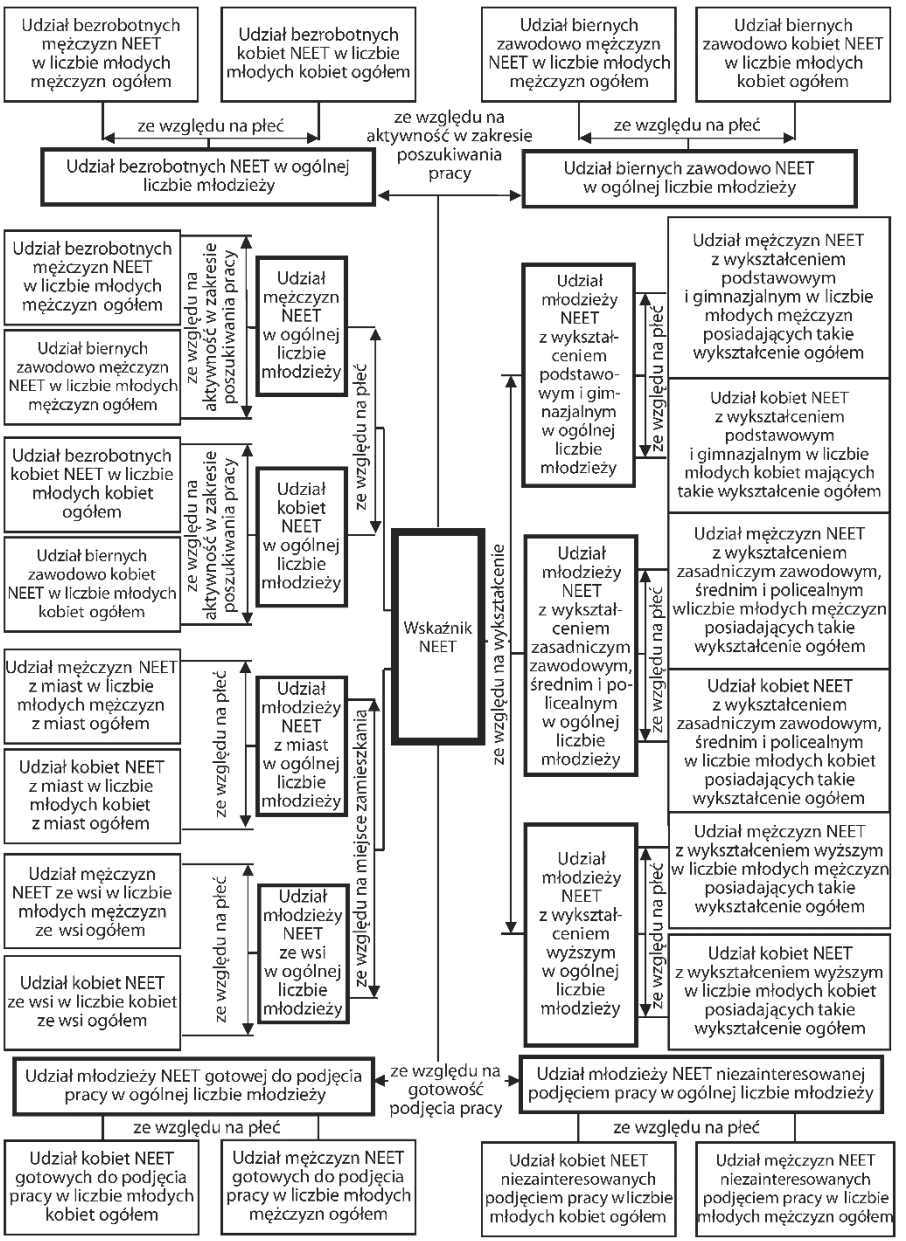
LBZK NEET – liczba biernych zawodowo kobiet NEET,

OLMM – ogólna liczba młodych mężczyzn,

OLMK – ogólna liczba młodych kobiet.

W wyniku dalszej dezagregacji wskaźnika NEET według cech społeczno-demograficznych i ekonomicznych otrzymano jego poszczególne komponenty. Na podstawie kryteriów dostępności, porównywalności, kompletności i aktualności danych statystycznych, podawanych przez Eurostat dla krajów UE, komponenty wskaźnika NEET zapisano jako 29 zmiennych diagnostycznych. Zależności pomiędzy nimi zaprezentowano na schemacie.

Schemat. Zestaw wskaźników składowych odnoszących się do osób młodych (15–29 lat) niepracujących, nieuczących się i niedokształcających się (NEET) z uwzględnieniem cech społeczno-demograficznych i ekonomicznych



Źródło: opracowanie własne na podstawie danych Eurostatu.

Każdy z 29 komponentów wskaźnika NEET w krajach UE zapisano w osobnym prostokącie. Punkt wyjścia stanowił wskaźnik NEET (dotyczący osób w wieku od 15 do 29 lat), który umieszczono w prostokącie o grubym obramowaniu. W wyniku dekompozycji ze względu na różne kryteria (płeć, miejsce zamieszkania, posiadane wykształcenie, aktywność w zakresie poszukiwania pracy i gotowość do podjęcia pracy) otrzymano 11 zmiennych diagnostycznych, które zapisano w prostokątach o cieńszym obramowaniu. Dalszy podział pozwolił uzyskać kolejne 18 zmiennych diagnostycznych, które zapisano w prostokątach o najcieńszym obramowaniu. Można zatem stwierdzić, że na wskaźnik NEET składa się wiele komponentów, które można wyrazić w postaci zmiennych o charakterze wskaźnikowym[3].

3. Metoda badania

Analizę porównawczą komponentów wskaźnika NEET w krajach UE w 2021 r. w stosunku do 2019 r. przeprowadzono z wykorzystaniem testu T^2 Hotellinga dla pomiarów zależnych (Volodarsky i in., 2018, s. 581). Do analizy wielowymiarowej wybrano 11 zmiennych metrycznych (zestawienie), otrzymanych w wyniku dekompozycji pierwszego rzędu (na schemacie zapisano je w prostokątach o średnio pogrubionym obramowaniu).

Zestawienie. Wybrane komponenty wskaźnika NEET wykorzystane w analizie porównawczej

Symbol	Opis zmiennej metrycznej
X_1	odsetek bezrobotnych NEET
X_2	odsetek biernych zawodowo NEET
X_3	odsetek młodzieży NEET gotowej do podjęcia pracy[a]
X_4	odsetek młodzieży NEET niezainteresowanej podjęciem pracy[b]
X_5	odsetek kobiet NEET
X_6	odsetek mężczyzn NEET
X_7	odsetek młodzieży NEET mieszkającej w mieście
X_8	odsetek młodzieży NEET mieszkającej na wsi
X_9	odsetek młodzieży NEET z wykształceniem podstawowym i gimnazjalnym
X_{10}	odsetek młodzieży NEET z wykształceniem zasadniczym zawodowym, średnim i policealnym
X_{11}	odsetek młodzieży NEET z wykształceniem wyższym

a Niezależnie od tego, czy aktywnie jej poszukują, czy nie. b Osoby, które nie poszukują pracy ani nie są gotowe do jej podjęcia, oraz osoby, które poszukują pracy, ale nie są gotowe do jej podjęcia. Zmienna ta różni się od zmiennej X_2 tym, że nie obejmuje osób, które nie poszukują pracy, ale są gotowe do jej podjęcia.

Źródło: opracowanie własne.

[3] Zaprezentowany na schemacie zestaw wskaźników składowych nie jest wyczerpujący. Bazując na danych jednostkowych możliwych do uzyskania z LFS, można wyznaczyć wskaźniki według innych przekrojów, np. stanu cywilnego, stanu rodzinnego / posiadania dzieci lub nie i źródeł dochodu gospodarstwa domowego.

Wskaźnik NEET jest współliniowy ze zmiennymi $X_1 - X_{11}$. Innymi słowy, odpowiednia kombinacja liniowa wybranych zmiennych daje dokładnie wielkość wskaźnika NEET. Zmienne przekazują informację zawartą we wskaźniku NEET w sposób bardziej złożony.

Aby ocenić stabilność i podobieństwo otrzymanych rankingów, obliczono współczynnik korelacji rang Spearmana pomiędzy wskaźnikiem NEET w latach 2019 i 2021. Do obliczeń wykorzystano dane Eurostatu.

Test T^2 Hotellinga dla pomiarów zależnych służy do porównywania wektorów wartości przeciętnych w dwóch momentach. Statystyka ta jest uogólnieniem statystyki t-Studenta wykorzystywanej do porównywania średnich w dwóch momentach. W omawianym badaniu zmienną zależną stanowiły komponenty wskaźnika NEET, a zmienną grupującą był czas (lata 2019 i 2021). Dla pewnych zmiennych (współrzędnych wektora) wartości oczekiwane mogą być jednakowe, a dla innych – znacznie się różnić. W tym drugim przypadku ważne jest nie tylko stwierdzenie, że wektory wartości oczekiwanych nie są jednakowe, lecz także wskazanie, dla których zmiennych występują różnice w wartościach oczekiwanych (Stelmach i Kończak, 2013).

W celu zbadania efektu zmian w czasie z zastosowaniem testu T^2 Hotellinga przyjęto zgodnie z postawioną hipotezą badawczą, że

$$H_0: \begin{bmatrix} \mu_{1,2019} \\ \vdots \\ \mu_{p,2019} \end{bmatrix} = \begin{bmatrix} \mu_{1,2021} \\ \vdots \\ \mu_{p,2021} \end{bmatrix}, \qquad H_1: \begin{bmatrix} \mu_{1,2019} \\ \vdots \\ \mu_{p,2019} \end{bmatrix} \neq \begin{bmatrix} \mu_{1,2021} \\ \vdots \\ \mu_{p,2021} \end{bmatrix}, \tag{5}$$

przy czym zmienna losowa $\boldsymbol{X}$ – różnica pomiędzy latami, czyli $\boldsymbol{X} = \boldsymbol{X}_{2019} - \boldsymbol{X}_{2021}$ – pochodzi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$. Wówczas statystyka testowa przyjmuje postać:

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0), \tag{6}$$

gdzie:

$\boldsymbol{\mu}_0 = (0, \ldots, 0)$,

$\boldsymbol{S}$ – próbkowe oszacowanie macierzy kowariancji $\boldsymbol{\Sigma}$.

Przy założeniu prawdziwości hipotezy H_0 rozkład statystyki testowej jest rozkładem Hotellinga. Badając istotność statystyczną, sprawdzono, czy:

$$T^2 > T^2_{\alpha,p,n-1}, \tag{7}$$

gdzie:
$T^2_{\alpha,p,n-1}$ – kwantyl rozkładu T^2 Hotellinga rzędu $1-\alpha$ z $n-1$ stopniami swobody,
α – poziom istotności testu,
p – liczba elementów wektora $\boldsymbol{X}$,
n – liczebność próby.

Aby przeprowadzić analizę porównawczą komponentów wskaźnika NEET w krajach UE dla lat 2019 i 2021, sprawdzono założenia stosowalności testu T^2 Hotellinga dla pomiarów zależnych. W przypadku porównywania dwóch wektorów należało sprawdzić wielowymiarową normalność w obu momentach i jednorodność macierzy kowariancji. Do badania wielowymiarowej normalności zastosowano test Shapiro-Wilka, a do weryfikacji jednorodności macierzy kowariancji – test M-Boxa.

Test Shapiro-Wilka wykazał odchylenie od normalności w obu analizowanych latach, natomiast test M-Boxa nie dał podstaw do odrzucenia hipotezy o równości macierzy kowariancji pomiędzy komponentami wskaźnika NEET. Biorąc pod uwagę to, że test T^2 Hotellinga charakteryzuje się dużą odpornością na odchylenie od normalności, zdecydowano się wykorzystać test w klasycznej postaci.

4. Wyniki badania

4.1. Analiza porównawcza wskaźnika NEET w krajach UE dla lat 2019 i 2021

Ocena zmiany sytuacji młodzieży niepracującej, nieuczącej się i niedokształcającej jest zazwyczaj przeprowadzana na podstawie wskaźnika NEET ogółem, czyli pojedynczej miary względnej. W tabl. 1 pokazano zmiany poziomu i zróżnicowanie wskaźnika NEET w krajach UE w 2021 r. w stosunku do 2019 r. Kraje uszeregowano według wartości wskaźnika NEET – od najmniejszej (pierwsza, najlepsza pozycja) do największej (ostatnia, najgorsza pozycja).

Tabl. 1. Zmiany poziomu oraz zróżnicowanie wartości wskaźnika NEET w krajach UE – ranking według wartości wskaźnika NEET

Pozycja	2019		2021		Zmiana pozycji[a] w 2021 r. w stosunku do 2019 r.	Przyrost absolutny wartości wskaźnika NEET[a] w 2021 r. w stosunku do 2019 r. w p.proc.
	kraje	wskaźnik NEET w %	kraje	wskaźnik NEET w %		
1.	Holandia	5,7	Holandia	5,5	bez zmian	–0,2
2.	Szwecja	6,3	Szwecja	6,0	bez zmian	–0,3
3.	Luksemburg	6,5	Słowenia	7,3	–4	–1,5
4.	Niemcy	7,6	Dania	8,4	–6	–1,2
5.	Malta	7,9	Luksemburg	8,8	2	2,3
6.	Austria	8,3	Niemcy	9,2	2	1,6
7.	Słowenia	8,8	Finlandia	9,3	–2	–0,2
8.	Portugalia	9,2	Austria	9,4	2	1,1
9.	Finlandia	9,5	Portugalia	9,5	1	0,3
10.	Dania	9,6	Irlandia	9,8	–5	–1,6
11.	Czechy	9,8	Malta	9,9	6	2,0
12.	Łotwa	10,3	Belgia	10,1	–4	–1,7
13.	Estonia	10,6	Czechy	10,9	2	1,1
14.	Litwa	10,9	Estonia	11,2	1	0,6
15.	Irlandia	11,4	Węgry	11,7	–4	–1,5
16.	Belgia	11,8	Łotwa	12,1	4	1,8
17.	Polska	12,0	Litwa	12,7	3	1,8
18.	Francja	13,0	Francja	12,8	bez zmian	–0,2
19.	Węgry	13,2	Polska	13,4	2	1,4
20.	Cypr	14,1	Słowacja	14,2	–2	–0,3
21.	Chorwacja	14,2	Chorwacja	14,9	bez zmian	0,7
22.	Słowacja	14,5	Hiszpania	15,2	–1	0,3
23.	Hiszpania	14,9	Cypr	15,4	3	1,3
24.	Bułgaria	16,7	Grecja	17,3	–2	–0,4
25.	Rumunia	16,8	Bułgaria	17,6	1	0,9
26.	Grecja	17,7	Rumunia	20,3	1	3,5
27.	Włochy	22,2	Włochy	23,1	bez zmian	0,9
.	UE-27	12,6	UE-27	13,1	.	0,5

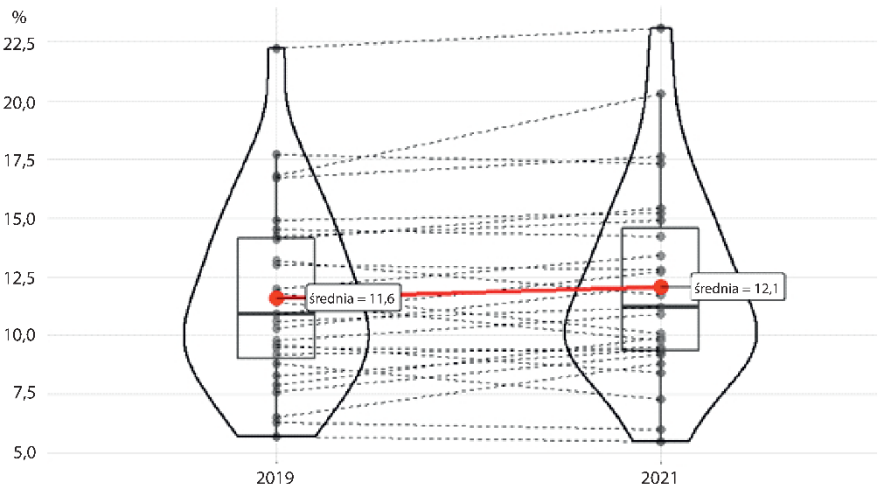
a Dotyczy kraju z kolumny 4.

Źródło: opracowanie własne na podstawie danych Eurostatu.

Dane przedstawione w tabl. 1 wskazują na występowanie wyraźnych dysproporcji pomiędzy krajami UE, jeśli chodzi o sytuację młodzieży NEET. Ich skalę dobrze oddaje miara rozproszenia taka jak rozstęp. Najniższy poziom wskaźnika NEET w obu porównywanych latach zaobserwowano w Holandii i Szwecji. W tych krajach okazał się on ponad trzy razy niższy niż we Włoszech, zajmujących ostatnią pozycję w rankingu. Obliczona wartość współczynnika korelacji rang Spearmana (0,9341) świadczy o bardzo dużej zgodności uporządkowania krajów UE pod względem poziomu wskaźnika NEET w rozpatrywanych latach.

W Polsce wartość wskaźnika NEET w 2021 r. wynosiła 13,4%, a więc o 1,4 p.proc. więcej niż w 2019 r. Zmiana poziomu wskaźnika NEET młodzieży pomiędzy 2019 r. a 2021 r. była widoczna we wszystkich krajach UE (wykr. 1).

Wykr 1. Wskaźnik NEET w krajach UE (kombinacja wykresu ramka-wąsy i wykresu wiolinowego)



Źródło: opracowanie własne na podstawie danych Eurostatu.

Wykorzystując statystykę testu *t*-Studenta dla pomiarów zależnych, weryfikowano hipotezę o braku różnicy pomiędzy średnim wskaźnikiem NEET w krajach UE w latach 2019 i 2021.

