Cena 15,00 zł
(VAT 8%)

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

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

PAŹDZIERNIK / OCTOBER
ROCZNIK / VOLUME 68

2023 | 10

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

POLSKIE TOWARZYSTWO STATYSTYCZNE
POLISH STATISTICAL ASSOCIATION

GUS

TPS

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

PAŹDZIERNIK / OCTOBER
ROCZNIK / VOLUME 68

2023 | 10 (749)

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. 276/2023 – nakład 285 egz.

SPIS TREŚCI
CONTENTS

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), IV–VI
Od redakcji / From the editorial team

OD REDAKCJI

Październikowy numer „Wiadomości Statystycznych. The Polish Statistician" zawiera dwa artykuły dotyczące różnorodnych zastosowań metod statystycznych oraz pracę z zakresu organizacji i metodologii spisów powszechnych.

Artykuł dr. hab. Dominika Krężołka, prof. UE w Katowicach, pt. *Zastosowanie jednowskaźnikowego semiparametrycznego modelu ekonometrycznego w analizie ryzyka inwestycyjnego* jest poświęcony ocenie możliwości zastosowania tytułowego modelu do pomiaru ryzyka inwestycyjnego na Giełdzie Papierów Wartościowych w Warszawie (GPW). Model semiparametryczny charakteryzuje się tym, że niektóre jego założenia są mniej restrykcyjne niż założenia modeli parametrycznych dla funkcji średniej warunkowej, a jednocześnie zachowanych jest wiele pożądanych cech modelu liniowego i metody najmniejszych kwadratów. Autor wykorzystuje jednowskaźnikowy semiparametryczny model ekonometryczny do pomiaru ryzyka dla 10 wybranych spółek sektora informatycznego notowanych na GPW w latach 2018–2021. Jako czynniki ryzyka przyjmuje stopy zwrotu dwóch indeksów giełdowych: WIG20 (Polska) i S&P 500 (Stany Zjednoczone). Uzyskane wyniki wskazują, że polski rynek znacznie silniej niż amerykański determinuje zmienność stóp zwrotu badanych podmiotów oraz że modele semiparametryczne są bardziej elastyczne niż parametryczne w odniesieniu do założeń teoretycznych. Ponadto – mimo pewnych wad, takich jak czasochłonność obliczeń i wrażliwość na wybór metody estymacji lub parametryzację – modele te pozwalają na uwzględnienie zmienności ryzyka systematycznego w czasie, co ma znaczenie w dynamicznie zmieniającym się środowisku giełdowym. W warunkach napływu ogromnej ilości informacji omówione narzędzia mogą ułatwiać podejmowanie właściwych decyzji inwestycyjnych.

W pracy *Mass appraisal: a statistical approach to determining the impact of a property's attributes on its value* dr Anna Gdakowicz i dr Ewa Putek-Szeląg prezentują modyfikację Szczecińskiego Algorytmu Masowej Wyceny Nieruchomości (SAMWN), polegającą na obiektywnym obliczaniu wpływu cech nieruchomości na jej wartość z wykorzystaniem współczynników zależności. Autorki stawiają sobie za cel wskazanie tych współczynników zależności, które mogą być użyte do wyceny nieruchomości opartej na SAMWN i analizy nieruchomości wykorzystującej atrybuty mierzone na skali porządkowej, a także wyłonienie takiej kombinacji metod uwzględniania wpływu i wag atrybutów, której zastosowanie w procedurze SAMWN dałoby wyniki najbliższe wartościom nieruchomości oszacowanym przez rzeczoznawców majątkowych. Na podstawie analiz empirycznych obejmujących 405 nieruchomości gruntowych w Szczecinie przeznaczonych na cele mieszkaniowe, które na potrzeby badania zostały indywidualnie wycenione przez rzeczoznawców, autorki proponują cztery rodzaje współczynników zależności i ich wartości cząstkowych oraz cztery sposoby uwzględniania tych współczynników w procedurze SAMWN. Ponadto przeprowadzają ocenę zastosowania sześciu metod ważenia proponowanych miar. Optymalną kombinację rozpatrywanych wariantów algorytmu SAMWN uzyskują poprzez obliczenie miary błędów wyceny, a następnie przeprowadzenie procedury porządkowania liniowego. Z badania wynika, że na rezultaty wyceny największy wpływ ma formuła zastosowana do wyliczenia wag wpływu poszczególnych atrybutów na wartość nieruchomości w SAMWN. Potwierdza to przydatność analizowanego algorytmu w praktyce, ale jednocześnie konieczność dostosowywania jego procedury do specyfiki wycenianej nieruchomości.

Dr Miladin Kovačević, Mira Nikić, Branko Josipović, Snežana Lakčević, Vesna Pantelić, Nevena Mitrović, Adil Kolaković i Petar Korović w artykule *Digital population and housing census – the experience of Serbia* omawiają doświadczenia Serbii w zakresie organizacji Powszechnego Spisu Ludności, Gospodarstw Domowych i Mieszkań 2022, ze szczególnym uwzględnieniem zagadnień dotyczących zatrudnienia personelu, ram prawnych i finansowania tego badania. Uwagę skupiają na strategicznych decyzjach w sprawie zbierania danych i zastosowania technik informatycznych,

takich jak: wykorzystanie danych przestrzennych, cyfrowe metody uzyskiwania danych, uczenie maszynowe, łączenie rekordów czy system monitorujący, mających na celu sprostanie wyzwaniom związanym ze spisem. Autorzy poruszają także kwestie niedostatecznego pokrycia spisu oraz wykorzystania rejestrów administracyjnych do imputacji danych. Ponadto poświęcają uwagę opracowaniu i udoskonalaniu statystycznej ewidencji ludności, dokładności danych oraz obniżaniu kosztów i zwiększaniu efektywności badania. Podkreślają, że cyfrowe przeprowadzenie spisu zwiększyło jego dokładność, obniżyło koszty, a także usprawniło proces gromadzenia, przetwarzania i rozpowszechniania danych. Ponadto spis dostarczył bardziej szczegółowych informacji o charakterystyce demograficznej, takich jak cechy etniczno-kulturowe oraz skład gospodarstwa domowego i rodziny.