Na podstawie testu *t*-Studenta dla pomiarów zależnych można zauważyć, że średnie wartości wskaźnika NEET w krajach UE w latach 2019 i 2021 nie różnią się istotnie statystycznie ($t = -1{,}82$; wartość $p = 0{,}08$; $p > 0{,}05$). Nie ma zatem podstaw, aby twierdzić, że sytuacja młodzieży niepracującej, nieuczącej się i niedokształcającej się w 2021 r. w stosunku do 2019 r. uległa istotnej zmianie (pogorszyła się lub się poprawiła).

4.2. Analiza porównawcza komponentów wskaźnika NEET w krajach UE dla lat 2019 i 2021

W tej części badania przeprowadzono analizę porównawczą komponentów wskaźnika NEET w krajach UE dla lat 2019 i 2021. W tabl. 2 i 3 przedstawiono wartości zmiennych $X_1 - X_{11}$. Uwzględniono także współczynnik zmienności, który informuje o stopniu zróżnicowania wartości poszczególnych zmiennych w krajach UE. Dane dla 2019 r. zaprezentowano w tabl. 2.

Tabl. 2. Statystyki przedstawiające dekompozycję wskaźnika NEET w krajach UE w 2019 r.

Kraje	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
	w %										
Austria	3,0	5,3	6,0	2,3	9,8	6,8	10,7	5,6	11,6	7,7	5,5
Belgia	3,7	8,1	6,7	5,1	12,5	11,2	16,3	7,8	14,9	11,7	7,9
Bułgaria	3,0	13,7	5,8	10,9	20,3	13,2	10,9	27,4	23,6	13,3	10,5
Chorwacja	6,3	7,9	10,1	4,1	16,3	12,2	11,3	15,9	7,8	16,2	17,1
Cypr	5,9	8,2	7,6	6,5	15,9	12,2	13,5	14,0	11,7	16,0	14,0
Czechy	1,6	8,2	2,6	7,2	15,9	4,0	8,7	9,5	8,9	10,2	10,4
Dania	3,0	6,6	6,4	3,2	9,6	9,5	8,1	10,8	10,8	8,8	7,8
Estonia	2,8	7,8	5,2	5,4	14,1	7,3	6,9	13,1[a]	10,8	10,9	9,5
Finlandia	3,2	6,4	4,9	4,6	10,4	8,6	8,1	11,1	9,6	10,4	6,5
Francja	6,0	6,9	8,7	4,3	13,8	12,1	13,4	12,4	15,3	13,9	8,8
Grecja	11,3	6,4	12,8	5,0	19,1	16,4	15,0	22,5	9,5	19,5	25,9
Hiszpania	8,7	6,2	10,6	4,3	15,4	14,4	14,0	14,6	18,9	11,5	12,4
Holandia	1,4	4,3	3,0	2,7	5,8	5,6	6,2	4,8	7,6	4,9	3,4
Irlandia	3,6	7,8	6,4	5,0	12,4	10,5	9,4	12,5	12,5	12,9	7,6
Litwa	4,1	6,8	5,4	5,6	11,0	10,9	7,6	13,2	7,5	14,1	9,2
Luksemburg	3,5	3,0	5,3	1,2	5,9	7,1	6,3	5,8	6,5	8,3	4,5
Łotwa	3,9	6,3	6,0	4,2	10,6	10,0	6,7	14,6	9,8	12,4	6,1
Malta	3,0	4,8	3,9	4,0	9,6	6,3	9,8	4,6	19,5	5,3	3,3
Niemcy	2,2	5,5	4,2	3,5	9,5	5,9	8,3	5,9	10,7	6,1	4,4
Polska	3,0	9,0	6,6	5,4	16,6	7,6	7,9	14,2	8,6	15,8	8,0
Portugalia	4,8	4,4	7,1	2,1	10,1	8,3	8,8	10,9	9,1	9,1	9,6
Rumunia	4,5	12,3	6,0	10,8	22,1	11,8	9,5	20,3	20,6	15,8	8,6
Słowacja	4,7	9,8	6,1	8,4	19,5	9,7	8,0	14,9	18,5	12,6	12,6
Słowenia	3,5	5,3	5,3	3,5	11,2	6,6	9,9	8,4	7,8	9,8	7,8
Szwecja	2,5	3,8	3,6	2,7	6,6	6,0	5,2	7,2	7,1	6,9	3,6
Węgry	3,5	9,7	6,2	7,0	17,9	8,7	8,0	16,9	17,1	11,4	10,3
Włochy	8,2	14,0	16,0	6,2	24,3	20,2	22,9	22,9	21,6	23,4	19,5
Współczynnik zmienności w %	52,9	37,6	44,4	47,2	36,1	37,4	37,9	46,0	40,4	36,3	54,2

a Dane szacunkowe.

Uwaga. Kolorem pomarańczowym zaznaczono największe (najbardziej niepożądane) wartości danej zmiennej, a zielonym – najmniejsze (najmniej niepożądane).

Źródło: opracowanie własne na podstawie danych Eurostatu.

Największe wartości poszczególnych komponentów wskaźnika NEET w 2019 r. zaobserwowano we Włoszech, w Bułgarii i Grecji. Należy jednak zaznaczyć, że żadna gospodarka nie wyróżnia się jednoznacznie ani najmniej, ani najbardziej pożądanym poziomem komponentów wskaźnika NEET niż pozostałe. Luksemburg i Holandia charakteryzują się najniższym poziomem trzech analizowanych komponentów wskaźnika NEET, Czechy i Malta – dwóch, a Szwecja – jednego.

Na podstawie obliczonych wartości współczynnika zmienności można zauważyć, że wyraźnie największa zmienność dotyczy odsetka bezrobotnych NEET (X_1) i odsetka młodzieży NEET z wykształceniem wyższym (X_{11}). Najmniejszym zróżnico-

waniem cechuje się natomiast odsetek kobiet NEET (X_5) i odsetek młodzieży NEET z wykształceniem zasadniczym zawodowym, średnim i policealnym (X_{10}).

W tabl. 3 zaprezentowano dane za 2021 r.

Tabl. 3. Statystyki przedstawiające dekompozycję wskaźnika NEET w krajach UE w 2021 r.

Kraje	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
	w %										
Austria	3,8	5,6	6,8	2,6	10,3	8,5	11,0	7,4	12,9	9,1	5,9
Belgia	4,1	6,0	5,9	4,2	9,7	10,4	13,4	7,6	12,5	10,8	6,5
Bułgaria	3,9	13,7	7,9	9,7	20,9	14,5	11,1	29,3	24,4	14,5	11,1
Chorwacja	7,5	7,5	10,0	5,0	16,2	13,8	11,3	16,8	8,8	18,0	14,2
Cypr	7,7	7,7	9,6	5,7	16,3	14,5	14,5	14,1	12,0	18,2	14,9
Czechy	2,2	8,7	2,9	7,9	17,3	4,8	10,1	11,3	9,2	11,9	11,1
Dania	2,4	6,0	5,1	3,3	8,9	8,0	7,0	9,9	9,7	7,7	7,2
Estonia	3,8	7,4	5,8	5,4	11,7	10,7	10,2	13,5[a]	10,3	13,7	7,6
Finlandia	3,1	6,2	5,1	4,2	8,8	9,7	7,6	10,1	9,3	10,6	4,7
Francja	5,8	7,0	8,8	3,9	13,0	12,5	12,8	11,6	15,0	13,7	9,1
Grecja	10,3	7,0	12,9	4,4	18,1	16,6	13,9	22,8	7,9	19,0	26,8
Hiszpania	9,1	6,0	11,6	3,6	14,8	15,5	14,6	14,4	22,3	10,9	11,7
Holandia	1,7	3,8	4,3	1,3	6,0	5,1	6,1	3,9	9,4	4,2	3,1
Irlandia	3,6	6,2	6,2	3,7	9,7	10,0	9,0	9,8	7,8	12,4	7,1
Litwa	4,4	8,3	5,5	7,2	13,5	11,9	8,4	15,5	10,6	16,9	7,3
Luksemburg	2,5	6,3	6,2	2,5	8,5	9,1	8,9	8,3	11,3	9,0	5,7
Łotwa	4,3	7,8	7,7	4,3	12,6	11,5	8,7	16,4	7,8	17,6	7,3
Malta	3,2	6,7	3,8	6,1	11,4	8,6	12,1	1,3	20,3	8,3	4,7
Niemcy	2,4	6,7	5,5	3,6	10,6	7,9	10,2	6,2	13,4	7,1	4,8
Polska	3,2	10,2	5,1	8,3	16,9	10,1	10,9	15,0	12,5	15,6	8,4
Portugalia	5,0	4,5	7,6	2,0	9,7	9,3	8,6	10,6	8,9	10,0	9,4
Rumunia	5,3	15,0	7,7	12,6	26,3	14,6	9,4	27,1	32,7	15,9	10,1
Słowacja	6,0	8,2	7,9	6,4	17,5	11,1	8,7	13,6	16,6	13,5	12,0
Słowenia	2,8	4,5	3,7	3,6	8,2	6,6	7,8	7,3	7,3	7,9	5,9
Szwecja	2,9	3,1	4,0	2,0	6,3	5,8	5,5	6,5	6,4	6,7	4,0
Węgry	3,9	7,7	6,4	5,3	14,8	8,7	6,2	16,8	15,4	10,3	7,3
Włochy	7,7	15,4	16,0	7,1	25,0	21,2	24,5	20,8	23,0	24,9	17,3
Współczynnik zmienności w %	49,0	40,0	42,5	50,8	38,8	35,0	36,0	51,3	47,7	37,4	54,7

a Dane szacunkowe.

Uwaga. Jak przy tabl. 2.

Źródło: opracowanie własne na podstawie danych Eurostatu.

Po przeanalizowaniu danych za 2021 r. można dojść do wniosku, że wysoka wartość wskaźnika NEET w Polsce to w głównej mierze konsekwencja wysokiego odsetka biernych zawodowo (X_2). W 2019 r. co jedenasta, a w 2021 r. już co dziesiąta młoda osoba w Polsce pozostawała w bierności zawodowej, której nie można wytłumaczyć ani pobieraniem nauki, ani uczestnictwem w szkoleniach. Skali problemu, czyli liczby młodych ludzi nieobecnych na rynku pracy, nie odzwierciedla stopa

bezrobocia – najpopularniejsza miara wykorzystywana przy opracowywaniu polityk społecznych. Dalsze analizy prowadzą do wniosku, że w Polsce w większym stopniu status NEET dotyczy: ze względu na płeć – kobiet (X_5), z uwagi na miejsce zamieszkania – mieszkańców obszarów wiejskich (X_8), a biorąc pod uwagę wykształcenie – osób posiadających wykształcenie zasadnicze zawodowe, średnie i policealne (X_{10}).

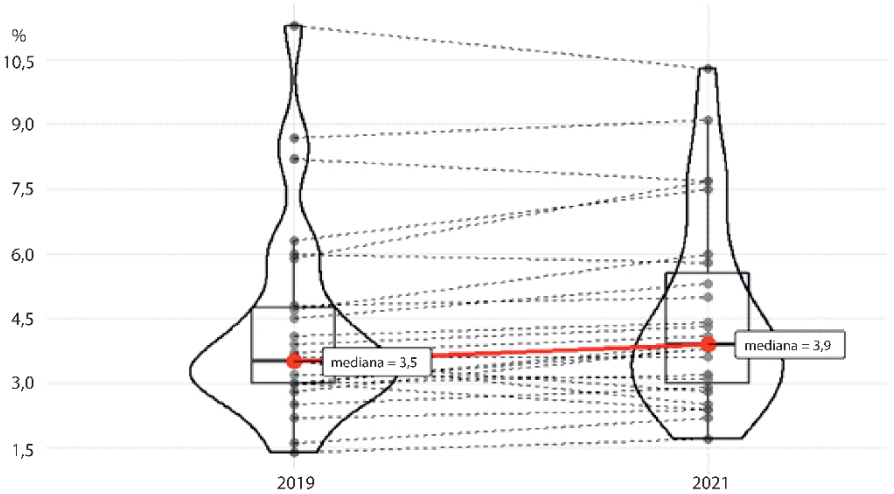
Wartość statystyki testu T^2 Hotellinga dla pomiarów zależnych wynosi 3,2377, a wyznaczona na jej podstawie wartość p – 0,01645 ($p < 0{,}05$). Na podstawie wyniku testu T^2 Hotellinga dla pomiarów zależnych należy zatem odrzucić hipotezę o równości średnich wektorów $\boldsymbol{X}_1 - \boldsymbol{X}_{11}$. Z tego powodu, wykorzystując testy brzegowe, sprawdzono, które komponenty wskaźnika NEET są istotnie zróżnicowane w czasie. Porównań brzegowych dokonano za pomocą testu t-Studenta dla pomiarów zależnych lub testu kolejności par Wilcoxona dla pomiarów zależnych (w zależności od tego, czy zostało spełnione założenie o normalności rozkładu zmiennych i jednorodności wariancji, czy też nie). W badaniu normalności rozkładu zmiennych posłużono się testem Shapiro-Wilka, a do weryfikacji jednorodności wariancji użyto testu F (Fishera-Snedecora). Uzyskane wyniki pokazują, że tylko pięć zmiennych przyjmuje rozkład normalny: X_4, X_5, X_6, X_8 i X_{10}. W ich przypadku nie ma także podstaw do odrzucenia hipotezy o równości wariancji. Z tego powodu do porównań tych zmiennych wykorzystano test t-Studenta dla pomiarów zależnych, a w przypadku pozostałych zmiennych – nieparametryczny test kolejności par Wilcoxona dla pomiarów zależnych.

Następnie uzyskano wyniki testów brzegowych (szczegółowych) tych zmiennych, dla których można było zaobserwować statystycznie istotne różnice w 2021 r. w stosunku do 2019 r. Pierwszą taką zmienną jest odsetek bezrobotnych NEET (X_1), przedstawiony na wykr. 2. W przypadku tej zmiennej statystyka testowa kolejności par Wilcoxona dla pomiarów zależnych (z poprawką na ciągłość) wynosi 96, a odpowiadająca jej wartość $p = 0{,}0446$ ($p < 0{,}05$).

Drugą zmienną istotną statystycznie jest odsetek młodzieży NEET gotowej do podjęcia pracy (X_3), przedstawiony na wykr. 3. W przypadku tej zmiennej statystyka testowa kolejności par Wilcoxona dla pomiarów zależnych (z poprawką na ciągłość) wynosi 87, a odpowiadająca jej wartość $p = 0{,}025$ ($p < 0{,}05$).

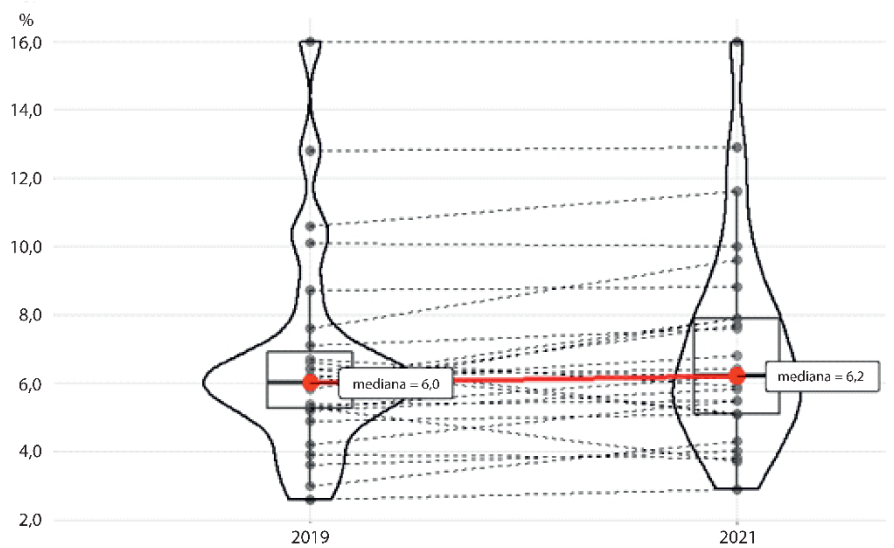
Trzecią zmienną, która wykazuje statystycznie istotną różnicę w 2021 r. w stosunku do 2019 r., jest odsetek mężczyzn NEET (X_6), przedstawiony na wykr. 4.

Wykr 2. Odsetek bezrobotnych NEET (kombinacja wykresu ramka-wąsy i wykresu wiolinowego)



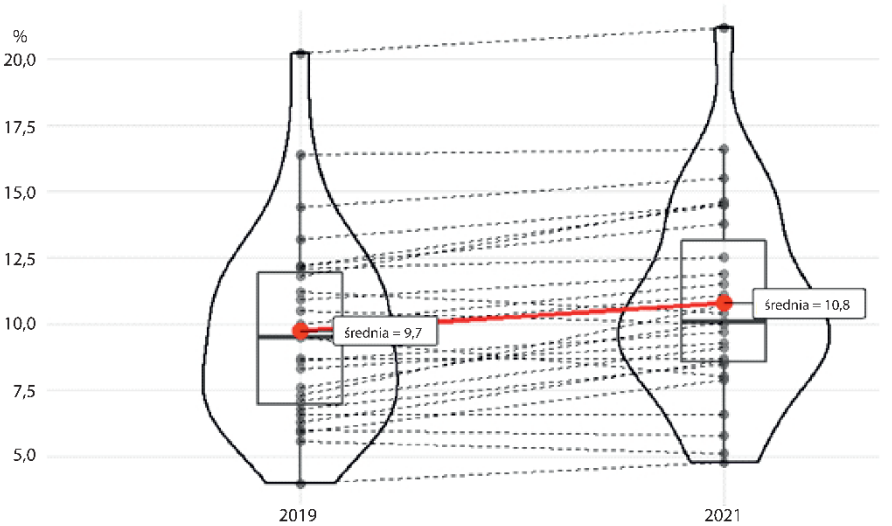
Źródło: opracowanie własne na podstawie danych Eurostatu.