Ponadto w tym numerze prof. dr hab. Krzysztof Malaga, dr hab. inż. Jarogniew Rykowski, prof. UEP, dr hab. Marcin Szymkowiak, prof. UEP, i dr hab. Krzysztof Węcel, prof. UEP, przedstawiają sprawozdanie z tegorocznej XII Ogólnopolskiej Konferencji Naukowej im. Profesora Zbigniewa Czerwińskiego „Matematyka i informatyka na usługach ekonomii", poświęconej zastosowaniom metod matematycznych i statystycznych oraz informatyki w ekonomii.

Na koniec o najnowszych publikacjach GUS pisze Joanna Sadowy.

Życzymy miłej lektury.

FROM THE EDITORIAL TEAM

The October issue of *Wiadomości Statystyczne. The Polish Statistician* contains two articles relating to different applications of statistical methods and a paper on the organisation and methodology of a census.

In his article, entitled *Application of single-index semiparametric econometric model in investment risk analysis*, Dominik Krężołek, PhD, DSc, professor at the University of Economics in Katowice, assesses the possibility of using the model contained in the title to measure investment risk on the Warsaw Stock Exchange (WSE). The semiparametric model is characterised by the fact that some of its assumptions are less restrictive than those of parametric models for the conditional mean function; additionally, it retains many desirable features of a linear model and the least squares method. The author uses a single-index semiparametric econometric model to measure the risk for 10 selected companies from the IT sector listed on the WSE in the years 2018–2021. Rates of return of two stock market indices were adopted as the risk factors, namely WIG20 (Poland) and S&P 500 (USA). The obtained results demonstrate that the Polish market determines the volatility of the returns of the analysed companies to a much greater extent than the US market. Moreover, semiparametric models prove more flexible than parametric models with respect to theoretical assumptions. Despite having some disadvantages, such as time-consuming calculations and sensitivity to the choice of the estimation method or parametrisation, these models additionally allow the variability of systematic risk over time to be taken into account, which is important in the dynamically changing stock market environment. In the event of a large inflow of information, the considered tools may facilitate making the right investment decisions.

In the work *Mass appraisal: a statistical approach to determining the impact of a property's attributes on its value*, Anna Gdakowicz, PhD, and Ewa Putek-Szeląg, PhD, present a modification of the Szczecin Mass Real Estate Valuation Algorithm (Pol. *Szczeciński Algorytm Masowej Wyceny Nieruchomości* – SAMWN), involving the objective calculation of the influence of attributes on the value of a property using dependency coefficients. The authors aim to identify those dependency coefficients that can be used for an SAMWN-based real estate valuation and real estate analysis using attributes measured on an ordinal scale. Additionally, the paper aims to identify such a combination of methods for considering the influence and weights of attributes, whose application in the SAMWN procedure would produce results closest to those provided by real

estate appraisers. On the basis of empirical analyses covering 405 residential land properties in Szczecin, individually valued by appraisers for the purpose of the study, the authors propose four types of dependency coefficients and their partial values. They also propose four methods of considering these coefficients in the SAMWN procedure. Additionally, the authors assess the application of six weighting methods for the proposed measures. They obtain an optimal combination of the considered variants of the SAMWN algorithm by calculating the appraisal error measures, then followed by linear ordering procedures. The study shows that the valuation results are most affected by the formula used to calculate the weights of the influence of individual attributes on the value of the property in SAMWN. This confirms the usefulness of the analysed algorithm in practice, but at the same time the need to adapt its procedure to the specifics of the valued property.

Dr Miladin Kovačević, Mira Nikić, Branko Josipović, Snežana Lakčević, Vesna Pantelić, Nevena Mitrović, Adil Kolaković and Petar Korović, in their paper entitled *Digital population and housing census – the experience of Serbia*, discuss the experience of the Serbia in organising the 2022 Census of Population, Households and Dwellings, particularly focusing on the employment of staff, the legal framework and the financing of the survey. They discuss the strategic decisions relating to data collection and the use of information technology, such as geospatial data, digital data collecting methods, machine learning, record linkage and monitoring system, applied to meet the challenges of the census. The authors also address the issue of census undercoverage and the use of administrative records for data imputation. In addition, they devote attention to developing and improving the statistical population registers, data accuracy, reducing costs and increasing the efficiency of the survey. They underline that conducting the census digitally has increased its accuracy, reduced the costs and improved the process of collecting, processing and disseminating data. In addition, the census provided more detailed information on demographic characteristics, such as ethno-cultural characteristics and household and family composition.

In this issue, Krzysztof Malaga, PhD, DSc, ProfTit, Jarogniew Rykowski, BEng, PhD, DSc, professor at the Poznań University of Economics and Business, Marcin Szymkowiak, PhD, DSc, professor at the Poznań University of Economics and Business, and Krzysztof Węcel, PhD, DSc, professor at the Poznań University of Economics and Business, present their report on this year's, 12th Professor Zbigniew Czerwiński National Scientific Conference 'Mathematics and IT at the service of economics', relating to the application of mathematical and statistical methods, and computer science in economics.

The issue closes with a discussion on Statistics Poland's latest publications, prepared by Joanna Sadowy.

We wish you a pleasant reading.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 1–23 https://doi.org/10.59139/ws.2023.10.1
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 20.09.2022, poprawiony/revised: 5.07.2023, zaakceptowany/accepted: 26.08.2023

Zastosowanie jednowskaźnikowego semiparametrycznego modelu ekonometrycznego w analizie ryzyka inwestycyjnego[1]

Dominik Krężołek[a]

Streszczenie. Jednym z zadań ekonometrii stosowanej i statystyki jest estymacja funkcji średniej warunkowej w założonym modelu. Dostępne metody estymacji takiej funkcji i wyniki oszacowania zależą głównie od przyjętych *a priori* założeń dotyczących populacji lub procesu, który generuje dane. Celem badania omówionego w artykule jest ocena możliwości zastosowania jednowskaźnikowego semiparametrycznego modelu ekonometrycznego do pomiaru ryzyka inwestycyjnego na Giełdzie Papierów Wartościowych w Warszawie (GPW). Taki model charakteryzuje się tym, że niektóre jego założenia są mniej restrykcyjne niż założenia modeli parametrycznych dla funkcji średniej warunkowej (takich jak model liniowy i binarny model probitowy), a jednocześnie zachowuje wiele pożądanych cech modelu liniowego i metody najmniejszych kwadratów.

W przeprowadzonym badaniu semiparametryczny model jednowskaźnikowy został wykorzystany do pomiaru ryzyka dla 10 wybranych spółek sektora informatycznego notowanych na GPW od 2018 r. do 2021 r. Dane pobrano z serwisu finansowego Bloomberg. Jako czynniki ryzyka przyjęto stopy zwrotu dwóch indeksów giełdowych: WIG20 (Polska) i S&P 500 (Stany Zjednoczone). Uzyskane wyniki wskazują, że polski rynek znacznie silniej niż amerykański determinuje zmienność stóp zwrotu badanych spółek. Z badania wynika również, że modele semiparametryczne są bardziej elastyczne niż modele parametryczne w odniesieniu do założeń teoretycznych, co w warunkach napływu ogromnej ilości informacji może ułatwiać podejmowanie właściwych decyzji inwestycyjnych.

Słowa kluczowe: semiparametryczny model jednowskaźnikowy, analiza ryzyka, Giełda Papierów Wartościowych w Warszawie, GPW

JEL: C14, C50, C58

Application of single-index semiparametric econometric model in investment risk analysis

Abstract. One of the tasks of applied econometrics and statistics is the estimation of the conditional mean function of the assumed model. The available methods for estimating such a function and the estimation results depend mostly on the *a priori* assumptions about the

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na XV Międzynarodowej Konferencji Naukowej im. Profesora Aleksandra Zeliasia „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", która odbyła się w dniach 9–12 maja 2022 r. w Zakopanem. / The article is based on a paper delivered at the 15th Professor Alexander Zelias International Conference on 'Modeling and Forecasting of Socio-economic Phenomena', held on 9–12 May 2022 in Zakopane.

[a] Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Polska / University of Economics in Katowice, Faculty of Informatics and Communication, Poland.
ORCID: https://orcid.org/0000-0002-4333-9405. E-mail: dominik.krezolek@ue.katowice.pl.

population or process that generates the data. The aim of the research presented in this paper is to apply single-index semiparametric econometric model to measure investment risk on the Warsaw Stock Exchange (WSE). Such a model is characterised by some assumptions less restrictive than in the case of other parametric models for the conditional mean function, such as a linear model or a binary probit model. At the same time, a single-index model retains many of the desirable features of a linear model and a least squares method.

The presented model was used to measure investment risk for 10 IT companies quoted on the WSE in the period from 2018 to 2021. The data came from the Bloomberg financial service. Rates of return of two stock market indices, WIG20 (Poland) and S&P 500 (USA), were adopted as risk factors. The results indicate that the Polish market determines the volatility of returns of the analysed companies to a much larger extent than is the case with the US market. Furthermore, semiparametric models proved more flexible than the parametric ones regarding theoretical assumptions, which in the event of a large inflow of information might facilitate making correct investment decisions.

Keywords: semiparametric single-index model, risk analysis, Warsaw Stock Exchange

1. Wprowadzenie

Metody nieparametryczne to techniki statystyczne i ekonometryczne, które nie wymagają od badacza określenia postaci funkcji dla szacowanych modeli. Dane prezentowane w formie nieparametrycznej informują o kształtowaniu się jakiegoś zjawiska. Najprostszą taką formą jest histogram empiryczny (w metodach parametrycznych dopasowuje się do niego odpowiednią funkcję gęstości). W przypadku modeli regresji podejście nieparametryczne jest znane jako *regresja nieparametryczna* lub *wygładzanie nieparametryczne*. Wśród metod nieparametrycznych szczególną rolę odgrywa estymacja jądrowa, która pozwala empirycznie oszacować funkcję gęstości lub dystrybuantę analizowanej zmiennej losowej.

Metody nieparametryczne są coraz szerzej wykorzystywane w analizie danych użytkowych. Najlepiej sprawdzają się w pracy z dużymi zbiorami danych, co pierwotnie spotykało się z krytyką, głównie ze względu na złożoność obliczeniową, niedokładność obliczeń czy brak statystycznej istotności wyników (Wasserman, 2006). Z tych powodów z metod nieparametrycznych często korzysta się, gdy powszechnie stosowane specyfikacje parametryczne okazują się nieodpowiednie w przypadku badanego problemu. Dzieje się tak zwłaszcza wtedy, gdy formalne odrzucenie modelu parametrycznego na podstawie testów specyfikacji nie daje żadnych wskazówek co do kierunku poszukiwań ulepszonego modelu parametrycznego. Popularność metod nieparametrycznych wynika z tego, że nie wymagają spełnienia niektórych sztywnych założeń nałożonych na proces generowania danych, co jest konieczne w przypadku modeli parametrycznych. Innymi słowy, metody nieparametryczne pozwalają, aby to dane same określiły odpowiedni model.