Wykr. 3. Odsetek młodzieży NEET gotowej do podjęcia pracy (kombinacja wykresu ramka-wąsy i wykresu wiolinowego)



Źródło: opracowanie własne na podstawie danych Eurostatu.

Wykr. 4. Odsetek mężczyzn NEET (kombinacja wykresu ramka-wąsy i wykresu wiolinowego)



Źródło: opracowanie własne na podstawie danych Eurostatu.

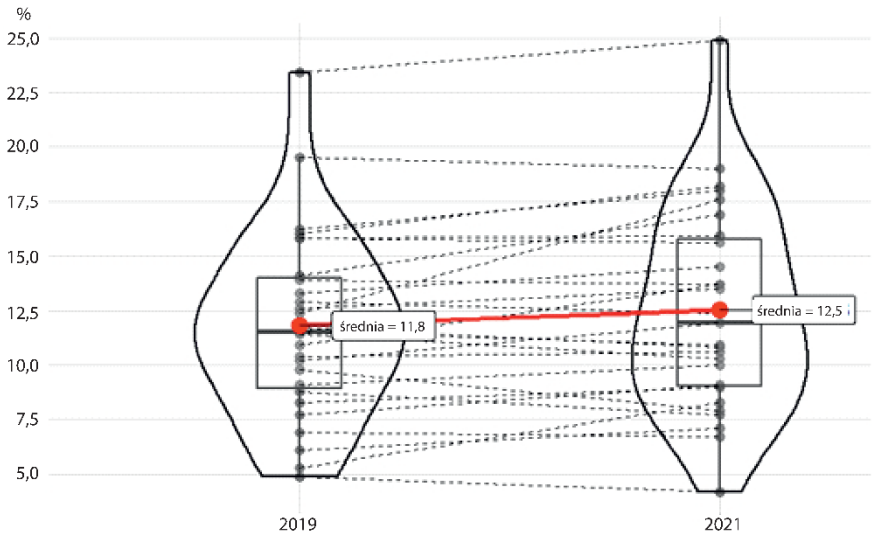
Na wykr. 4 widać, że w 2021 r. średnia wartość odsetka mężczyzn NEET w krajach UE wynosiła 10,8%. Oznacza to, że w 2021 r. średnio co dziesiąty młody mężczyzna mieszkający w UE jednocześnie nie pracował i nie uczył się ani nie uczestniczył w szkoleniach. Dla porównania, średni odsetek kobiet NEET ukształtował się na poziomie 13,4%.

Analizę porównawczą zmiennej X_6 w krajach UE w 2021 r. w stosunku do 2019 r. przeprowadzono z wykorzystaniem testu t-Studenta dla pomiarów zależnych. Obliczona wartość statystyki testowej wynosi –4,54, a wyznaczona na jej podstawie wartość $p = 0{,}0001$ ($p < 0{,}05$).

Ostatnią zmienną, która wykazała statystycznie istotną różnicę w 2021 r. w stosunku do 2019 r., jest odsetek młodzieży z wykształceniem zasadniczym zawodowym, średnim i policealnym (X_{10}), przedstawiony na wykr. 5. Do analizy statystycznej tej zmiennej również użyto parametrycznego testu t-Studenta dla pomiarów zależnych. Statystyka testowa t-Studenta dla X_{10} wynosi –2,36, a wyznaczona na jej podstawie wartość $p = 0{,}0260$ ($p < 0{,}05$).

Zbiorcze wyniki porównań brzegowych wszystkich komponentów (zarówno istotnych, jak i nieistotnych statystycznie) wskaźnika NEET dla lat 2019 i 2021 zamieszczono w tabl. 4.

Wykr. 5. Odsetek młodzieży NEET z wykształceniem zasadniczym zawodowym, średnim i policealnym (kombinacja wykresu ramka-wąsy i wykresu wiolinowego)



Źródło: opracowanie własne na podstawie danych Eurostatu.

Tabl. 4. Wyniki porównań brzegowych poszczególnych komponentów wskaźnika NEET w krajach UE

Zmienne	2019		2021		Wartość statystyki testowej		Wartość *p*
	średnia arytmetyczna	mediana	średnia arytmetyczna	mediana			
	w %						
X_1	4,26	3,5	4,54	3,9	(b)	96,00	0,0446*
X_2	7,38	6,8	7,53	7,0	(b)	164,00	0,7799
X_3	6,61	6,0	7,04	6,2	(b)	87,00	0,0254*
X_4	5,01	4,6	5,03	4,3	(a)	−0,11	0,9117
X_5	13,56	12,5	13,44	12,6	(a)	0,33	0,7445
X_6	9,74	9,5	10,78	10,1	(a)	−4,54	0,0001*
X_7	10,05	8,8	10,46	10,1	(b)	123,50	0,1908
X_8	12,66	12,5	12,89	11,6	(a)	−0,60	0,5565
X_9	12,51	10,8	13,25	11,3	(b)	154,00	0,4071
X_{10}	11,81	11,5	12,53	11,9	(a)	−2,36	0,0260*
X_{11}	9,44	8,6	9,08	7,3	(b)	146,50	0,3128

Uwaga. Statystyka testowa: (a) – test *t*-Studenta dla pomiarów zależnych, (b) – test kolejności par Wilcoxona dla pomiarów zależnych (z poprawką na ciągłość). * $p < 0,05$.

Źródło: opracowanie własne na podstawie danych Eurostatu.

Wyniki testów brzegowych przedstawione w tabl. 4 pozwalają zaobserwować statystycznie istotne różnice w zakresie kształtowania się wartości zmiennych X_1, X_3, X_6 i X_{10} w latach 2019 i 2021. Warto przy tym dodać, że w przypadku X_1 i X_3 zostały one potwierdzone nieparametrycznym testem kolejności par Wilcoxona dla pomiarów zależnych (z poprawką na ciągłość), a w przypadku zmiennych X_6 i X_{10} – parametrycznym testem t-Studenta dla pomiarów zależnych. Różnice dotyczące pozostałych zmiennych w analizowanych latach okazały się nieistotne statystycznie.

5. Podsumowanie

Do oceny zmiany sytuacji młodzieży na rynku pracy wykorzystywane są różne miary. W UE w ostatnich kilkunastu latach szczególną wagę przywiązuje się do wskaźnika NEET. W badaniu omawianym w artykule dokonano porównania zarówno zmiany poziomu wskaźnika NEET ogółem (co było już przedmiotem wielu badań, w tym realizowanych przez instytucje międzynarodowe), jak i wektorów średnich wartości zmiennych składających się na ten wskaźnik, co jest autorskim wkładem w badania zjawiska NEET. Analizowano zmianę, jaka nastąpiła pomiędzy rokiem 2019, bezpośrednio poprzedzającym wybuch pandemii, a pełnym rokiem jej trwania – 2021. Do oceny zmiany wykorzystano parametryczny test T^2 Hotellinga dla pomiarów zależnych.

Wyniki badania pokazują, że ocena zmiany sytuacji młodzieży jednocześnie niepracującej, nieuczącej się i niedokształcającej się w krajach UE w 2021 r. w stosunku do 2019 r. dokonana za pomocą porównania jedynie wartości wskaźnika NEET ogółem prowadzi do wniosku, że nie uległa ona istotnej zmianie (nie pogorszyła się ani się nie poprawiła). Porównanie komponentów wskaźnika NEET za pomocą testu T^2 Hotellinga dla pomiarów zależnych prowadzi z kolei do wniosku, że sytuacja młodzieży NEET w rozpatrywanych latach była różna. Testy wielokrotnych porównań wskazały ponadto, że w 2021 r. w stosunku do 2019 r. istotnie statystycznie zwiększyły się wartości odsetka: bezrobotnych NEET, młodzieży NEET gotowej do podjęcia pracy, mężczyzn NEET i młodzieży NEET z wykształceniem zasadniczym zawodowym, średnim i policealnym. Na tej podstawie można przyjąć, że testowana hipoteza – głosząca, że podejście wielowymiarowe pozwala na bardziej szczegółową ocenę zmiany sytuacji młodzieży NEET na rynku pracy niż użycie samego wskaźnika NEET ogółem – jest prawdziwa.

Furlong (2006), biorąc pod uwagę heterogeniczność młodzieży NEET, stwierdza, że dalsze badania i działania polityczne należy rozpocząć od dezagregacji populacji tych osób, tak aby móc określić odrębne cechy i potrzeby różnych jej podgrup. Wyniki badania omówionego w artykule dowodzą, że równie ważne są dezagregacja wskaźnika NEET i przeanalizowanie zmian jego poszczególnych komponentów.

Pozwala to uzyskać pełniejszy obraz sytuacji młodzieży NEET i zaobserwować rzeczywisty wpływ podejmowanych działań – przede wszystkim w ramach aktywnej polityki dotyczącej rynku pracy – na zmiany w tym zakresie. Jak bowiem wskazuje Dahlke (2016), mimo że działania te cechują się dużą efektywnością zatrudnieniową, to ich rzeczywisty wpływ na sytuację na rynku pracy wydaje się ograniczony.

Bibliografia

Bacher, J., Koblbauer, C., Leitgöb, H., Tamesberger, D. (2017). Small differences matter: how regional distinctions in educational and labour market policy account for heterogeneity in NEET rates. *Journal for Labour Market Research*, *51*(4), 1–20. https://doi.org/10.1186/s12651-017-0232-6.

Balcerowicz-Szkutnik, M., Wąsowicz, J. (2017). Pokolenie NEETs na rynku pracy – aktualne problemy. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, *9*(312), 7–17.

Dahlke, M. (2016). Aktywna polityka rynku pracy i jej znaczenie. W: K. Pujer (red.), *Rynek pracy w Polsce – szanse i zagrożenia* (s. 71–80). Exante.

Elder, S. (2015). *What does NEETs mean and why is the concept so easily misinterpreted?* (Work4Youth Technical Brief No. 1). International Labour Organization. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_343153.pdf.

Eurofound. (2016). *Exploring the diversity of NEETs.* Publications Office of the European Union. https://www.eurofound.europa.eu/sites/default/files/ef_publication/field_ef_document/ef1602en.pdf.

European Commission. (2011). *Youth neither in employment nor education and training (NEET). Presentation of data for the 27 Member States*. https://www.ec.europa.eu/social/BlobServlet?docId=6602&langId=en.

Furlong, A. (2006). Not a very NEET solution: representing problematic labour market transitions among early school-leavers. *Work, Employment and Society*, *20*(3), 553–569. https://doi.org/10.1177/0950017006067001.

Holte, B. H. (2018). Counting and Meeting NEET Young People: Methodology, Perspective and Meaning in research on marginalized youth. *Young*, *26*(1), 1–16. https://doi.org/10.1177/1103308816677618.

International Labour Organization. (2013). *Global employment trends for youth 2013. A generation at risk*. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_212423.pdf.

Kryńska, E., Poliwczak, I., Sobocka-Szczapa, H. (1999). *Młodzież na rynku pracy województwa łódzkiego*. Instytut Pracy i Spraw Socjalnych.

Quintano, C., Mazzocchi, P., Rocca, A. (2018). The determinants of Italian NEETs and the effects of the economic crisis. *Genus, 74*(5), 1–24. https://doi.org/10.1186/s41118-018-0031-0.

Ripamonti, E., Barberis, S. (2021). The association of economic and cultural capital with the NEET rate: differential geographical and temporal patterns. *Journal for Labour Market Research*, *55*(13), 1–17. https://doi.org/10.1186/s12651-021-00296-y.

Serracant, P. (2014). A Brute Indicator for a NEET Case: Genesis and Evolution of Problematic Concept and Results from an Alternative Indicator. *Social Indicators Research*, *117*(2), 401–419. https://doi.org/10.1007/s11205-013-0352-5.

Social Exclusion Unit [Office of the Deputy Prime Minister, London]. (1999). *Bridging the gap: New opportunities for 16–18 year olds not in education, employment or training*. Stationery Office.

Stelmach, J., Kończak, G. (2013). O porównywaniu dwóch populacji wielowymiarowych z wykorzystaniem objętości elipsoid ufności. W: J. Kolonko, G. Kończak (red.), *Metody wnioskowania statystycznego w badaniach ekonomicznych* (s. 146–157). Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Tamesberger, D., Bacher, J. (2014). NEET youth in Austria: a typology including socio-demography, labour market behaviour and permanence. *Journal of Youth Studies*, *17*(9), 1239–1259. https://doi.org/10.1080/13676261.2014.901492.

Turek, D., Wojtczuk-Turek, A., Marczak, K. (2014). *Wsparcie młodych osób na mazowieckim rynku pracy*. Wojewódzki Urząd Pracy w Warszawie. https://obserwatorium.mazowsze.pl/pliki/files/Raport_finalny_z_badania_NEET_okl.pdf.

Volodarsky, E., Warsza, Z., Kosheva, L., Dobrolyubova, M. (2018). Zastosowanie kart kontrolnych Hotellinga w kontroli jakości wieloparametrowego procesu technologicznego. *Przemysł Chemiczny*, *97*(4), 579–583. https://doi.org/10.15199/62.2018.4.13.

Zagórowska, A. M., Goleński, W. R. (2016). Problem młodzieży niepracującej, nieuczącej się, niedokształcającej się (NEET) w województwie opolskim – przyczynek do dalszych badań. W: A. M. Zagórowska (red.), *Edukacja młodzieży a rynek pracy* (s. 78–91). Uniwersytet Opolski, Wojewódzki Urząd Pracy w Opolu.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), 40–57 https://doi.org/10.59139/ws.2023.05.3
Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges
Zgłoszony/submitted: 20.12.2021, poprawiony/revised: 31.03.2023, zaakceptowany/accepted: 20.04.2023

New algorithm for determining the number of features for the effective sentiment-classification of text documents

Adam Idczak,[a] Jerzy Korzeniewski[b]

Abstract. Sentiment analysis of text documents is a very important part of contemporary text mining. The purpose of this article is to present a new technique of text sentiment analysis which can be used with any type of a document-sentiment-classification method. The proposed technique involves feature selection independently of a classifier, which reduces the size of the feature space. Its advantages include intuitiveness and computational non-complexity. The most important element of the proposed technique is a novel algorithm for the determination of the number of features to be selected sufficient for the effective classification. The algorithm is based on the analysis of the correlation between single features and document labels. A statistical approach, featuring a naive Bayes classifier and logistic regression, was employed to verify the usefulness of the proposed technique. They were applied to three document sets composed of 1,169 opinions of bank clients, obtained in 2020 from a Poland-based bank. The documents were written in Polish. The research demonstrated that reducing the number of terms over 10-fold by means of the proposed algorithm in most cases improves the effectiveness of classification.

Keywords: sentiment analysis, document sentiment classification, text mining, logistic regression, naive Bayes classifier, feature selection, correlation

JEL: C52, C81, M31

Nowy algorytm ustalania liczby zmiennych potrzebnych do klasyfikacji dokumentów tekstowych ze względu na ich wydźwięk emocjonalny

Streszczenie. Analiza sentymentu, czyli wydźwięku emocjonalnego, dokumentów tekstowych stanowi bardzo ważną część współczesnej eksploracji tekstu (ang. *text mining*). Celem artykułu jest przedstawienie nowej techniki analizy sentymentu tekstu, która może znaleźć zastosowanie w dowolnej metodzie klasyfikacji dokumentów ze względu na ich wydźwięk emocjonalny. Proponowana technika polega na niezależnym od klasyfikatora doborze cech, co skutkuje zmniejszeniem rozmiaru ich przestrzeni. Zaletami tej propozycji są intuicyjność i prostota obli-

[a] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Polska / University of Lodz, Faculty of Economics and Sociology, Poland. ORCID: https://orcid.org/0000-0001-9676-2410. Autor korespondencyjny / Corresponding author: adam.idczak@uni.lodz.pl.

[b] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Polska / University of Lodz, Faculty of Economics and Sociology, Poland. ORCID: https://orcid.org/0000-0001-6526-5921.
E-mail: jerzy.korzeniewski@uni.lodz.pl.

czeniowa. Zasadniczym elementem omawianej techniki jest nowatorski algorytm ustalania liczby terminów wystarczających do efektywnej klasyfikacji, który opiera się na analizie korelacji pomiędzy pojedynczymi cechami dokumentów a ich wydźwiękiem. W celu weryfikacji przydatności proponowanej techniki zastosowano podejście statystyczne. Wykorzystano dwie metody: naiwny klasyfikator Bayesa i regresję logistyczną. Za ich pomocą zbadano trzy zbiory dokumentów składające się z 1169 opinii klientów jednego z banków działających na terenie Polski uzyskanych w 2020 r. Dokumenty zostały napisane w języku polskim. Badanie pokazało, że kilkunastokrotne zmniejszenie liczby terminów przy zastosowaniu proponowanej techniki na ogół poprawia jakość klasyfikacji.

Słowa kluczowe: analiza sentymentu, klasyfikacja dokumentów ze względu na wydźwięk emocjonalny, eksploracja tekstu, regresja logistyczna, naiwny klasyfikator Bayesa, dobór cech, korelacja

1. Introduction

Text sentiment classification relates mainly to establishing the sentiment of the opinion expressed in a text. In this context, what proves very important is the text presentation, i.e. converting unstructured text into a machine-readable format using selected features. Therefore, feature selection plays a crucial role in text presentation. Existing literature on text sentiment typically uses a simple textual presentation, namely the bag-of-words (BOW), in which each document is represented by single terms along with their frequencies. The primary downside of the BOW is a huge number of features it produces, which very easily leads to the curse-of-dimensionality. Our study proposes a simple framework based on unigram models, which essentially consists in feature-filtering using distance-based correlation. The correlation is measured for each term-feature between the distances between text document sentiment labels and the distances between frequencies of the terms' occurrence, across all documents.