W ciągu ostatnich kilkudziesięciu lat metody nieparametryczne, jak również semiparametryczne stały się przedmiotem dużego zainteresowania statystyków. Model semiparametryczny stanowi połączenie modeli parametrycznego i niepara-

metrycznego – jedna jego część jest określona przez znane parametry (jak w modelach parametrycznych), a druga – elastyczna i nieokreślona przez stałą liczbę parametrów (jak w modelach nieparametrycznych).

W latach 50. XX w. opublikowano pierwszą pracę na temat estymacji jądrowej (Rosenblatt, 1956). Rozwiązanie to zaproponowano w raporcie technicznym Sił Powietrznych Stanów Zjednoczonych jako sposób na uwolnienie analizy dyskryminacyjnej od sztywnych specyfikacji parametrycznych (Fix i Hodges, 1951). Na przestrzeni lat nastąpił gwałtowny rozwój omawianej dziedziny, zarówno w wymiarze teoretycznym, jak i praktycznym.

Celem badania omawianego w artykule jest ocena możliwości zastosowania jednowskaźnikowego semiparametrycznego modelu ekonometrycznego do pomiaru ryzyka inwestycyjnego na Giełdzie Papierów Wartościowych w Warszawie (GPW). Analizę przeprowadzono dla wybranych spółek sektora informatycznego notowanych na GWP w latach 2018–2021. Spółki te są składowymi indeksu WIG-informatyka, a jako czynniki ryzyka przyjęto stopy zwrotu dwóch indeksów giełdowych: polskiego WIG20 i amerykańskiego S&P 500. Badanie przeprowadzono na podstawie dziennych logarytmicznych stóp zwrotu analizowanych walorów. Postawiono hipotezę, że semiparametryczny model jednowskaźnikowy umożliwia szacowanie ryzyka determinowanego zmiennością rynku przy relatywnie niskich wartościach średniego błędu absolutnego (ang. *mean absolute error* – MAE). Ponadto przyjęto, że spółki sektora informatycznego w Polsce w większym stopniu zależą od zmienności obserwowanej na giełdzie polskiej niż amerykańskiej.

2. Semiparametryczny model jednowskaźnikowy

Metody semiparametryczne to jedne z najpopularniejszych metod bazujących na estymacji elastycznej, odnoszącej się do wyznaczenia funkcji, która – inaczej niż w tradycyjnych modelach parametrycznych – nie jest ściśle określona przez stałą liczbę parametrów. Funkcja jest elastyczna i można ją dostosowywać do skomplikowanych wzorców w zbiorze danych bez konieczności określania *a priori* jej formy. Dzięki temu w modelu można uwzględnić nieliniowe zależności w zbiorze danych, co mogłoby się okazać trudne w przypadku zastosowania typowych modeli parametrycznych (Silverman, 1986). Modele semiparametryczne, powstałe dzięki połączeniu modeli parametrycznych i nieparametrycznych, są przydatne w sytuacjach, w których narzędzia w pełni nieparametryczne nie dają właściwych oszacowań. Do takiej sytuacji dochodzi, gdy nadwymiarowość analizowanego zbioru danych prowadzi do bardzo zróżnicowanych oszacowań lub gdy nie jest znana postać funkcyjna określonej zależności w odniesieniu do podzbioru regresorów lub rozkładu składnika losowego. Wspomniana nadmiarowość dotyczy zbiorów danych charakteryzujących się dużą liczbą cech i zmiennych. Może się też zdarzyć, że niektóre regresory

występują jako funkcja liniowa względem zmiennych, ale postać funkcyjna parametrów w odniesieniu do innych zmiennych nie jest znana, lub że funkcja regresji jest nieparametryczna, ale struktura składnika losowego ma postać parametryczną (Li i Racine, 2007).

Modele semiparametryczne są postrzegane jako rodzaj kompromisowego rozwiązania między specyfikacjami w pełni nieparametrycznymi i w pełni parametrycznymi. Opierają się na założeniach parametrycznych i dlatego mogą być błędnie określane, tak jak ich parametryczne odpowiedniki. Znanych jest wiele metod semiparametrycznych. W dalszej części artykułu ograniczono się do opisu modelu typu regresyjnego, a dokładnej – modelu jednowskaźnikowego.

Dokonując analizy dużych zbiorów danych, badacze niejednokrotnie spotykają się ze zjawiskiem nadwymiarowości w zbiorze danych. Jednym ze sposobów jej redukcji jest zastosowanie modelu jednowskaźnikowego. W modelu tego typu występuje jednowymiarowy wskaźnik (skalar), który stanowi element nieznanej funkcji nieparametrycznej. Modele jednowskaźnikowe zawdzięczają swoją popularność temu, że działają podobnie jak wiele znanych modeli parametrycznych, np. regresja Poissona, modele logitowe, probitowe i tobitowe (Härdle i Stoker, 1989).

Model jednowskaźnikowy to jeden z najpopularniejszych modeli semiparametrycznych stosowanych w naukach ilościowych. Estymacja wektora $\boldsymbol{\beta}$ współczynników modelu była analizowana przez wielu badaczy, którzy rozważali kwestię efektywności. Wśród wykorzystywanych do tego celu technik należy wymienić metody: średniej pochodnej (Powell i in., 1989; Stoker, 1986), odwróconej regresji krokowej (Duan i Li, 1991; Li, 1991), najmniejszych kwadratów (Härdle i in., 1993; Li i in., 2014; Yang i Song, 2014), minimalnej uśrednionej warunkowej wariancji (Xia, 2006; Xia i in., 2002) i empirycznego prawdopodobieństwa (Xue, 2013; Xue i Zhu, 2006; Zhu i Xue, 2006). Model jednowskaźnikowy należy uznać za atrakcyjne i efektywne podejście do nieparametrycznego wielowymiarowego modelowania, ponieważ redukuje on problem nadwymiarowości zbioru danych przy jednoczesnym zachowaniu interpretowalności modeli liniowych.