One can find a comprehensive overview of some techniques of sentiment analysis in Medhat et al. (2014). In some works, the SentiWordNet is used, which basically is a WordNet-based lexicon classifying each term as positive, negative or neutral. In Khan et al. (2011), the SentiWordNet was used to calculate the score and determine the polarity of either subjective or objective sentences from reviews and blog comments. The authors showed that their proposal slightly outperformed maximum-likelihood-based methods. Agarwal et al. (2011) carried out a sentiment analysis on Twitter data. They introduced polarity features, where the polarity was measured by means of the SentiWordNet. Kouloumpis et al. (2011) investigated the usefulness of linguistic features for the establishment of the sentiment of Twitter messages. The authors used a subjectivity lexicon to this end. Davies and Ghahramani (2011) presented a language-independent model for the sentiment analysis of short texts using emoticons as sentiment indicators. Their method slightly outperformed the naive Bayes classifier. Njølstad et al. (2014) proposed and evaluated four different feature categories composed of 26 article features for the

sentiment analysis. Then they used different machine-learning (ML) methods to train sentiment classifier of Norwegian on-line financial news. They achieved classification precision of 71%. Govindarajan (2013) proposed an ensemble classification method that used arcing classifier, naive Bayes and genetic algorithm. The evaluation of the performance of these classifiers was carried out by means of different performance-quality metrics on datasets of film reviews. However, the gain classifiers accuracy was very low.

Yazdani et al. (2017) concentrated on overcoming the limits of the BOW method by studying bigram models and using lexicon-based methods to capture the semantics of words. Iqbal et al. (2019) proposed an approach similar to that of Khan et al. (2011), focusing on maximum-likelihood and lexicon-based methods. The genetic algorithm was used for optimised feature selection. As regards feature space, Pintas et al. (2021) did an extensive review of the literature on this subject. The authors described over a hundred methods and algorithms along with a relatively detailed account of their merits and disadvantages. According to them, feature selection methods could be classified into three categories: filter, wrapper, and embedded methods. Filter methods are independent of the learning process and applicable prior to this process. Wrapper methods collaborate closely with the classifier trying to optimise the whole classification process. Embedded methods involve feature selection at the training stage. The important issues connected with feature selection are: measuring feature relevance, subset search and globalisation. The above-mentioned research demonstrates that the scientific achievements in this field enable the user to choose from amongst a plethora of different algorithms, classifiers, and optimisations. A full search of the whole feature space is impossible due to its high dimensionality.

For this reason, researchers try to develop efficient algorithms that would measure the relevance of single features and eliminate the irrelevant ones. However, the superiority of the new proposals (as claimed by their authors) over the classical methods of feature selection has been marginal. For example, Elakkiya and Selvakumar (2020) or Yassir et al. (2020) achieved the classification accuracy 1–2% higher than the accuracy obtained by means of one of the well-established methods. Moreover, most of the efficient methods are lexicon-based and operate only in English. Some are also very expensive and complicated in computational terms. In particular, as far as the selection of features is considered, no good method of establishing an adequate number of terms has been proposed so far.

The aim of the study is to propose a novel algorithm for the text sentiment analysis which can be used with any type of document sentiment classification method. The proposed technique determines the number of terms most important for the classification purposes and verifies the efficiency of two classifying methods in a reduced term space in a corpus of documents consisting of the opinions of bank clients.

2. Classification algorithms used

As mentioned before, we employed a statistical approach to assess the performance of the proposed technique. More specifically, two methods were used, i.e. naive Bayes classifier and logistic regression (see Idczak, 2021). These methods require the establishment of input data as a set of features, which are derived from a corpus and presented as frequencies of particular terms in particular documents. This type of representation is called unigram and might be presented as the following document-term matrix (DTM):

$$\mathbf{x} = \left[x_{ij}\right], \tag{1}$$

where:
$\mathbf{x}$ is the document-term matrix,
x_{ij} is the frequency of the j-th term occurrence in the i-th document,
$i = 1, \ldots, I$ (I is the total number of documents in a training set),
$j = 1, \ldots, J$ (J is the total number of terms in a training set).

The features or terms or words, i.e. the columns of matrix $\mathbf{x}$ will be denoted by $\boldsymbol{w}_j$, which is the j-th feature or word (j-th column of matrix $\mathbf{x}$). The true class labels of documents will be denoted by $\boldsymbol{w}_0$.

2.1. Naive Bayes

Bayes' rule (Domański & Pruska, 2000) for the document sentiment classification defines a conditional probability that document $\boldsymbol{x}_i$ belongs to class C_k:

$$P(C_k|\boldsymbol{x}_i) = \frac{p_k f(\boldsymbol{x}_i|C_k)}{\sum_{k=1}^{K} p_k f(\boldsymbol{x}_i|C_k)}, \tag{2}$$

where:
C_k is the k-th class, $k = 1, \ldots, K$,
$\boldsymbol{x}_i$ is the i-th document (i-th row of matrix $\mathbf{x}$) with features J,
p_k is the prior probability that the document belongs to class C_k,
$f(\boldsymbol{x}_i|C_k)$ is the probability of occurrence of document $\boldsymbol{x}_i$, providing it belongs to class C_k.

A naive Bayes classifier assigns document $\boldsymbol{x}_i$ to class C_k if the following equation is satisfied:

$$\mathrm{P}(C_k|\boldsymbol{x}_i) = \max_k \mathrm{P}(C_k|\boldsymbol{x}_i), \tag{3}$$

which is equivalent to:

$$\mathrm{P}(C_k|\boldsymbol{x}_i) = \max_k[p_k f(\boldsymbol{x}_i|C_k)]. \tag{4}$$

The above-mentioned classification rule assumes that $\boldsymbol{w}_j$ terms are independently distributed, given the k-th class:

$$f(\boldsymbol{x}_i|C_k) = \prod_{j=1}^{J} f\left(x_{ij}\middle|C_k\right). \tag{5}$$

In order to train a naive Bayes classifier, p_k will be calculated using the following relative-frequency estimation:

$$\hat{p}_k = \frac{n_k}{I}, \tag{6}$$

where n_k is the number of documents that belongs to the k-th class,

while $f(\boldsymbol{x}_i|C_k)$ will be calculated using the subsequent relative-frequency estimation:

$$\hat{p}\left(w_j = x_{ij}|C_k\right) = \frac{n_{ijk}}{n_{jk}}, \tag{7}$$

where:
n_{ijk} is the frequency of the i-th value of the j-th term in the k-th class,
n_{jk} is the frequency of the j-th term in the k-th class.

2.2. Logistic regression

Let us assume that C is the Bernoulli random variable:

$$C \sim Bernoulli(p), \tag{8}$$

that might take one of the two values:

$$C = \begin{cases} 0, \text{when the sentiment of a document is negative,} \\ 1, \text{when the sentiment of a document is positive.} \end{cases} \tag{9}$$

Then the logistic regression (Hosmer et al., 2013) can be written as follows:

$$p = p(C = 0|\boldsymbol{x}_i) = \frac{e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i}}, \tag{10}$$

where β_0 is an intercept, and $\boldsymbol{\beta}$ is a vector of estimated parameters.

It is convenient to apply logit transformation on (10) to obtain some desirable properties of a linear model:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \boldsymbol{\beta}^t \boldsymbol{x}_i. \tag{11}$$

As a result, the above-mentioned equation is linear in its parameters, so betas have a handy interpretation in terms of the odds ratio $\left(\frac{e^{\beta_0+\boldsymbol{\beta}^t x_{i'}}}{e^{\beta_0+\boldsymbol{\beta}^t x_i}}\right)$.[1] This means that if one element of $\boldsymbol{x}_i$, x_{ij} increases by one unit (ceteris paribus), the odds ratio will increase by e^{β_j}, i.e. the odds that a document has a negative sentiment (given the increased x_{ij}) increase (decrease) by $(e^{\beta_j} - 1) \cdot 100\%$.

$p(C = 0|\boldsymbol{x}_i)$ in (10) is the probability that document $\boldsymbol{x}_i$ has a negative sentiment, thus the probability that document $\boldsymbol{x}_i$ has a positive sentiment is calculated by the following equation:

$$p(C = 1|\boldsymbol{x}_i) = 1 - p(C = 0|\boldsymbol{x}_i). \tag{12}$$

Document $\boldsymbol{x}_i$ is classified as negative if the following equation is satisfied:

$$p(C = 0|\boldsymbol{x}_i) = max[p(C = 0|\boldsymbol{x}_i),\ p(C = 1|\boldsymbol{x}_i)]; \tag{13}$$

otherwise, the document is considered positive.

Parameters from equation (10) can be estimated by means of the maximum-likelihood method by maximising the following likelihood function:

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{I} p(C_1|\boldsymbol{x}_i)^{C_i} [1 - p(C_1|\boldsymbol{x}_i)]^{1-C_i}, \tag{14}$$

with respect to parameters β_0 and $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmax}}\, \mathrm{L}(\boldsymbol{\beta}). \tag{15}$$

[1] $e^{\beta_0+\boldsymbol{\beta}^t x_{i'}}$ denotes the odds for feature $\boldsymbol{w}_j$ to be increased by one unit.

3. General description of the methodology

The approach adopted in this study involves feature selection, i.e. reducing the number of terms used in the document classification by selecting the most important ones. Thus, the method is independent of the classification method used subsequently. How to choose the most important terms or arrange all terms in the order of importance? We propose a technique which is connected with a distance-based correlation between features. The method is relatively uncomplicated and, what is most important, may be further used to determine the number of features to be selected. Within its framework, a term-priority list is created on the basis of the correlation between the terms and the document's sentiment. It is not easy to measure the correlation of this kind, therefore we tried to use a distance-based correlation coefficient. The higher the correlation between the document's sentiment distances (i.e. class label distances) and distances between the unigram (1) representation of a given term, the more important the term is. This approach proved very successful in cluster analysis with relation to distance-based correlation between sets of features (see Korzeniewski, 2012), therefore it should work in the classification of documents as well. If the 'jumps' in the document sentiment distances are positively correlated with the 'jumps' in the unigram representation of a given term, this situation resembles a positive correlation between two sets of features, and the importance of both sets for creating a possible cluster structure is emphasised. Formally, the distance-based correlation coefficient (*DBCorr*) between two sets of features A, B is given by the formula:

$$DBCorr(A,B,l) = \frac{\frac{1}{l}\sum_{t=1}^{l}(d_t^A d_t^B) - \bar{d}^A \bar{d}^B}{s^A s^B}, \tag{16}$$

where:

$1 \leq l \leq n$ denotes the number of document pairs drawn without replacement from all pairs of documents,

d_t^A, d_t^B denote distances for the t-th pair of documents coming from sets A and B, respectively, based on relevant features,

$\bar{d}^A$, $\bar{d}^B$, s^A, s^B denote arithmetic means and standard deviations computed from all l distances on both sets of features, respectively.

In the current structure of document classification, both sets of features, A and B, will be one-feature sets. Set A will consist of a feature classifying documents to one of the two classes, and set B will consist of one I-dimensional feature which is a unigram representation of a given term across all the documents in the training set. Let us establish that we will use formula (16) for $l = 50$ with the value of $DBCorr(A,B,l)$ being the arithmetic mean of 1,000 repetitions of drawing sets of documents consisting of 50 items. Thus in the further part of the text, the notation

$DBCorr(A, B)$ will be used instead of $DBCorr(A, B, l)$. We will use the Manhattan distance (a sum of absolute differences across all coordinates) on both features (or sets of features).

We propose the following formal description of the procedure of creating the feature-priority list:

- find $DBCorr(\boldsymbol{w}_0, \boldsymbol{w}_j)$ for each feature $\boldsymbol{w}_j$ present in the training set;
- rank all terms $\boldsymbol{w}_j$ according to the decreasing order of $DBCorr(\boldsymbol{w}_0, \boldsymbol{w}_j)$.

In order to present the newly-proposed algorithm for determining the number of features to be selected, at first the whole classification procedure should be run. The evaluation of the results should enable the formulation of the proposal. Therefore, our paper is further arranged as follows:

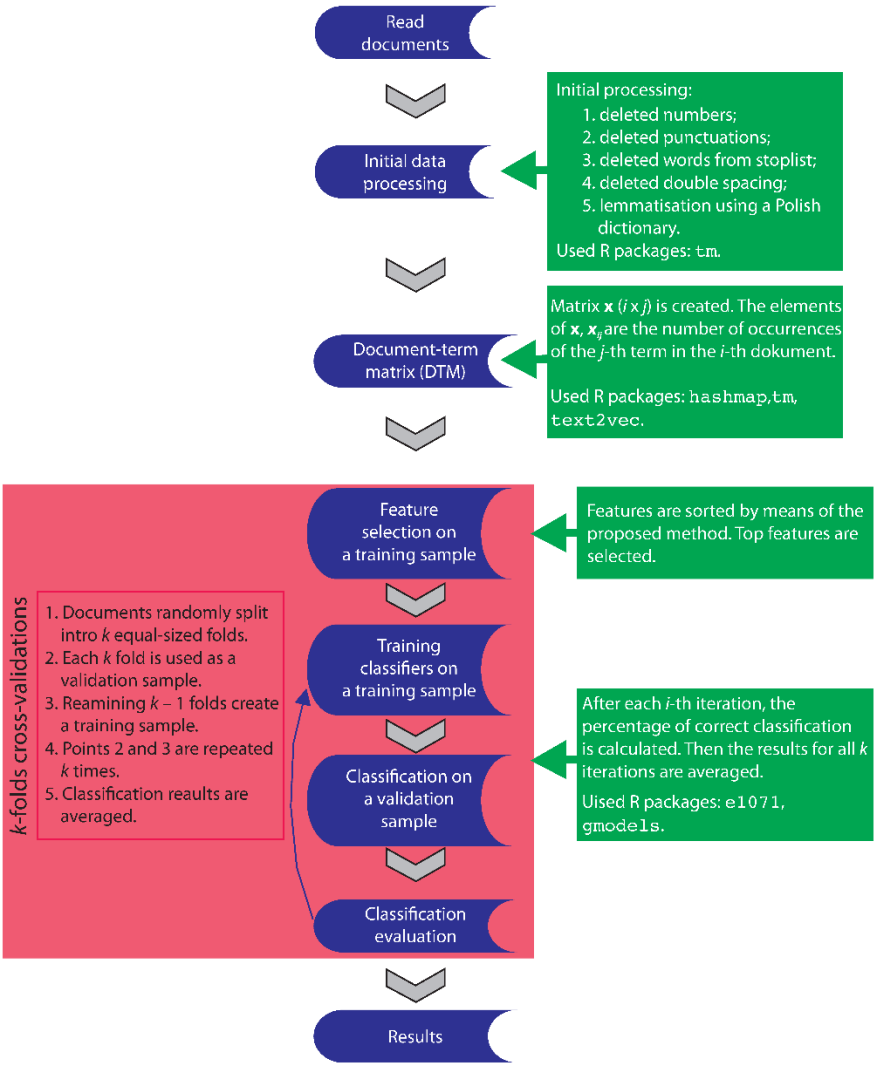
- presentation of the document classification procedure with respect to documents' sentiment;
- introduction of the new algorithm for determining the number of features (to be selected);
- assessment of the quality of the proposed algorithm by means of its application to the data sets used in the sentiment classification experiment.

4. Classification experiment

4.1. Experiment set-up

In order to assess the effectiveness of the proposed technique for the classification of a document sentiment, we conducted an experiment in line with the algorithm presented in Figure 1. All calculations were made by means of R software. First, the documents analysed were read into the memory and then initially processed, i.e. the unwanted numbers, punctuation marks and words were deleted. Lemmatisation was another, and a very important part of this step. This process groups the inflected forms of the word so that they can be analysed as a single item (word's lemma), e.g. *płakać* is lemma for *płakał, płakaliśmy, płacze*. It is especially important in the case of the Polish language, which is inflected. The lemmatisation was performed by means of functions from `tm` package in R and a Polish dictionary. This step might have a crucial impact on features (and their number) in the document-term matrix. For the purpose of this study, unigrams were considered. The DTM matrix was calculated by means of `hashmap`, `tm` and `tex2vec` packages. After the DTM was created, our novel technique was employed to sort features according to their relevance. Then the matrix was used for the purpose of 10-fold cross-validation, as shown in Figure 1, where a naive Bayes classifier and logistic regression were taught on a training sample which consisted of the most relevant features. The classification was assessed on a validation sample. This part of the algorithm was handled by `e1071` and `gmodels` packages. The classification was evaluated in terms of accuracy. A pseudocode is presented below as a supplement to Figure 1.

Figure 1. Algorithmic description of the experiment



 # Initial processing

Step 1. Delete numbers.
Step 2. Delete words from stoplist.
Step 3. Delete punctuation marks.
Step 4. Delete double spacing.
Step 5. Lemmatise each term using the Polish dictionary.
 # Creating document-term matrix
Step 6. Create document-term matrix **x** containing unigrams.
 # Creating k-fold validation structure
Step 7. Split all documents in k equal-sized folds. Each fold will be used as a validation set and the remaining k–1 folds will create a training set.

 # Ordering features with respect to their importance to documents' sentiment
Step 8. $x_{ij} \leftarrow frequency\ of\ term\ j\ in\ document\ i;$
$x_{i0} \leftarrow sentiment\ of\ document\ i; either\ 1\ or\ 0;$
Step 9. Draw dependently set of 50 documents from the training set.
$d_{\cdot}^{j} \leftarrow distance\ on\ feature\ j\ between\ a\ pair\ documents\ in\ the\ set;$
$d_{\cdot}^{0} \leftarrow distance\ between\ class\ labels\ for\ a\ pair\ of\ documents\ in\ the\ set;$
$DBCorr(\boldsymbol{w}_j, \boldsymbol{w}_0) \leftarrow \frac{\frac{1}{I}\sum_{t=1}^{I}(d_t^j d_t^0) - \bar{d}^j \bar{d}^0}{s^j s^0}$
Step 10. Repeat Step 9 1,000 times to find the average $\overline{DBCorr}(\boldsymbol{w}_j, \boldsymbol{w}_0)$ for each $j = 1, \ldots, J$.
Step 11. Arrange all J features in decreasing order of $\overline{DBCorr}(\boldsymbol{w}_j, \boldsymbol{w}_0)$
Step 12. Train classifier on the training set.
Step 13. Classify documents from the validation set. Find the percentage of correct classifications.
Step 14. Repeat Step 8 to Step 13 for each fold k. Find the average percentage of correct classifications.
Source: authors' work.