Semiparametryczne modele jednowskaźnikowe mają przewagę nad klasycznymi modelami liniowymi w kilku kluczowych aspektach (Pagan i Ullah, 1999). Pierwszy z nich to elastyczność dotycząca kształtu funkcji. W modelu liniowym zakłada się zachodzenie liniowej relacji między zmiennymi objaśniającymi a zmienną zależną, co może prowadzić do niedoszacowania lub przeszacowania efektów zmiennych objaśniających, jeśli rzeczywista relacja jest nieliniowa. W jednowskaźnikowym semiparametrycznym modelu ekonometrycznym można uwzględnić nieliniowe zależności między zmiennymi, co pozwala na dokładniejsze oszacowanie efektów zmiennych objaśniających. Istotną cechą tego modelu jest także brak założeń dotyczących rozkładu składnika losowego. W modelu liniowym wymaga się, aby błędy

miały rozkład normalny, co w niektórych przypadkach może być niewłaściwe. W jednowskaźnikowym semiparametrycznym modelu ekonometrycznym taki warunek nie jest wymagany, co czyni ten model bardziej odpornym na nieodpowiednie założenia. Inną właściwością podejścia semiparametrycznego jest brak założeń dotyczących homoskedastyczności składnika losowego. W modelu liniowym zakłada się, że błędy mają stałą wariancję. Ten warunek może okazać się nieodpowiedni w przypadku, gdy wariancja błędów zależy od wartości zmiennych objaśniających (Racine, 2008). Jednowskaźnikowy semiparametryczny model ekonometryczny nie wymaga takiego założenia i może dostarczyć bardziej wiarygodnych oszacowań, gdy homoskedastyczność wariancji składnika losowego nie występuje. Ponadto istnieje możliwość uwzględnienia efektów nieliniowych zmiennych w wariancji błędów, a także nieliniowych zależności między zmiennymi objaśniającymi a błędami, co może się przyczyniać do zwiększania dokładności oszacowań. Za kolejną zaletę modelu jednowskaźnikowego należy uznać to, że oszacowania składnika parametrycznego są asymptotycznie zbieżne z rzeczywistymi w tempie podobnym do modelu w pełni parametrycznego. Te korzyści wiążą się jednak z pewnymi kosztami, np. trudnościami z włączeniem do wskaźnika (skalara) zmiennych dyskretnych lub w ogóle brakiem takiej możliwości czy problemami z określeniem właściwej struktury modelu, głównie w kontekście zastosowanej funkcji łączącej i funkcji składowych takiego modelu (Henderson i Parmeter, 2015).

Ogólna postać modelu jednowskaźnikowego jest następująca (Horowitz, 2009):

$$y_i = m\big(\varphi(\boldsymbol{x}_i, \boldsymbol{\beta})\big) + u_i, \tag{1}$$

gdzie:

$m(\cdot)$ – nieznana funkcja wygładzająca,

$\varphi(\cdot,\cdot)$ – znana funkcja parametryczna q zmiennych objaśniających $\boldsymbol{x}_i$ i p-wymiarowego wektora $\boldsymbol{\beta}$ współczynników,

u_i – składnik losowy nieskorelowany z $\boldsymbol{x}_i$,

$i = 1, 2, \dots, n$.

Nawet jeśli $\varphi(\boldsymbol{x}_i, \boldsymbol{\beta})$ jest funkcją nieliniową, to uważa się ją za pojedynczy wskaźnik, ponieważ $\varphi(\boldsymbol{x}_i, \boldsymbol{\beta})$ jest skalarem.

W większości badań ekonomicznych przyjmuje się, że mamy do czynienia z pojedynczym liniowym wskaźnikiem (skalarem), a zatem funkcja $\varphi(\cdot,\cdot)$ przyjmuje postać:

$$\varphi(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi}, \tag{2}$$

gdzie liczba zmiennych niezależnych jest równa liczbie szacowanych parametrów, tj. $p = q$.

Postać macierzowa modelu jednowskaźnikowego ma więc formę (Ramanathan, 1989):

$$y_i = m(\boldsymbol{x}_i\boldsymbol{\beta}) + u_i, \tag{3}$$

gdzie:
$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})$,
$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$,
$i = 1, 2, \ldots, n$.

Warto zauważyć, że w modelu (2) nie występuje wyraz wolny, ponieważ nie można zidentyfikować punktu przecięcia bez ograniczeń nałożonych na nieznaną funkcję, a wektor parametrów $\boldsymbol{\beta}$ zazwyczaj nie może być zidentyfikowany bez normalizacji (podobnie jak w przypadku modeli logitowych i probitowych). W literaturze najczęściej występują dwie normalizacje. Pierwsza z nich polega na ustawieniu współczynnika pierwszego regresora na poziomie 1 (przy założeniu, że prawdziwy parametr jest różny od 0), a w drugiej zakłada się, że wektor $\boldsymbol{\beta}$ ma długość jednostkową (tj. jego norma euklidesowa wynosi 1). Więcej szczegółowych informacji na temat normalizacji parametrów w modelach jednowskaźnikowych można znaleźć w pracy Camerona i Trivediego (2005).

Podstawowa różnica pomiędzy modelem parametrycznym a semiparametrycznym polega na wykorzystaniu funkcji wygładzającej $m(\cdot)$. W modelu parametrycznym ta funkcja jest znana i możliwe jest zastosowanie estymacji nieliniowej metodą najmniejszych kwadratów lub estymacji metodą największej wiarygodności. W przypadku modelu semiparametrycznego funkcja $m(\cdot)$ jest nieznana i należy ją oszacować. Warto zauważyć, że w sensie ogólnym $m(\cdot)$ jest średnią warunkową y względem $\boldsymbol{x\beta}$. Wynika z tego wybór przeprowadzenia nieparametrycznej regresji y względem $\boldsymbol{x\beta}$. Gdyby wektor $\boldsymbol{\beta}$ był znany, taka regresja byłaby bardzo prosta, jednak w praktyce jest on nieznany, dlatego należy oszacować zarówno $m(\cdot)$, jak i $\boldsymbol{\beta}$.