4.2. Data sets

The proposed method was verified on three datasets. The documents were client reviews on one of Poland-based banks. Each document was labelled with a positive or negative sentiment (positive or negative class). These labels were assigned manually by an opinion holder (a client) by choosing a happy- or a sad-face icon. Each dataset consisted of 302 features. There were 192 positive and 198 negative documents in the first dataset, 198 positive and 192 negative documents in the second dataset, and 188 positive and 201 negative documents in the third dataset.

4.3. Classification results

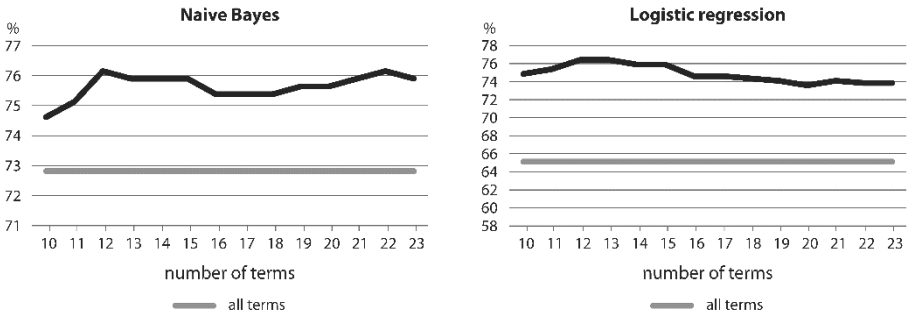
A novel feature-ordering method was applied to the three datasets comprising clients' reviews. The classification was performed by means of the naive Bayes classifier and the logistic regression. The results show that the proposed feature selection method for $k = 10, 11, \ldots, 23$ outperformed the classification based on all terms from all the three datasets (see figures 2–4), although there were some differences. The effectiveness of classification was assessed in terms of accuracy:

$$accuracy = \frac{TP + TN}{I}, \tag{17}$$

where:
TP is the number of documents with a positive sentiment, classified as positive,
TN is the number of documents with a negative sentiment, classified as negative,
I is the number of classified documents.

According to Figure 2, both the naive Bayes and the logistic regression on $k = 10, 11, \ldots, 23$ most relevant features from dataset 1 achieved better results (on average 75.64% for the naive Bayes and 74.85% for the logistic regression) than the classification based on all terms from this set (72.82% and 65.13%, respectively).

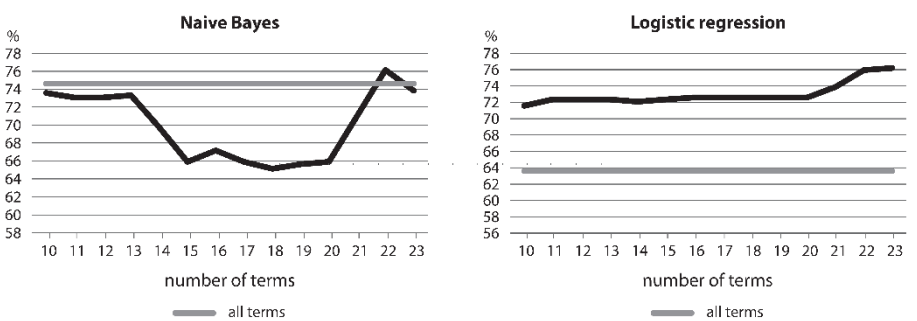
Figure 2. Accuracy of classification – dataset 1



Source: authors' calculation based on dataset 1.

The proposed feature selection method performed slightly poorer than the naive Bayes (on average 69.96% vs 74.62%) on dataset 2 (see Figure 3). It is worth mentioning that for $k = 22$, the adopted procedure still outperformed the classification based on all terms (76.15% vs 74.62%). On the other hand, the results of logistic regression were more consistent, i.e. k most relevant features yielded 72.97% vs 63.59% on average.

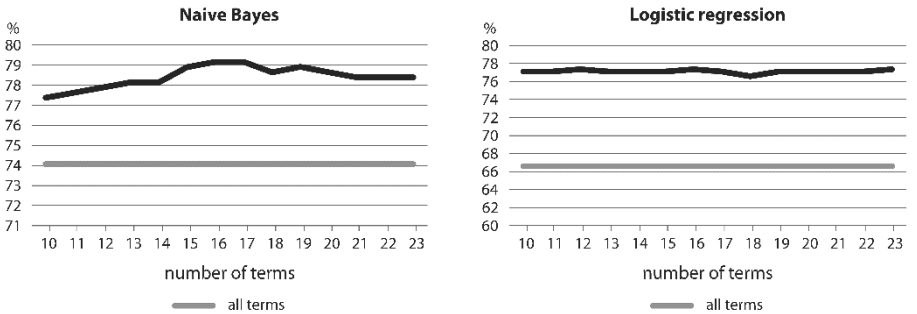
Figure 3. Accuracy of classification – dataset 2



Source: authors' calculation based on dataset 2.

The results obtained from dataset 3 (Figure 4) also confirm the usefulness of the method. In all the cases ($k = 10, 11, \ldots, 23$), both the naive Bayes and the logistic regression performed better than the classification based on all terms. Naive Bayes classified 78.41% of the documents correctly, while the all-terms approach did so in 74.06% cases. Logistic regression was also successful, having classified 77.13% of documents correctly, compared to 66.59% correct classifications by the all-terms approach.

Figure 4. Accuracy of classification – dataset 3



Source: authors' calculation based on dataset 3.

Detailed results of the classification are presented in Table 1. The highest accuracy scores for a given classifier are bolded.

Table 1. Detailed classification results on three datasets

Number of features (*k*)	Dataset 1		Dataset 2		Dataset 3	
	naive Bayes	logistic regression	naive Bayes	logistic regression	naive Bayes	logistic regression
10	74.62	74.87	73.59	71.54	77.37	77.11
11	75.13	75.38	73.08	72.31	77.62	77.11
12	**76.15**	**76.41**	73.08	72.31	77.88	**77.36**
13	75.90	**76.41**	73.33	72.31	78.14	77.11
14	75.90	75.90	69.74	72.05	78.14	77.11
15	75.90	75.90	65.90	72.31	78.90	77.11
16	75.38	74.62	67.18	72.56	**79.16**	**77.36**
17	75.38	74.62	65.90	72.56	**79.16**	77.11
18	75.38	74.36	65.13	72.56	78.64	76.59
19	75.64	74.10	65.64	72.56	78.91	77.11
20	75.64	73.59	65.90	72.56	78.65	77.11
21	75.90	74.10	71.03	73.85	78.39	77.11
22	**76.15**	73.85	**76.15**	75.90	78.39	77.11
23	75.90	73.85	73.85	**76.15**	78.39	**77.36**
All features	72.82	65.13	74.62	63.59	74.06	66.59

Source: authors' calculation based on datasets 1–3.

Table 1 shows that within dataset 1, the highest accuracy for the naive Bayes was obtained for $k = 12$ and $k = 22$ (76.15%), and it was higher by about 3.3 p.p. than the accuracy achieved by means of the all-terms classification. As for logistic regression, the best result was obtained for $k = 12$ and $k = 13$ (76.41%), and it was higher by approximately 11.3 p.p. than the result of the all-terms classification.

Within dataset 2, the highest accuracy for the naive Bayes was obtained for $k = 22$ (76.15%), which was higher by about 1.5 p.p. than in the case of the all-terms

classification. As for logistic regression, the highest accuracy was achieved for $k = 23$ (76.15%), which was higher by approximately 12.6 p.p. than the highest accuracy obtained by means of the all-terms classification.

As regards dataset 3, the highest accuracy for the naive Bayes was obtained for $k = 16$ and $k = 17$ (79.16%), which was higher by about 5.1 p.p. than the accuracy achieved by the all-terms classification. As for logistic regression, the highest accuracy was obtained for $k = 12, 16, 23$ (77.36%), and it was higher by approximately 10.8 p.p. than the accuracy achieved by the all-terms classification.

5. Algorithm for determining the number of features

How to find the adequate number k of features that should be selected? Let us assume that the criterion for the evaluation of the effectiveness of such an algorithm is based solely on the accuracy of the subsequent document classification. The already-verified fact that the classification based on a narrower number of selected features yielded better results than the one based on all features does not conclude this research. There are numerous other algorithms for reducing feature space in the document sentiment classification, and some of them might even give better results than our approach, but none of them is able to predict the adequate number of features to be selected. Distance-based correlation will be further used in such a way as to allow the prediction of the optimal number of features to be selected. The subject of determining the smallest possible number of features has not been investigated extensively in the literature so far, and to our best knowledge, no effective method to this end has yet been discovered.

As the use of single-feature sets is too weak to trace any connections between the number of features and the document sentiment labels, we propose computing distance-based correlations given by formula (16) for subsets A, consisting of two features, and B, consisting of one feature (document sentiment labels). Generally, investigating distance-based correlation for feature subsets consisting of several features does not make sense due to the curse of dimensionality. But in the current structure, there are only two features in one set and one feature in the other. Secondly, the features are already ordered in the decreasing order with respect to their one-to-one correlations with the sentiment labels. If there is an optimal number k of features, then, logically, the distance-based correlations between the 'better' (more meaningful) features positioned to the left of k should be much stronger than those of the features positioned to the right of k. Therefore, we propose to compute distance-based correlations for all the pairs of features from the set of $k-1$ features to the left of k and find the arithmetic mean $r1$ of these

correlations. In a similar manner, distance-based correlations for all pairs of features from the set of $k-1$ features to the right of k should be calculated, and the arithmetic mean $r2$ of these correlations should be computed.

Subsequently, we have to investigate the flow of $r1-r2$, i.e. the differences between the two correlations for $k = 5, 6, \ldots, 23$, and choose the k corresponding to the last local maximum. The algorithm formulated in this way needs determining the range of features arranged in a sequence from which the best candidate will be selected. In the case of our three data sets, the range $k = 5, 6, \ldots, 23$ was chosen because further features have very weak (below 0.005) distance-based correlation with document class labels.

Figure 5. The pseudocode of the algorithm for determining the optimal number of features

 # Selecting the optimal number of initial features from the set of all features arranged with respect to # their decreasing relevance to the document's sentiment

Step 1. Draw dependently a set of 50 documents from the training set.

$d_t \leftarrow distance\ on\ two\ features\ i, j\ between\ documents\ from\ pair\ t;$
$\bar{d} \leftarrow mean\ of\ distances\ d_t\ from\ all\ pairs\ t;$
$s \leftarrow standard\ deviation\ of\ distances\ d_t\ from\ all\ pairs\ t;$
$d_t^0 \leftarrow distance\ between\ class\ labels\ for\ pair\ t\ of\ documents;$
$\bar{d}^0 \leftarrow mean\ of\ distances\ d_t^0\ from\ all\ pairs\ t;$
$s^0 \leftarrow standard\ deviation\ of\ distances\ d_t^0\ from\ all\ pairs\ t;$

$$DBCorr(w_{ij}, \boldsymbol{w}_0) \leftarrow \frac{\frac{1}{l}\sum_{t=1}^{l}(d_t d_t^0) - \bar{d}\bar{d}^0}{s \cdot s^0}$$

Step 2. Repeat Step 1 1,000 times to find the average $\overline{DBCorr}(w_{ij}, \boldsymbol{w}_0)$.

Step 3. Find the 'left' mean $r1 = \frac{\sum_{i<j<k} \overline{DBCorr}(w_{ij}, \boldsymbol{w}_0)}{0.5 \cdot (k-1) \cdot (k-2)}$ for every $k = 11, 12, \ldots, 23$.

Step 4. Find the 'right' mean $r2 = \frac{\sum_{k<i<j\leq 2k} \overline{DBCorr}(w_{ij}, \boldsymbol{w}_0)}{0.5 \cdot (k-1) \cdot (k-2)}$ for every $k = 11, 12, \ldots, 23$.

Step 5. Select $k \in \{11, 12, \ldots, 23\}$ which corresponds to the last local maximum of $r1 - r2$.

Source: authors' work.

6. Assessment of the algorithm

The new algorithm was be applied to the three data sets that had already been classified. In Table 2, we present the detailed differences between mean correlations and single correlations for the most meaningful features. It is much easier, however, to investigate these numbers in graphical form (we took this into account and Figure 5 consists of three graphs).

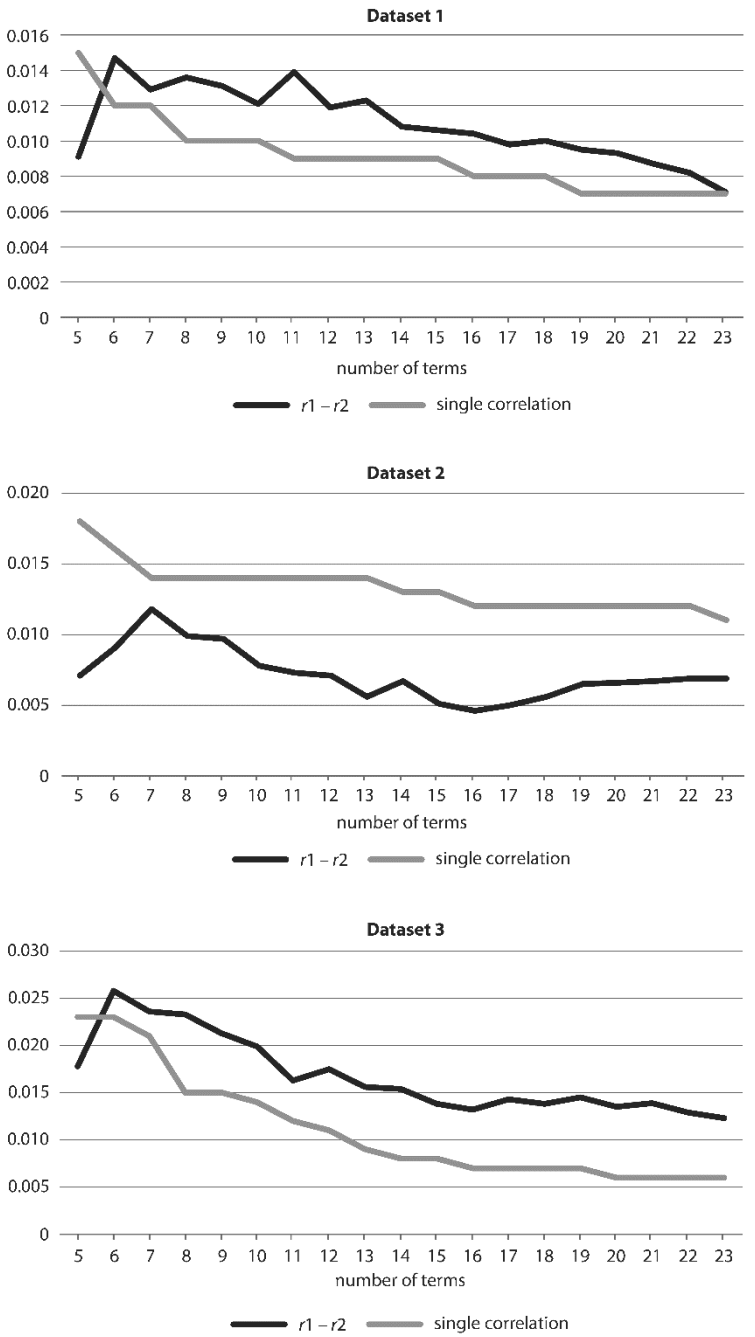
Table 2. Differences between means of correlations and single-feature correlations

Number of features (k)	Dataset 1		Dataset 2		Dataset 3	
	$r1 - r2$	single correlations	$r1 - r2$	single correlations	$r1 - r2$	single correlations
5	0.0091	0.015	0.0071	0.018	0.0178	0.023
6	0.0147	0.012	0.0091	0.016	0.0258	0.023
7	0.0129	0.012	0.0118	0.014	0.0236	0.021
8	0.0136	0.010	0.0099	0.014	0.0233	0.015
9	0.0131	0.010	0.0097	0.014	0.0213	0.015
10	0.0121	0.010	0.0078	0.014	0.0199	0.014
11	0.0139	0.009	0.0073	0.014	0.0163	0.012
12	0.0119	0.009	0.0071	0.014	0.0175	0.011
13	0.0123	0.009	0.0056	0.014	0.0156	0.009
14	0.0108	0.009	0.0067	0.013	0.0154	0.008
15	0.0106	0.009	0.0051	0.013	0.0138	0.008
16	0.0104	0.008	0.0046	0.012	0.0132	0.007
17	0.0098	0.008	0.0050	0.012	0.0143	0.007
18	0.0100	0.008	0.0056	0.012	0.0138	0.007
19	0.0095	0.007	0.0065	0.012	0.0145	0.007
20	0.0093	0.007	0.0066	0.012	0.0135	0.006
21	0.0087	0.007	0.0067	0.012	0.0139	0.006
22	0.0082	0.007	0.0069	0.012	0.0129	0.006
23	0.0071	0.007	0.0069	0.011	0.0123	0.006

Source: authors' calculations based on datasets 1–3.