2.1. Estymacja parametrów semiparametrycznego modelu jednowskaźnikowego

Przy założeniu, że funkcja wygładzająca $m(\cdot)$ jest znana, oszacowanie wektora $\boldsymbol{\beta}$ można uzyskać za pomocą nieliniowej metody najmniejszych kwadratów. W takim przypadku należy minimalizować funkcję celu postaci:

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - m(\boldsymbol{x}_i\boldsymbol{\beta})]^2 \tag{4}$$

względem wektora $\boldsymbol{\beta}$ (Ichimura, 1993). Jednak w sytuacji, kiedy funkcja $m(\cdot)$ nie jest znana, należy oszacować zarówno $m(\cdot)$, jak i $\boldsymbol{\beta}$. W pracy Ichimury (1993) zaproponowano wykorzystanie w tym celu estymatora typu *leave-one-out* średniej warunkowej zmiennej $\boldsymbol{y}$ względem $\boldsymbol{x\beta}$ przy jednoczesnej minimalizacji sumy kwadratów reszt względem q-wymiarowego wektora $\boldsymbol{\beta}$. W statystyce estymacja typu *leave-one-out* odnosi się do specyficznego schematu estymacji, w którym pojedyncza obserwacja jest pomijana przy estymacji parametru. Następnie tej pominiętej obserwacji używa się do obliczenia błędu estymacji.

Estymator typu *leave-one-out* średniej warunkowej $\boldsymbol{y}$ względem $\boldsymbol{x\beta}$ ma następującą postać:

$$E_{-i}(y_i|\boldsymbol{x}_i\boldsymbol{\beta}) = \frac{\sum_{j\neq i}^{n} y_j K_h(\boldsymbol{x}_j\boldsymbol{\beta}, \boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j\neq i}^{n} K_h(\boldsymbol{x}_j\boldsymbol{\beta}, \boldsymbol{x}_i\boldsymbol{\beta})}, \tag{5}$$

gdzie:

$K_h(\boldsymbol{x}_j\boldsymbol{\beta}, \boldsymbol{x}_i\boldsymbol{\beta}) = \prod_{d=1}^{q} k\left(\frac{x_{dj}\boldsymbol{\beta}_d - x_{di}\boldsymbol{\beta}_d}{h_d}\right)$ – iloczyn funkcji jądrowej $k(\cdot)$[2],

h_d – szerokość pasma (parametr wygładzania).

Problem optymalizacyjny dany wzorem (4) przyjmuje ostatecznie postać:

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{\sum_{j\neq i}^{n} y_j K_h(\boldsymbol{x}_j\boldsymbol{\beta}, \boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j\neq i}^{n} K_h(\boldsymbol{x}_j\boldsymbol{\beta}, \boldsymbol{x}_i\boldsymbol{\beta})}\right]^2. \tag{6}$$

Estymator wektora $\boldsymbol{\beta}$ jest definiowany jako wartości wektora oszacowań $\widehat{\boldsymbol{\beta}}$, które minimalizują funkcję celu określoną wzorem (6).

W badaniach teoretycznych nad modelem jednowskaźnikowym funkcja celu określona wzorem (6) różni się na ogół w dwóch aspektach od funkcji danej wzorem (4). Po pierwsze, biorąc pod uwagę, że lokalnie stały estymator najmniejszych kwadratów jest zbieżny jednostajnie po zbiorach zwartych, niezbędna jest pewna funkcja ucinająca, aby wyeliminować małe wartości w mianowniku. Po drugie, funkcja celu szacowana jest często za pomocą funkcji wagowej. Ichimura (1993) sugeruje wykorzystanie jako funkcji wagowej odwrotności wariancji niejednorodnej. Funkcja wagowa nie jest niezbędna do osiągnięcia tempa zbieżności rzędu $\sqrt{n}$ dla współczynników kierunkowych, gdy składnik losowy jest heteroskedastyczny, ale jest wymagana do osiągnięcia granicy efektywności dla oszacowań semiparametrycznych.

[2] Dla ciągu $X_1, X_2, \ldots, X_n$ realizacji zmiennej losowej X funkcja $k\left(\frac{x-X_i}{h}\right)$ jest funkcją jądrową o szerokości pasma h.

2.2. Wybór optymalnej szerokości pasma semiparametrycznego modelu jednowskaźnikowego

Funkcja jądrowa $k(\cdot)$ występująca we wzorze (5) jest funkcją nieujemną i powinna spełniać następujące warunki:

- $\int_{-\infty}^{+\infty} k(\psi)d\psi = 1$;
- $\int_{-\infty}^{+\infty} \psi\, k(\psi)d\psi = 0$;
- $\int_{-\infty}^{+\infty} \psi^2\, k(\psi)d\psi = \kappa_2 < \infty$, gdzie κ_2 to moment centralny rzędu drugiego.

Dodatkowo funkcja $k(\psi)$ powinna być symetryczna i przyjmować duże (małe) wartości dla małych (dużych) ψ. Wśród najpopularniejszych funkcji jądra wyróżnia się funkcje (Henderson i Parmeter, 2015):

- gaussowską: $k(\psi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\psi^2}{2}\right)$;
- Epanecznikowa: $k(\psi) = \frac{3}{4\sqrt{5}}\left(1 - \frac{\psi^2}{5}\right)$ dla $|\psi| < \sqrt{5}$;
- trójkątną: $k(\psi) = 1 - |\psi|$ dla $|\psi| < 1$;
- jednostajną: $k(\psi) = \frac{1}{2}$ dla $|\psi| < 1$;
- dwuwagową: $k(\psi) = \frac{15}{16}(1 - \psi^2)^2$ dla $|\psi| < 1$;
- trójwagową: $k(\psi) = \frac{35}{32}(1 - \psi^2)^3$ dla $|\psi| < 1$.