The first striking observation is that the only set with single-feature correlations higher than differences $r1 - r2$ is dataset 2. This is not very surprising given that (as e.g. the naive Bayes classifier suggests) this set might prove quite difficult for classification (due to the uncertainty as to what number of features should be chosen). For dataset 1, the algorithm performed well, because it was clear that the last local maximum corresponded to $k = 18$. The choice of 18 was probably not the best one; 13, 14 or even 22 would be better in view of the numbers from Table 1. However, the classification accuracy for the most relevant initial features of $k = 18$ was higher than the classification accuracy for all the features. As regards dataset 2, the results were slightly more difficult to interpret, because the first region of the local maxima started with $k = 7$ and ended with $k = 14$ or $k = 15$. Later on, however, i.e. for $k \geq 22$, the numbers started growing again. Thus, the assumption that $k = 22$ or $k = 23$ runs in line with the algorithm formulation. Numbers from Table 1 confirm the above. For example, in the case of the naive Bayes classifier, selecting 22 was one of few options that guaranteed relatively effective classification compared to the performance of all the features. For the logistic regression likewise – the best possible choice was 22 or 23. For dataset 3, the algorithm selected $k = 21$, which, by no means being the best choice, was anyway better than the result for all the features in the light of numbers from Table 1 for both classifiers.

Figure 6. Comparison of differences between the means and individual feature correlations



Source: authors' calculation based on datasets 1–3.

7. Conclusions

In this paper, we propose a simple technique for the text sentiment classification based on unigram models. Essentially, it consists in feature-filtering by means of distance-based correlation. The correlation is measured for each term-feature $\boldsymbol{w}_i$ between the distances between text document sentiment labels and the distances between the frequencies of occurrence of terms across all documents. It is possible to order the terms, so that the impact of the curse of dimensionality becomes softened. This is because it is sufficient to find the correlation for two groups of terms positioned prior and posterior to the currently-assessed term-feature $\boldsymbol{w}_i$. Thus, the proposed procedure is computationally non-complex. Its other advantages include the fact that it enables users to customise the algorithm and choose any classifier, and that it is lexicon-free and easy to use.

The experimental results based on three datasets of clients' reviews show that the proposed method yields better results than the standard full bag-of-word approach. Moreover, the proposed distance-based correlation algorithm for ordering features according to their significance for determining the document sentiment allows the application of distance-based correlation also to the choice of the best number of initial (most meaningful) features that would ensure the effective classification. This algorithm consists in comparing distance-based correlation for two-feature subsets positioned to the left and to the right of any single feature that already has a place in the sequence of features.

The only limitation of this algorithm are situations in which the corpus of documents and the resulting set of features are so large that distance-based correlations between single features and the documents sentiment labels are too weak (< 0.005) to make reasonable ordering of features possible.

Our technique may be used to upgrade any text classifier with respect to text sentiment. We believe the fact that our approach uses distance-based correlations between terms and text-sentiment labels opens it for further development. More specifically, it might be applied to larger text corpora and sets of terms, providing that that one carefully uses the number of terms included in the set of terms for the correlation assessment. Ever-faster algorithms become increasingly desirable in modern societies, where time efficiency is crucial for the effective performance of the growing amount of online work (both clerical and analytical). Such work is often based on the recognition of the sentiment of text documents.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. W: *LSM '11: Proceedings of the Workshop on Languages in Social Media* (s. 30–38). Association for Computational Linguistics.

Davies, A., & Ghahramani, Z. (2011). Language-independent Bayesian sentiment mining of Twitter. W: *The fifth SNAKDD Workshop 2011 on Social Network Mining and Analysis* (s. 99–106).

Domański, C., & Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne.

Elakkiya, E., Selvakumar, S. (2020). GAMEFEST: Genetic Algorithmic Multi Evaluation measure based FEature Selection Technique for social network spam detection. *Multimed Tools and Application*, *79*(11–12), 7193–7225. https://doi.org/10.1007/s11042-019-08334-1.

Govindarajan, M. (2013). Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm. *International Journal of Advanced Computer Research*, *3*(4), 139–145. https://accentsjournals.org/PaperDirectory/Journal/IJACR/2013/12/21.pdf.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387.

Idczak, A. P. (2021). Sentiment Classification of Bank Clients' Reviews Written in the Polish Language. *Acta Universitatis Lodziensis. Folia Oeconomica*, (2), 43–56. https://doi.org/10.18778/0208-6018.353.03.

Iqbal, F., Hashmi, J. M., Fung, B. C. M., Batool, R., Khattak, A. M., Aleem, S., & Hung, P. C. K. (2019). A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction. *IEEE Access, 7*, 14637–14652. http://doi.org/10.1109/ACCESS.2019.2892852.

Khan, A., Baharudin, B., & Khan, K. (2011). Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews. *Trends in Applied Sciences Research*, *6*(10), 1141–1157. https://doi.org/10.3923/tasr.2011.1141.1157.

Korzeniewski, J. (2012). *Metody selekcji zmiennych w analizie skupień. Nowe procedury.* Wydawnictwo Uniwersytetu Łódzkiego. http://dx.doi.org/10.18778/7525-695-6.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG!. *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media*, *5*(1), 538–541. https://doi.org/10.1609/icwsm.v5i1.14185.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Njølstad, P. C. S., Høysæter, L. S., Wei, W., & Gulla, J. A. (2014). Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News. W: *WI-IAT '14: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (p. 71–78). IEEE. https://doi.org/10.1109/WI-IAT.2014.82.

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. *Artificial Intelligence Review*, *54*(8), 6149–6200. https://doi.org/10.1007/s10462-021-09970-6.

Yassir, A. H., Mohammed, A. A., Alkhazraji, A. A. J., Hameed, M. E., Talib, M. S., & Ali, M. F. (2020). Sentimental classification analysis of polarity multi-view textual data using data mining techniques. *International Journal of Electrical & Computer Engineering* (2088–8708), *10*(5), 5526–5533. http://doi.org/10.11591/ijece.v10i5.pp5526-5534.

Yazdani, S. F., Murad, M. A. A., Sharef, N. M., Singh, Y. P., & Latiff, A. R. A. (2017). Sentiment Classification of Financial News Using Statistical Features. *International Journal of Pattern Recognition and Artificial Intelligence, 31*(3), 1–34. https://doi.org/10.1142/S0218001417500069.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), 58–60
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

NOWOŚCI WYDAWNICZE W ZBIORACH CENTRALNEJ BIBLIOTEKI STATYSTYCZNEJ
NEW PUBLICATIONS IN THE CENTRAL STATISTICAL LIBRARY RESOURCES

W zbiorach Centralnej Biblioteki Statystycznej im. Stefana Szulca dostępne są następujące, warte polecenia publikacje:
The resources of The Stefan Szulc Central Statistical Library offer the following, highly recommendable publications:

Olga Smalej

Kultura bezdzietności. Niska dzietność i bezdzietność z wyboru w perspektywie społeczno-ekonomicznej
Culture of childlessness. Low fertility and childlessness by choice in a socio-economic perspective

W monografii przedstawiono zagadnienie niskiego poziomu urodzeń w krajach europejskich w kontekście ich rozwoju społeczno-gospodarczego.

Język: polski
Language: Polish
Wydawnictwo/Publisher: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej
Miejsce i rok wydania / Place and year of publication: Lublin 2022
Liczba stron / Number of pages: 243

We współczesnej Europie obserwujemy rosnącą elastyczność podejścia do tradycyjnego systemu wartości, pełnienia funkcji w społeczeństwie i wzorców zachowania. Posiadanie dzieci nie jest już naturalnym biegiem wydarzeń, lecz staje się kwestią wyboru uwarunkowanego wieloaspektową analizą kosztów i korzyści. Autorka *Kultury bezdzietności* przybliża czytelnikom temat bezdzietności intencjonalnej, czyli celowego zaniechania prokreacji.

Niski wskaźnik urodzeń, cechujący większość krajów Europy, jest silnie powiązany z bezdzietnością. W opracowaniu omówiono oba te zjawiska na przykładzie wybranych krajów w szerokiej perspektywie społeczno-gospodarczej. Wskazano główne przyczyny spadku urodzeń: wysoce skuteczną antykoncepcję, wymogi rynku pracy i trudności w pogodzeniu obowiązków zawodowych z życiem rodzinnym, niedostosowanie polityki państwa do potrzeb społeczeństwa (m.in. związanych z rosną-

cymi kosztami wychowania dzieci i opieki nad nimi), zmiany cyklu życia jednostek i ról społecznych, a także niestabilność związków.

Książka składa się z trzech rozdziałów. W pierwszym wskazano determinanty spadku wartości współczynnika dzietności w skali mikro i makro oraz możliwe konsekwencje tego procesu. W drugim omówiono zjawisko bezdzietności z wyboru i przeanalizowano jego uwarunkowania. Trzeci rozdział poświęcono opinii społecznej na temat bezdzietności intencjonalnej. Olga Smalej – na podstawie wyników własnego badania ankietowego – przedstawiła motywy decyzji o nieposiadaniu dzieci i odczucia społeczne związane z tym zjawiskiem oraz określiła wielkość dystansu społecznego do osób bezdzietnych z wyboru. Przytoczyła także opinie respondentów dotyczące pomocy psychologicznej i propozycji wprowadzenia dodatkowego podatku dla osób świadomie rezygnujących z rodzicielstwa.

W monografii wykorzystano m.in. analizy i dane statystyczne Banku Światowego, Organizacji Współpracy Gospodarczej i Rozwoju, Eurostatu i Głównego Urzędu Statystycznego oraz bogatą literaturę. Publikację uzupełniają materiały graficzne i zestawienia tabelaryczne.

Klara Dygaszewicz, Piotr Zapadka

Prawo statystyki publicznej
Law of public statistics

W publikacji omówiono prawne aspekty funkcjonowania polskiej statystyki publicznej.

Język: polski
Language: Polish
Wydawnictwo/Publisher: Polskie Wydawnictwo Ekonomiczne
Miejsce i rok wydania / Place and year of publication: Warszawa 2021
Liczba stron / Number of pages: 192

Statystyka publiczna dostarcza wiarygodne i rzetelne dane statystyczne, a tym samym odgrywa istotną rolę w rozwoju społeczeństwa informacyjnego. Między innymi wspiera strategiczne procesy zarządzania państwem w zakresie życia społecznego, gospodarczego i kwestii demograficznych. Podstawowym aktem prawnym, który reguluje prawno-instytucjonalne funkcjonowanie statystyki publicznej w Polsce, jest Ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej. Określono w niej zasady rzetelnego, obiektywnego, profesjonalnego i niezależnego prowadzenia badań staty-

stycznych, których wyniki mają charakter oficjalnych danych, a także organizację i tryb prowadzenia tych badań i związane z nimi obowiązki.

Publikacja składa się z sześciu rozdziałów. W pierwszym zaprezentowano przepisy ogólne dotyczące regulacji, źródeł i celów prawa statystyki publicznej, także w odniesieniu do prawodawstwa międzynarodowego. Drugi rozdział został poświęcony organizacji badań statystycznych prowadzonych w ramach statystyki publicznej. Omówiono m.in. program tych badań, źródła ich finansowania, strukturę służb statystyki publicznej oraz zakres współdziałania Prezesa Głównego Urzędu Statystycznego z Prezesem Narodowego Banku Polskiego. W trzecim rozdziale wyjaśniono kwestie związane z tajemnicą statystyczną. Podkreślono jej rolę w budowaniu zaufania społecznego do statystyki publicznej i konieczność przestrzegania tajemnicy w kontekście prowadzenia badań. W czwartym rozdziale podjęto szczegółowe zagadnienia badań statystycznych, takie jak ich przedmiot, dostępne źródła danych oraz proces gromadzenia, opracowywania i udostępniania informacji statystycznych. Przetwarzanie danych osobowych do celów statystycznych i ochrona danych to tematy ujęte w rozdziale piątym. W ostatnim rozdziale skupiono się na przepisach karnych zawartych w ustawie o statystyce publicznej. W podsumowaniu autorzy podkreślili znaczenie oficjalnych statystyk dla funkcjonowania państwa w różnych jego aspektach.

Opracowanie zawiera bogatą bibliografię, składającą się m.in. z aktów prawa międzynarodowego, europejskiego i krajowego, a także wykazu orzeczeń.

Dorota Kierska
Centralna Biblioteka Statystyczna, Polska / Central Statistical Library, Poland

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(5), 61–62
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

WYDAWNICTWA GUS. KWIECIEŃ 2023
PUBLICATIONS OF STATISTICS POLAND. APRIL 2023

W ofercie wydawniczej Głównego Urzędu Statystycznego z ubiegłego miesiąca warto zwrócić uwagę na następującą publikację:
Among Statistics Poland's last month's publications, we would like to recommend:

Przedsiębiorstwa niefinansowe powstałe w latach 2017–2021
Non-financial enterprises established in 2017–2021

Najnowsza edycja publikacji poświęconej funkcjonowaniu na rynku przedsiębiorstw niefinansowych w ciągu pierwszych pięciu lat ich istnienia.

Język: polski, angielski
Language: Polish, English
Seria: Analizy statystyczne
Series: Statistical analyses
Dostępne wersje: drukowana i elektroniczna z tablicami w formacie Excel
Available in: printed and electronic form with Excel tables

Opracowanie składa się z dwóch rozdziałów. W pierwszym podano informacje na temat całej zbiorowości przedsiębiorstw niefinansowych powstałych w latach 2017–2021. Drugi (podzielony na pięć podrozdziałów dotyczących podmiotów jedno-, dwu-, trzy-, cztero- i pięcioletnich) zawiera dane o warunkach powstania, działaniu i perspektywach rozwojowych przedsiębiorstw niefinansowych. Populacje przedsiębiorstw zostały scharakteryzowane według przeważającego rodzaju działalności, formy prawnej, wielkości zatrudnienia i siedziby jednostki. Podano m.in. informacje o liczbie przedsiębiorstw, wskaźnikach przeżycia, zasięgu prowadzonej działalności, napotkanych trudnościach, planowanych zmianach oraz sytuacji finansowej. W podrozdziale poświęconym przedsiębiorstwom powstałym w 2020 r. dodatkowo uwzględniono czynniki przedsiębiorczości.

Dane wykorzystane w publikacji pochodzą m.in. z rocznego badania działalności gospodarczej przedsiębiorstw.

W kwietniu br. ukazały się ponadto:
- *Aktywność ekonomiczna ludności Polski – 4 kwartał 2022 roku*;
- „Biuletyn statystyczny" nr 3/2023;
- *Ceny robót budowlano-montażowych i obiektów budowlanych (luty 2023 r.)*;
- *Działalność przedsiębiorstw posiadających jednostki zagraniczne w 2021 r.*;

- *Koniunktura w przetwórstwie przemysłowym, budownictwie, handlu i usługach 2000–2023 (kwiecień 2023)*;
- *Ludność. Stan i struktura ludności oraz ruch naturalny w przekroju terytorialnym w 2022 r. Stan w dniu 31 grudnia*;
- *Nakłady i wyniki przemysłu w 2022 r.*;
- *Produkcja upraw rolnych i ogrodniczych w 2022 r.*;
- *Produkcja ważniejszych wyrobów przemysłowych w marcu 2023 r.*;
- *Sytuacja społeczno-gospodarcza kraju w 1 kwartale 2023 r.*;
- *Sytuacja społeczno-gospodarcza województw Nr 4/2022*;
- *Zatrudnienie i wynagrodzenia w gospodarce narodowej w 2022 r.*;
- *Zeszyt metodologiczny. Rachunek podaży i wykorzystania wyrobów i usług.*

Joanna Sadowy
Główny Urząd Statystyczny, Departament Opracowań Statystycznych, Polska
Statistics Poland, Statistical Products Department, Poland

Wszystkie publikacje GUS w wersji elektronicznej są dostępne na stronie stat.gov.pl/publikacje/publikacje-a-z. Wersje drukowane (jeśli zostały wydane) można zamawiać pod adresem: zws-sprzedaz@stat.gov.pl.
All the publications of Statistics Poland available in electronic form can be accessed at stat.gov.pl/en/publications. Printed versions (if available) may be ordered at: zws-sprzedaz@stat.gov.pl.

DLA AUTORÓW
FOR THE AUTHORS

(for the English translation of the information given below, please visit ws.stat.gov.pl/ForAuthors)

W „Wiadomościach Statystycznych. The Polish Statistician" („WS") zamieszczane są artykuły o charakterze naukowym poświęcone teorii i praktyce statystycznej, które prezentują wyniki oryginalnych badań teoretycznych lub analitycznych wykorzystujących metody statystyki matematycznej, opisowej bądź ekonometrii. Ukazują się również artykuły przeglądowe, recenzje publikacji naukowych oraz inne opracowania informacyjne. W czasopiśmie publikowane są prace w języku polskim i angielskim.

Od 2007 r. „WS" znajdują się na liście czasopism naukowych MEiN. Zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych „WS" otrzymały 70 punktów.

„Wiadomości Statystyczne. The Polish Statistician" są udostępniane w następujących bazach, repozytoriach, katalogach i wyszukiwarkach: Agro, BazEkon, Biblioteka Nauki, Central and Eastern European Academic Source (CEEAS), Central and Eastern European Online Library (CEEOL), Central European Journal of Social Sciences and Humanities (CEJSH), Directory of Open Access Journals (DOAJ), EBSCO Discovery Service, European Reference Index for the Humanities and Social Sciences (ERIH Plus), Exlibris Primo, Google Scholar, ICI Journals Master List, ICI World of Journals, Norwegian Register for Scientific Journals and Publishers (The Nordic List) oraz Summon.

Za publikację artykułów na łamach „WS" autorzy nie otrzymują honorariów ani nie wnoszą opłat.

1. Zgłaszanie artykułów

Prace przeznaczone do opublikowania w „WS" należy przesyłać za pośrednictwem platformy Editorial System: www.editorialsystem.com/ws.