Szerokość pasma h_d występująca we wzorze (5) określa stopień wygładzenia funkcji $m(\cdot)$. Optymalna szerokość pasma h to taka, która minimalizuje określoną miarę dokładności estymatora, np. błąd średniokwadratowy (ang. *mean squared error* – MSE) lub scałkowany błąd średniokwadratowy (ang. *integrated mean squared error* – IMSE). W semiparametrycznym modelu jednowskaźnikowym wybór szerokości pasma jest dość skomplikowany. Nie jest oczywiste, że optymalna szerokość pasma najlepsza do oszacowania średniej warunkowej jest jednocześnie najlepsza do estymacji parametrów modeli parametrycznych. Nie jest również ściśle określone, czy szerokość pasma należy szacować w jednym kroku, czy w większej liczbie kroków. W pracy Härdle'a i in. (1993) zaproponowano procedurę wyboru optymalnej wartości h. Wykazano przy tym, że równoczesna estymacja szerokości pasma h i wektora $\boldsymbol{\beta}$ jest bardzo zbliżona do minimalizacji sumy kwadratów reszt oraz zwykłej funkcji walidacji krzyżowej dla osobnej estymacji h i $\boldsymbol{\beta}$. W rzeczywistości prowadzi to do uzyskania $\sqrt{n}$-zgodnych oszacowań wektora $\boldsymbol{\beta}$ współczynników, jak również asymptotycznie optymalnych oszacowań szerokości pasma h.

Funkcja celu wykorzystana do wyznaczenia optymalnej szerokości pasma otrzymywana jest metodą walidacji krzyżowej i przyjmuje postać

$$CV(\widehat{\boldsymbol{\beta}},\, h) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - m_{-i}(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\right]^2, \tag{7}$$

gdzie $m_{-i}(\cdot)$ jest estymatorem typu *leave-one-out* funkcji wygładzającej $m(\cdot)$.

Funkcja celu $CV(\widehat{\boldsymbol{\beta}}, h)$ jest minimalizowana zarówno względem $\widehat{\boldsymbol{\beta}}$, jak i względem h. Udowodniono (Härdle i in., 1993), że funkcję daną wzorem (7) można przybliżyć jako

$$CV(\widehat{\boldsymbol{\beta}}, h) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - m_{-i}(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\right]^2 \approx \frac{1}{n}\sum_{i=1}^{n}\left[y_i - m(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\right]^2 + \\ + \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{m}_{-i}(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}) - m(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}})\right]^2. \tag{8}$$

Analizując wzór (8), można zauważyć, że pierwszy składnik jest problemem optymalizacyjnym najmniejszych kwadratów dla oszacowania wektora parametrów, gdy funkcja $m(\cdot)$ jest znana, a drugi składnik jest problemem optymalizacyjnym najmniejszych kwadratów dla oszacowania szerokości pasma nieznanej funkcji $m(\cdot)$, przy założeniu że wektor parametrów jest znany. Funkcja walidacji krzyżowej zachowuje się więc tak, jakby każdy składnik modelu estymowano osobno.

2.3. Ocena semiparametrycznego modelu jednowskaźnikowego

Zastosowanie semiparametrycznego modelu jednowskaźnikowego wymaga oceny stopnia, w jakim dopasowuje się on do danych. Wśród najważniejszych kryteriów takiej oceny można wymienić miary jakości, zgodności i odporności modelu.

Modele parametryczne mają ustalony zestaw parametrów opisujących ich strukturę, np. w modelu regresji liniowej parametrami są współczynniki nachylenia i przesunięcia. Modele semiparametryczne odznaczają się większą elastycznością i w większym stopniu mogą dostosowywać się do danych – niektóre części ich struktury są parametryczne, a inne nieparametryczne. Oznacza to, że pewne elementy modelu nie są ograniczone do określonych parametrów, co pozwala na lepszą adaptację do różnych form danych (Horowitz, 2009).

Jeśli chodzi o miary jakości, to różnica między rozpatrywanymi typami narzędzi może wynikać z elastyczności modeli semiparametrycznych. Modele parametryczne charakteryzują się bowiem zazwyczaj mniejszą liczbą parametrów i ściśle określoną strukturą, co może prowadzić do uproszczeń w zakresie interpretacji i oceny modelu. Jeśli jednak dane są bardziej skomplikowane lub nie spełniają założeń modelu parametrycznego, to modele semiparametryczne mogą się do nich lepiej dostosować (Yatchew, 2003).

Znanych jest kilka miar dobroci dopasowania, które mogą być wykorzystywane do oceny efektywności semiparametrycznego modelu jednowskaźnikowego. Wybrane z nich przedstawiono w zestawieniu (Henderson i Parmeter, 2015).

Zestawienie wybranych miar dobroci dopasowania

Miary dobroci dopasowania	Charakterystyka
Średni kwadrat błędu (MSE)	Miara średniej kwadratowej różnicy między wartościami przewidywanymi a rzeczywistymi; niskie wartości MSE świadczą o lepszym dopasowaniu
Kryterium informacyjne Akaikego (ang. *Akaike information criterion* – AIC)	Miara relacji między dopasowaniem modelu a jego złożonością; niższa wartość AIC wskazuje na lepszą równowagę między tymi cechami
Kryterium informacyjne Bayesa (ang. *Bayesian information criterion* – BIC)	Miara podobna do AIC, ale nakłada większą karę za złożoność modelu; niższa wartość BIC wskazuje na lepsze dopasowanie
R^2	Miara jakości dopasowania określająca proporcję zmienności zmiennej zależnej, która jest wyjaśniona przez zmienne niezależne; wyższy R^2 wskazuje na lepsze dopasowanie
Miara rozbieżności Kullbacka-Leiblera	Miara różnicy pomiędzy prawdziwym rozkładem danych a założonym rozkładem teoretycznym; niższe wartości miary wskazują na lepsze dopasowanie
Średni błąd bezwzględny (MAE)	Miara średniej różnicy między wartościami przewidywanymi a rzeczywistymi; niższe wartości MAE wskazują na lepsze dopasowanie modelu

Źródło: opracowanie własne na podstawie: Pagan i Ullah (1999).