Zgłaszany artykuł powinien być zanonimizowany, tj. pozbawiony informacji o autorze/autorach (również we właściwościach pliku), podziękowań i informacji o źródłach finansowania, a także innych informacji wskazujących na afiliację lub umożliwiających zidentyfikowanie autora. Jeżeli w pracy występują tablice, wykresy lub mapy, powinny być umieszczone w treści artykułu. Materiały graficzne, razem z danymi do nich, należy ponadto załączyć jako osobny plik / osobne pliki, najlepiej w formacie Excel. **Prosimy o niestosowanie stylów i ograniczenie formatowania do wymogów redakcyjnych.** Więcej informacji w pkt 4 *Wymogi redakcyjne*.

Razem z artykułem należy przesłać skan/zdjęcie oświadczenia o oryginalności pracy i niezłożeniu jej w innym wydawnictwie. **Załączenie oświadczenia jest warunkiem poddania pracy ocenie wstępnej i skierowania do recenzji.**

Zgłoszenie artykułu do opublikowania w „WS" oznacza zgodę na jego udostępnienie na licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0).

Autorzy mają prawo do samodzielnego umieszczania w wybranych przez siebie repozytoriach artykułu w wersji zarówno zgłoszonej do „WS", jak i zaakceptowanej do opublikowania

oraz opublikowanej, z zastrzeżeniem wymogu niezwłocznego podania w repozytorium informacji o numerze „WS", w którym praca się ukazała, wraz z linkiem do niej (DOI).

2. Przebieg prac redakcyjnych

Zgłoszony artykuł jest oceniany i opracowywany w czteroetapowym procesie:

1. **Ocena wstępna**, dokonywana przez redakcję. Polega na weryfikacji naukowego charakteru artykułu oraz jego struktury i zawartości pod kątem wymogów redakcyjnych, a także zgodności tematyki z profilem czasopisma. Autor uzupełnia i poprawia artykuł stosownie do uwag redakcji, a w przypadku nieuwzględnienia danej uwagi uzasadnia swoje stanowisko. Warunkiem skierowania pracy do recenzji jest potwierdzenie oryginalności tekstu uzyskane za pomocą systemu antyplagiatowego. W przypadku wykrycia znacznego podobieństwa do innych prac artykuł zostanie odrzucony.
2. **Ocena recenzentów**, dokonywana przez specjalistów w danej dziedzinie. Artykuł oceniają dwaj recenzenci spoza jednostki naukowej, przy której afiliowany jest autor; w przypadku pracy w języku angielskim co najmniej jeden recenzent jest afiliowany przy jednostce zagranicznej. W razie sprzecznych opinii dwóch recenzentów powoływany jest trzeci recenzent. Recenzenci kierują się kryteriami oryginalności i jakości opracowania zarówno w odniesieniu do treści, jak i formy artykułu.

   Autorzy artykułów, które otrzymały pozytywne oceny, wprowadzają poprawki zalecane przez recenzentów i przesyłają zmodyfikowaną wersję pracy. Jeśli pojawi się różnica zdań dotycząca zasadności proponowanych zmian, autorzy są zobligowani do uzasadnienia swojego stanowiska.
3. **Ocena Kolegium Redakcyjnego** (KR), decydująca o przyjęciu pracy do publikacji. Jest dokonywana na podstawie recenzji, z uwzględnieniem opinii redaktorów tematycznego i merytorycznego. Polega m.in. na weryfikacji dokonania przez autora zmian w artykule stosownie do uwag recenzentów. KR ocenia artykuł pod względem poprawności i spójności merytorycznej oraz zaleca autorowi wprowadzenie poprawek, jeśli są one konieczne, aby praca spełniała wymogi czasopisma. Autorowi przysługuje prawo do odwołania od decyzji o niepublikowaniu artykułu. W takim przypadku powinien on skontaktować się z redakcją „WS" i przedstawić uzasadnienie. Ostateczna decyzja w tej sprawie należy do redaktora naczelnego.

   W „WS" publikowane są wyłącznie te artykuły, które otrzymają pozytywną ocenę na każdym z wymienionych etapów i zostaną poprawione przez autora zgodnie z otrzymanymi uwagami (chyba że autor przedstawi argumenty uzasadniające nieuwzględnienie danej uwagi).

   Artykuły przyjęte przez KR do publikacji są zamieszczane na stronie internetowej czasopisma w zakładce Early View, gdzie znajdują się do czasu opublikowania w konkretnym wydaniu „WS".
4. **Opracowanie redakcyjne, autoryzacja i korekta.** Artykuł zakwalifikowany do druku jest poddawany opracowaniu merytorycznemu i językowemu. Redakcja zastrzega sobie prawo do zmiany tytułu i śródtytułów, modyfikowania tablic, wykresów i innych elementów graficznych oraz przeredagowania treści bez naruszenia zasadniczej myśli autora.

   Po opracowaniu redakcyjnym artykuł jest przesyłany do autoryzacji. Tekst zatwierdzony przez autora, po składzie i łamaniu, jest poddawany korekcie i rewizji (II korekcie).

Autor dokonuje korekty autorskiej tekstu na etapie rewizji. Wykresy i inne materiały graficzne są opracowywane na podstawie plików i danych przekazanych przez autora i poddawane korekcie i rewizji. Autor dokonuje ich akceptacji na etapie rewizji.

W przypadku odkrycia błędów w opublikowanym artykule zamieszcza się na łamach „WS" sprostowanie, a artykuł w wersji elektronicznej jest poprawiany i umieszczany na stronie internetowej „WS" ze stosownym wyjaśnieniem.

3. Zasady etyki publikacyjnej COPE

Redakcja „WS" podejmuje wszelkie starania w celu utrzymania najwyższych standardów etycznych zgodnie z wytycznymi Komitetu ds. Etyki Publikacyjnej (COPE), dostępnymi na stronie internetowej www.publicationethics.org, oraz wykorzystuje wszystkie możliwe środki mające na celu zapobieżenie nadużyciom i nierzetelności autorskiej. Przyjęte zasady postępowania obowiązują autorów, Radę Naukową, Kolegium Redakcyjne, redakcję, pracowników Wydziału Czasopism Naukowych GUS, recenzentów i wydawcę.

3.1. Odpowiedzialność autorów

1. Artykuły naukowe kierowane do opublikowania w „WS" powinny zawierać precyzyjny opis badanych zjawisk i stosowanych metod oraz autorskie wnioski i sugestie dotyczące rozwoju badań i analiz statystycznych. Autorzy powinni wyraźnie określić cel artykułu oraz jasno przedstawić wyniki przeprowadzonej analizy. Prezentacja efektów badań statystycznych zaprojektowanych i przeprowadzonych przez autorów wymaga opisania zastosowanej w nich metodologii. W przypadku nowatorskich metod analizy pożądane jest podanie przykładu ilustrującego ich zastosowanie w praktyce statystycznej. Autorzy ponoszą odpowiedzialność za treści prezentowane w artykułach. W razie zgłaszania przez czytelników zastrzeżeń odnoszących się do tych treści autorzy są zobligowani do udzielenia odpowiedzi za pośrednictwem redakcji.
2. Na autorach spoczywa obowiązek zapewnienia pełnej oryginalności przedłożonych prac. Redakcja nie toleruje przejawów nierzetelności naukowej autorów, takich jak:
   - duplikowanie publikacji – ponowne publikowanie własnego utworu lub jego części;
   - plagiat – przywłaszczenie cudzego utworu lub jego fragmentu bez podania informacji o źródle;
   - fabrykowanie danych – oparcie pracy naukowej na nieprawdziwych wynikach badań;
   - autorstwo widmo (*ghost authorship*) – nieujawnianie współautorów, mimo że wnieśli oni istotny wkład w powstanie artykułu;
   - autorstwo gościnne (*guest authorship*) – podawanie jako współautorów osób o znikomym udziale lub niebiorących udziału w opracowywaniu artykułu;
   - autorstwo grzecznościowe (*gift authorship*) – podawanie jako współautorów osób, których wkład jest oparty jedynie na słabym powiązaniu z badaniem.

   Autorzy deklarują w stosownym oświadczeniu, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i nie został złożony w innym wydawnictwie oraz że jest ich oryginalnym dziełem, i określają swój wkład w opracowanie artykułu. Jeżeli doszło do zaprezentowania podobnych materiałów podczas konferencji lub

sympozjum naukowego, to podczas składania tekstu do publikacji w „WS" autorzy są zobowiązani poinformować o tym redakcję.
3. Autorzy są zobowiązani do podania w treści artykułu wszelkich źródeł finansowania badań będących podstawą pracy.
4. Główną odpowiedzialność za rzetelność przekazanych informacji, łącznie z informacją na temat wkładu poszczególnych współautorów w powstanie artykułu, ponosi zgłaszający artykuł.
5. Autorzy zgłaszający artykuły do publikacji w „WS" biorą udział w procesie recenzji double-blind peer review, dokonywanej przez co najmniej dwóch niezależnych ekspertów z danej dziedziny. Po otrzymaniu pozytywnych recenzji autorzy wprowadzają zalecane przez recenzentów poprawki i dostarczają redakcji zaktualizowaną wersję opracowania wraz z pisemnym poświadczeniem uwzględnienia poprawek. Jeśli pojawi się różnica zdań co do zasadności proponowanych zmian, należy wyjaśnić, które poprawki zostały uwzględnione, a w przypadku ich nieuwzględnienia – uzasadnić swoje stanowisko.
6. Jeżeli autorzy odkryją w swoim maszynopisie lub tekście już opublikowanym błędy, nieścisłości bądź niewłaściwe dane, powinni niezwłocznie poinformować o tym redakcję w celu dokonania korekty, wycofania tekstu lub zamieszczenia sprostowania. W przypadku korekty artykułu już opublikowanego jego nowa wersja jest zamieszczana na stronie internetowej „WS" wraz ze stosownym wyjaśnieniem.

3.2. Odpowiedzialność Rady Naukowej, Kolegium Redakcyjnego i Wydziału Czasopism Naukowych GUS

1. Rada Naukowa (RN) kształtuje profil programowy czasopisma, określa kierunki jego rozwoju i konsultuje jego zakres merytoryczny.
2. Kolegium Redakcyjne (KR) podejmuje decyzję o publikacji danego artykułu z uwzględnieniem ocen recenzentów oraz opinii zespołu redakcyjnego. W swoich rozstrzygnięciach członkowie KR kierują się kryteriami merytorycznej oceny wartości artykułu, jego oryginalności i jasności przekazu, a także ścisłego związku z celem i zakresem tematycznym „WS". Oceniają artykuły niezależnie od płci, rasy, pochodzenia etnicznego, narodowości, religii, wyznania, światopoglądu, niepełnosprawności, wieku lub orientacji seksualnej ich autorów.
3. Zespół redakcyjny, wyodrębniony z KR, tworzą redaktor naczelny i jego zastępca, redaktorzy tematyczni i redaktor merytoryczny. Członkowie zespołu redakcyjnego weryfikują nadsyłane artykuły pod względem merytorycznym, oceniają ich zgodność z celem i zakresem tematycznym „WS" oraz sprawdzają spełnienie wymogów redakcyjnych i przestrzeganie zasad rzetelności naukowej. Ponadto wybierają recenzentów w taki sposób, aby nie wystąpił konflikt interesów, i dbają o zapewnienie uczciwego, bezstronnego i terminowego procesu recenzowania.
4. Za sprawny przebieg procesu wydawniczego, poinformowanie wszystkich jego uczestników o konieczności przestrzegania obowiązujących zasad i przygotowanie artykułów do publikacji odpowiadają pracownicy Wydziału Czasopism Naukowych (WCN) GUS. W celu uzyskania obiektywnej oceny oryginalności nadsyłanych artykułów przed skierowaniem ich do recenzji WCN wykorzystuje system antyplagiatowy. Informacje dotyczące

artykułu mogą być przekazywane przez WCN wyłącznie autorom, recenzentom, członkom RN i KR oraz wydawcy.

5. Zmiany dokonane w tekście na etapie przygotowania artykułu do publikacji nie mogą naruszać zasadniczej myśli autorów. Wszelkie modyfikacje o charakterze merytorycznym są z nimi konsultowane.
6. W przypadku podjęcia decyzji o niepublikowaniu artykułu nie może on zostać w żaden sposób wykorzystany przez wydawcę lub uczestników procesu wydawniczego bez pisemnej zgody autorów. Autorzy mogą się odwołać od decyzji o niepublikowaniu artykułu. W tym celu powinni się skontaktować z redaktorem naczelnym lub sekretarzem redakcji „WS" i przedstawić stosowną argumentację. Odwołania autorów są rozpatrywane przez redaktora naczelnego.
7. Członkowie RN i KR ani pracownicy WCN nie mogą pozostawać w jakimkolwiek konflikcie interesów w odniesieniu do artykułów zgłaszanych do publikacji. Przez konflikt interesów należy rozumieć sytuację, w której jakiekolwiek interesy lub zależności (służbowe, finansowe lub inne) mogą mieć wpływ na ocenę artykułu lub decyzję o jego publikacji.
8. W celu przeciwdziałania nierzetelności naukowej wymagane jest złożenie przez autorów oświadczenia, w którym deklarują, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i jest ich oryginalnym dziełem, a także określają swój wkład w opracowanie artykułu.
9. W celu zapewnienia wysokiej jakości recenzji wymagane jest złożenie przez recenzentów oświadczenia o przestrzeganiu zasad etyki recenzowania COPE i niewystępowaniu konfliktu interesów.
10. W przypadku uzasadnionego podejrzenia na jakimkolwiek etapie procesu wydawniczego, że autorzy dopuścili się nierzetelności naukowej (zob. pkt 3.1. Odpowiedzialność autorów), zespół redakcyjny skrupulatnie zbada sprawę ewentualnego nadużycia. Jeśli nierzetelność autorów zostanie udowodniona, to zgłoszony przez nich artykuł zostanie odrzucony przez KR, a autorzy otrzymają informację o podjętej decyzji wraz z jej uzasadnieniem.
11. Czytelnicy, którzy mają wobec autorów opublikowanego artykułu uzasadnione podejrzenia o nierzetelność naukową, powinni powiadomić o tym redaktora naczelnego lub sekretarza redakcji. Po zbadaniu sprawy ewentualnego nadużycia czytelnicy zostaną poinformowani o rezultacie przeprowadzonego postępowania. W przypadku potwierdzenia nadużycia, na łamach czasopisma zostanie zamieszczona stosowna informacja.

3.3. Odpowiedzialność recenzentów

1. Recenzenci przyjmują artykuł do recenzji tylko wtedy, gdy uznają, że:
   - posiadają odpowiednią wiedzę w określonej dziedzinie, aby rzetelnie ocenić pracę;
   - zgodnie z ich stanem wiedzy nie istnieje konflikt interesów w odniesieniu do autorów, przedstawionych w artykule badań i instytucji je finansujących, co potwierdzają w oświadczeniu;
   - mogą wywiązać się z terminu ustalonego przez redakcję, aby nie opóźniać publikacji.
2. Recenzenci są zobligowani do zachowania obiektywności i poufności oraz powstrzymania się od osobistej krytyki. Zawsze powinni uzasadnić swoją ocenę, przedstawiając stosowną argumentację.

3. Recenzenci powinni wskazać ważne dla wyników badań opublikowane prace, które w ich ocenie powinny zostać przywołane w ocenianym artykule.
4. W razie stwierdzenia wysokiego poziomu zbieżności treści recenzowanej pracy z innymi opublikowanymi materiałami lub podejrzenia innych przejawów nierzetelności naukowej recenzenci są zobowiązani poinformować o tym redakcję.
5. Po ukończeniu recenzji przechowywanie przesłanych przez redakcję materiałów (w jakiejkolwiek formie) oraz posługiwanie się nimi przez recenzentów jest niedozwolone.

3.4. Odpowiedzialność wydawcy

1. Materiały opublikowane w „WS" są chronione prawem autorskim.
2. Wydawca udostępnia pełną treść wszystkich artykułów w internecie w trybie otwartego dostępu, tj. bezpłatnie i bez technicznych ograniczeń, od 1 stycznia 2022 r. na licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0). W przypadku artykułów zgłoszonych do „WS" od 2022 r. dozwolone jest dzielenie się artykułem (kopiowanie i rozpowszechnianie go w dowolnym medium i formacie) oraz adaptowanie go (w dowolnym celu, także komercyjnym) na warunkach określonych w tej licencji. Z pozostałych artykułów zamieszczonych w czasopiśmie można korzystać w ramach otwartego dostępu, zgodnie z ustawą o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego.
3. Wydawca deklaruje gotowość do opublikowania poprawek, wyjaśnień oraz przeprosin.

4. Wymogi redakcyjne

Zgodnie z wymogami czasopisma omawiany w artykule problem badawczy powinien być jednoznacznie zdefiniowany oraz istotny dla oceny zjawisk społecznych lub gospodarczych. Artykuł powinien zawierać wyraźnie określony cel badania, precyzyjny opis badanych zjawisk i stosowanych metod, uzyskane wyniki przeprowadzonej analizy oraz autorskie wnioski.

4.1. Struktura i zawartość artykułu

Wymagane elementy artykułu:
1. Tytuł.
2. Dane autora: imię/imiona i nazwisko, afiliacja w języku polskim i angielskim, ORCID, wkład w powstanie artykułu, adres e-mail. Wśród autorów artykułu wieloautorskiego należy wskazać autora korespondencyjnego.
3. Streszczenie (zalecana objętość – do 1200 znaków ze spacjami, forma bezosobowa). W przypadku artykułu opisującego badanie empiryczne powinno zawierać: cel, przedmiot, okres i metodę badania, źródła danych i najważniejsze wnioski z badania. W przypadku artykułów o innym charakterze należy podać co najmniej cel pracy, jej przedmiot i najważniejsze wnioski.

   Streszczenie to podstawowe źródło informacji o artykule, warunkujące też decyzję czytelnika o zapoznaniu się z całą pracą. Dlatego powinno być przygotowane ze szczególną starannością i dbałością o umieszczenie w nim wszystkich wymaganych elementów.

4. Słowa kluczowe – najistotniejsze pojęcia lub wyrażenia użyte w pracy (nie mniej niż trzy). Powinny być zawarte w streszczeniu i/lub tytule.
5. Kod/kody z klasyfikacji Journal of Economic Literature (JEL).
6. Tłumaczenie tytułu, streszczenia i słów kluczowych (na język angielski w przypadku artykułu napisanego w języku polskim, a na język polski w przypadku artykułu napisanego w języku angielskim).
7. W artykule opisującym badanie empiryczne wymagane są następujące części:
   - wprowadzenie, zawierające syntetyczne przedstawienie zagadnień teoretycznych, uzasadnienie podjęcia danego problemu badawczego, cel badania i krytyczne odniesienie do literatury przedmiotu. W wyjątkowych przypadkach, kiedy istotne dla podjętego tematu jest obszerniejsze przedstawienie dyskusji toczącej się w literaturze, przegląd literatury może stanowić odrębną część artykułu;
   - metoda badania, uwzględniająca przedmiot i okres badania, źródła danych i zastosowane metody badawcze, w tym uzasadnienie ich wyboru;
   - wyniki badania – analiza danych oraz interpretacja wyników i odniesienie ich do rezultatów wcześniejszych badań (dyskusja). W uzasadnionych przypadkach dyskusja może stanowić odrębną część artykułu;
   - podsumowanie, które powinno być zwięzłe i odzwierciedlać istotę problemu badawczego przedstawionego w artykule, bez podawania danych liczbowych; końcowe wnioski powinny odnosić się do treści artykułu, a w szczególności do celu badania.

   Wszystkie części powinny być opatrzone numerami.
8. Bibliografia, zawierająca pełny wykaz prac i materiałów przywołanych w artykule, przygotowana zgodnie z wymogami czasopisma.

4.2. Przygotowanie artykułu

1. Artykuł powinien być utrzymany w formie bezosobowej.
2. Tekst należy zapisać alfabetem łacińskim. Nazwy własne, tytuły itp. oryginalnie zapisane innym alfabetem powinny być poddane transliteracji.
3. Nie należy stosować stylów; formatowanie należy ograniczyć do wymogów redakcyjnych.
4. Objętość artykułu łącznie ze streszczeniem, słowami kluczowymi, bibliografią, tablicami, wykresami i innymi materiałami graficznymi nie powinna być mniejsza niż 10 stron maszynopisu ani przekraczać 20 stron.
5. Edytor tekstu: Microsoft Word, format *.doc lub *.docx.
6. Krój czcionki:
   - Arial – tytuł, autor, streszczenie, słowa kluczowe, kody JEL, śródtytuły, elementy graficzne (tablice, zestawienia, wykresy, schematy), przypisy;
   - Times New Roman – tekst główny, bibliografia.
7. Wielkość czcionki:
   - 14 pkt – tytuł, autor, śródtytuły wyższego rzędu;
   - 12 pkt – tekst główny, śródtytuły niższego rzędu;
   - 10 pkt – pozostałe elementy.
8. Marginesy – 2,5 cm z każdej strony.

9. Interlinia – 1,5 wiersza; tablice i przypisy – 1 wiersz; przed tytułami rozdziałów i podrozdziałów oraz po nich – pusty wiersz.
10. Wcięcie akapitowe – 0,4 cm; bibliografia – bez wcięcia, wysunięcie 0,4 cm.
11. Przy wyliczeniach należy posłużyć się listą punktowaną z punktorami w postaci kropek (wysunięcie 0,4 cm, wcięcie 0 cm); wiersze (oprócz ostatniego) zakończone średnikiem.
12. Strony ponumerowane automatycznie.
13. Tablice i elementy graficzne (wykresy, mapy, schematy) muszą być przywołane w tekście.
14. Wykresy, mapy i schematy należy zamieścić w tekście głównym. Wykresy powinny być edytowalne (optymalnie wykonane w programie Excel; w przypadku wykonania w programie graficznym powinny mieć postać wektorową). Wykresy i inne materiały graficzne należy przekazać osobno, najlepiej w pliku programu Excel lub innym edytowalnym w pakiecie Microsoft Office.
15. Tablice muszą być edytowalne. Nie należy stosować rastrów, cieniowania, pogrubiania czy też podwójnych linii itp.
16. Wskazówki dotyczące opracowywania map znajdują się w publikacji *Mapy statystyczne. Opracowanie i prezentacja danych*, dostępnej na stronie internetowej GUS.
17. Pod tablicami i każdym elementem graficznym należy podać źródło opracowania, a także objaśnić użyte w nich skróty i symbole.
18. Literowe symbole liczb i innych wielkości niezłożonych należy zapisywać małą lub dużą literą i pismem pochyłym (np. a, A, $y(x)$, a_i); wektorów – pismem pochyłym i pogrubionym (np. $\boldsymbol{a}$, $\boldsymbol{A}$, $\boldsymbol{w}$, $\boldsymbol{y}(x)$, $\boldsymbol{w}_i$); macierzy – pismem prostym i pogrubionym (np. $\mathbf{A}$, $\mathbf{a}$, $\mathbf{M}$, $\mathbf{Y}(x)$, $\mathbf{M}_i$).
19. Objaśnienia znaków umownych i zapisów w tablicach: kreska (–) – zjawisko nie wystąpiło; zero (0) – zjawisko istniało w wielkości mniejszej od 0,5; (0,0) – zjawisko istniało w wielkości mniejszej od 0,05; kropka (.) – brak informacji, konieczność zachowania tajemnicy statystycznej, wypełnienie pozycji jest niemożliwe lub niecelowe; „w tym" – oznacza, że nie podaje się wszystkich składników sumy.
20. Stosowane są następujące skróty: tablica – tabl., wykres – wykr.
21. Wszystkie zawarte w artykule informacje, dane i stwierdzenia wykraczające poza wiedzę powszechną – np. wyniki badań innych autorów, zarówno o charakterze empirycznym, jak i koncepcyjnym – muszą być opatrzone przypisem bibliograficznym. Przez wiedzę powszechną należy rozumieć informacje ogólnie znane i niebudzące wątpliwości ani kontrowersji w danej grupie społecznej, np. utworzenie GUS w 1918 r. lub powstanie UE w 1993 r. na podstawie traktatu z Maastricht. Natomiast dane statystyczne udostępniane lub publikowane np. przez GUS lub Eurostat nie należą do takich informacji. Charakteru wiedzy powszechnej nie mają również stwierdzenia odnoszące się do idei, zjawisk i procesów społecznych, politycznych czy gospodarczych. Nawet pozornie zdroworozsądkowe idee zmieniają bowiem swój sens w zależności od kultury, języka lub dyscypliny naukowej, a także bywają w rozmaity sposób konceptualizowane, jak np. pojęcie poznania w naukach społecznych.

    Podanie źródła jest konieczne niezależnie od tego, czy informacje lub stwierdzenia są ujęte w ramy cytatu, czy przedstawione bez dosłownego przytoczenia, np. w formie parafrazy. Jeżeli stwierdzenie może budzić jakiekolwiek wątpliwości odbiorców, autor powinien wskazać stosowne źródło podawanej informacji.

22. Przypisy rzeczowe, słownikowe lub informacyjne należy umieszczać na dole strony. Przypisy bibliograficzne, zgodnie ze standardem APA (American Psychological Association), należy podawać w tekście głównym.
23. Bibliografię należy przygotować zgodnie ze standardem APA.

4.3. Zasady przywoływania publikacji w treści artykułu

Wyszczególnienie	Przykład przywołania	
	w odsyłaczu	w treści zdania
Autor indywidualny		
Jeden autor	(Iksiński, 2001)	Iksiński (2001)
Dwóch autorów	(Iksiński i Nowak, 1999)	Iksiński i Nowak (1999)
Trzech autorów lub więcej	(Jankiewicz i in., 2003)	Jankiewicz i in. (2003)
Autor instytucjonalny		
Nazwa funkcjonuje jako powszechnie znany skrótowiec: pierwsze przywołanie w tekście	(International Labour Organization [ILO], 2020)	International Labour Organization (ILO, 2020)
kolejne przywołanie	(ILO, 2020)	ILO (2020)
Pełna nazwa	(Stanford University, 1995)	Stanford University (1995)
Typ publikacji		
Publikacja bez ustalonego autorstwa	(*Skrócony tytuł* …, 2015)	*Pełny tytuł* (2015)
Publikacja bez roku wydania	(Iksiński, b.r.)	Iksiński (b.r.)
Akt prawny	(Pełny tytuł)	Pełny tytuł
Strona internetowa / Zbiór danych: znana data publikacji	(Iksiński, 2020) / (Nazwa instytucji, 2020)	Iksiński (2020) / Nazwa instytucji (2020)
nieznana data publikacji	(Iksiński, b.r.) / (Nazwa instytucji, b.r.)	Iksiński (b.r.) / Nazwa instytucji (b.r.)
Rodzaj przywołania		
Przywoływanie kilku prac (porządek prac – chronologiczny, porządek autorów – alfabetyczny)	(Iksiński, 1997, 1999, 2004a, 2004b; Nowak, 2002)	Iksiński (1997, 1999, 2004a, 2004b) i Nowak (2002)
Przywoływanie publikacji za innym autorem (uwaga: w bibliografii należy wymienić tylko pracę czytaną)	(Nowakowski, 1990, za: Zieniecka, 2007)	Nowakowski (1990, za: Zieniecka, 2007)

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

4.4. Przykłady opisu bibliograficznego

Bibliografia powinna być zamieszczona na końcu opracowania. Prace należy uszeregować alfabetycznie według nazwiska pierwszego autora. W przypadku dwóch lub więcej prac tego samego autora / tych samych autorów trzeba je uporządkować chronologicznie według roku publikacji. Jeśli kilka prac tego samego autora / tych samych autorów zostało opublikowanych w tym samym roku, należy podać je w kolejności alfabetycznej według tytułu i odpowiednio oznaczyć literami a, b, c itd.

Typ publikacji	Przykład opisu bibliograficznego
Artykuł w czasopiśmie	
W wersji drukowanej	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik* (zeszyt), strona początku–strona końca.
Dostępny w internecie, z DOI	Nazwisko, X., Nazwisko 2, Y. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://doi.org/xxx.
Dostępny w internecie, bez DOI	Nazwisko, X., Nazwisko 2, Y., Nazwisko 3, Z. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://xxx.
Maszynopis	
Niepublikowany / przygotowywany przez autora / zgłoszony do publikacji, ale jeszcze niezaakceptowany	Nazwisko, X. (rok). *Tytuł* [maszynopis niepublikowany / w przygotowaniu / zgłoszony do publikacji].
Zaakceptowany do publikacji	Nazwisko, X. (w druku). Tytuł artykułu. *Tytuł czasopisma.*
Opublikowany nieformalnie (np. na stronie internetowej autora)	Nazwisko, X., Nazwisko 2, Y. (rok). *Tytuł artykułu*. https://xxx.
Opublikowany w trybie online first (przed włączeniem do zeszytu)	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma.* Online first. https://xxx.
Książka	
W wersji drukowanej	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo.
Dostępna w internecie, z DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://doi.org/xxx.
Dostępna w internecie, bez DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://xxx.
W przekładzie	Nazwisko, X. (rok). *Tytuł książki* (tłum. Y. Nazwisko). Wydawnictwo.
Wydanie wielotomowe: tom zatytułowany	Nazwisko, X. (rok). *Tytuł książki: nr tomu. Tytuł tomu*. Wydawnictwo.
tom niezatytułowany	Nazwisko, X. (rok). *Tytuł książki* (nr tomu). Wydawnictwo.
Kolejne wydanie	Nazwisko, X. (rok). *Tytuł książki* (nr wydania). Wydawnictwo.
Pod redakcją: w języku polskim	Nazwisko, X. (red.). (rok). *Tytuł książki*. Wydawnictwo.
w języku angielskim	Nazwisko, X. (Ed.). (rok). *Tytuł książki*. Wydawnictwo.
Rozdział w pracy zbiorowej	Nazwisko, X. (rok). Tytuł rozdziału. W: Y. Nazwisko, Z. Nazwisko 2 (red.), *Tytuł książki* (s. strona początku–strona końca). Wydawnictwo. https://doi.org/xxx lub https://xxx.
Inne prace	
Raport: autor indywidualny	Nazwisko, X. (rok). *Tytuł raportu*. Wydawnictwo.
autor instytucjonalny	Nazwa instytucji. (rok). *Tytuł raportu*. Wydawnictwo.
Working Papers	Nazwisko, X. (rok). *Tytuł pracy* (nazwa serii i numer). https://doi.org/xxx lub https://xxx.
Sesja konferencyjna / prezentacja / referat	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł pracy* [typ wystąpienia, np. referat]. Nazwa konferencji, miejsce konferencji.
Rozprawa doktorska: nieopublikowana	Nazwisko, X. (rok). *Tytuł pracy* [niepublikowana rozprawa doktorska]. Nazwa instytucji nadającej tytuł doktorski.
opublikowana	Nazwisko, X. (rok). *Tytuł pracy* [rozprawa doktorska, nazwa instytucji nadającej tytuł doktorski]. https://xxx.
Akt prawny	Pełny tytuł aktu prawnego wraz z datą publikacji w dzienniku urzędowym.

Typ publikacji	Przykład opisu bibliograficznego
Strona internetowa	
Znana data publikacji, zawartość strony się nie zmienia	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł*. https://xxx.
Nieznana data publikacji, zawartość strony się zmienia	Nazwa instytucji. (b.r.). *Tytuł*. Pobrane dzień, miesiąc i rok pobrania z https://xxx.
Zbiór danych	
Surowe dane nieopublikowane	Nazwisko, X. (rok wydania pracy, w której dane są wykorzystywane) [opis danych, np. surowe dane nieopublikowane dotyczące…]. Źródło danych (np. nazwa uniwersytetu).
Dane opublikowane: znana data publikacji, zawartość zbioru się nie zmienia	Nazwisko, X. (rok). *Nazwa zbioru danych* [zbiór danych]. Wydawca. https://xxx.
nieznana data publikacji, zawartość zbioru się zmienia	Nazwa instytucji. (b.r.). *Nazwa zbioru danych* [zbiór danych]. Wydawca. Pobrane dzień, miesiąc i rok pobrania z https://xxx.

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

Praca przygotowana w sposób niezgodny z powyższymi wskazówkami będzie odesłana do autora z prośbą o dostosowanie formy artykułu do wymogów redakcyjnych.

DZIAŁY „WS" – TEMATYKA ARTYKUŁÓW
WS SECTIONS – TOPICS OF THE ARTICLES

Pełny opis zakresu tematycznego działów: ws.stat.gov.pl/AimScope
Description of the topics covered in each section: ws.stat.gov.pl/AimScope

Studia metodologiczne / Methodological studies

- Oryginalne teoretyczne rozwiązania metodologiczne ze wskazaniem ich praktycznej użyteczności
- Prace przeglądowe i porównawcze oraz dotyczące etyki w statystyce, które wnoszą pionierski wkład poznawczy do obecnego stanu wiedzy

Statystyka w praktyce / Statistics in practice

- Nowatorskie zastosowania narzędzi i modeli statystycznych oraz analiza i ocena statystyczna zjawisk społeczno-ekonomicznych i innych, prowadzona w szczególności na danych z zasobów statystyki publicznej
- Wykorzystanie narzędzi informatycznych do uzyskiwania i przetwarzania informacji statystycznych, naliczania danych wynikowych, ich prezentacji i rozpowszechniania
- Projektowanie badań statystycznych, uzyskiwanie, integracja i przetwarzanie danych oraz generowanie wynikowych informacji statystycznych i kontrola ich ujawniania

Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges

- Wyzwania badawcze wynikające z rosnących potrzeb użytkowników danych statystycznych i wymagające zaangażowania znacznych środków oraz rozwiązań z różnych dziedzin nauki i techniki
- Wykorzystanie technologii informacyjnych i komunikacyjnych, innowacyjność, przetwarzanie i analiza zagadnień związanych z data science i big data
- Wyniki badań prowadzonych przez przedstawicieli dyscyplin innych niż statystyka z wykorzystaniem metod statystycznych

Spisy powszechne – problemy i wyzwania / Issues and challenges in census taking

- Propozycje rozwiązań – zarówno organizacyjnych, jak i metodologicznych – możliwych do zastosowania w spisach oraz rezultaty analiz danych spisowych
- Praktyczne aspekty związane z gromadzeniem i udostępnianiem danych ze spisów, w tym dotyczące obciążenia odpowiedzi i ochrony tajemnicy statystycznej

Edukacja statystyczna / Statistical education

- Metody i efekty nauczania statystyki oraz popularyzacja myślenia statystycznego i rzetelnego posługiwania się informacjami statystycznymi
- Problemy związane z kształceniem w zakresie umiejętności stosowania statystyki na wszystkich poziomach edukacji, a także dotyczące wykorzystywania nowoczesnych koncepcji i metod dydaktycznych oraz pomocy naukowych w nauczaniu statystyki

Z dziejów statystyki / From the history of statistics

- Historia prowadzenia obserwacji statystycznych oraz rozwoju ich metodologii i narzędzi
- Życie i osiągnięcia zawodowe wybitnych statystyków, jak również działalność najważniejszych instytucji i organizacji statystycznych w Polsce i za granicą

In memoriam

- Nekrologi i artykuły wspomnieniowe

Informacje. Recenzje. Dyskusje / Discussions. Reviews. Information

- Teksty nierecenzowane i niemające charakteru artykułów naukowych: sprawozdania z konferencji naukowych i innych wydarzeń dotyczących statystyki, recenzje książek, omówienia nowości wydawniczych GUS, polemiki i dyskusje