Należy podkreślić, że zastosowanie jakiejkolwiek pojedynczej miary nie pozwala w pełni uchwycić dobroci dopasowania modelu, a zastosowanie różnych miar może prowadzić do odmiennych wniosków na temat wydajności modelu. Do oceny wydajności modelu zaleca się zatem stosowanie kilku miar.

Znane są również bardziej zaawansowane miary jakości, wykorzystywane do oceny dopasowania semiparametrycznych modeli ekonometrycznych do danych o bardziej skomplikowanej strukturze, czyli takich, które nie mogą być łatwo modelowane przy użyciu standardowych metod parametrycznych z powodu ich różnorodności, nieliniowości lub interakcji zmiennych. Należą do nich miary oparte na estymatorach jądrowych, metody bootstrapowe i różne testy nieparametryczne (Horowitz, 2009). Miary oparte na estymatorach jądrowych, takie jak jądrowa funkcja gęstości, wykorzystuje się do oceny dopasowania modelu do rozkładu danych, np. estymator jądrowy może być użyty do porównania empirycznego rozkładu danych z rozkładem przewidywanym przez model semiparametryczny. Metody bootstrapowe, czyli techniki próbkowania z powtórzeniami, pozwalają na estymację rozkładu próbkowego w przypadku analizowanych statystyk i są stosowane w semiparametrycznych modelach ekonometrycznych do oszacowania przedziałów ufności dla parametrów lub innych statystyk modelu. Testy nieparametryczne opierają się z kolei na rozkładzie empirycznym danych, który jest porównywany z rozkładem przewidywanym przez model. Za przykład mogą w tym wypadku posłużyć testy dopasowania rozkładu (np.

test Kołmogorowa-Smirnowa – K-S) lub testy porównujące grupy (np. test rangowy Wilcoxona-Manna-Whitneya). Wymienione miary jakości stosuje się w różnych kontekstach, w zależności od konkretnego semiparametrycznego modelu ekonometrycznego i charakterystyki danych. Wybór odpowiedniej miary jakości zależy od celu analizy i struktury danych.

Inną grupą miar, za pomocą których można dokonywać oceny modeli semiparametrycznych, są miary zgodności. Pomagają one określić, w jakim stopniu model odpowiada danym, czyli jak dalece pasuje do kształtowania się obserwacji. Miary zgodności stosowane w przypadku modeli semiparametrycznych mogą się różnić od tych wykorzystywanych w przypadku modeli parametrycznych. Wśród miar zgodności dla modeli semiparametrycznych można wyróżnić m.in. krzywą ROC (ang. *receiver operating characteristic*) i krzywą Kaplana-Meiera. Krzywa ROC znajduje zastosowanie m.in. w przypadku modeli klasyfikacyjnych, np. logistycznych. Im krzywa ROC jest położona wyżej i im większe jest pole powierzchni pod krzywą (AUC – *area under the curve*), tym większa zgodność modelu z danymi. Krzywa Kaplana-Meiera może być z kolei wykorzystana w analizie historii zdarzeń (analizie przeżycia). Przedstawia ona szanse przeżycia w zależności od czasu. Miara zgodności w postaci testu log-rank i różne metody porównawcze mogą być natomiast stosowane do porównywania różnych semiparametrycznych modeli przeżycia. Aby ocenić, który model semiparametryczny lepiej pasuje do danych, warto zastosować testy porównawcze, np. porównanie funkcji straty (ang. *loss function comparison*; Henderson i Parmeter, 2015).

Semiparametryczne modele ekonometryczne różnią się od modeli parametrycznych również pod względem miar odporności. Miary te służą do oceny stopnia stabilności modelu i jego niezależności od odstępstw od założeń lub skrajnych obserwacji w zbiorze danych. W przypadku modeli parametrycznych – o ściśle określonych założeniach dotyczących struktury modelu i parametrów – miary odporności mogą być wykorzystywane do identyfikacji obserwacji odstających, które mają duży wpływ na estymowane parametry lub dopasowanie modelu. Warto tutaj wymienić współczynnik wpływu Cooka, statystykę DFBETA (mierzącą zmianę w estymowanych parametrach modelu po usunięciu poszczególnych obserwacji) i błąd standardowy Hubera-White'a (odporna miara wariancji estymatora parametrów, która uwzględnia obserwacje odstające).

W przypadku semiparametrycznych modeli ekonometrycznych – o większej elastyczności w strukturze modelu i możliwości dopasowywania się do różnych form danych – miary odporności mogą być wykorzystywane w podobny sposób. W tym przypadku warto wykorzystać ocenę reszt lub odchyleń standardowych. Analiza reszt modelu może ujawnić obserwacje odstające lub niezgodności z założeniami modelu, a duże wartości odchylenia standardowego mogą wskazywać na obserwacje,

które mają znaczący wpływ na model. Inne miary bazują na estymacji kwantylowej, która pozwala na modelowanie rozkładu zmiennych zależnych w bardziej elastyczny sposób. Do analizy wpływu obserwacji ekstremalnych na model i estymowane parametry można stosować różne kwantyle, zależnie od problemu badawczego. W modelach semiparametrycznych przydatne są także techniki wykorzystujące modele odporne (takie jak odporna regresja lub odporna estymacja jądrowa), które minimalizują wpływ obserwacji odstających na estymację parametrów modelu. Ma to szczególne znaczenie w przypadku danych nietypowych lub obserwacji odstających, które mogą istotnie wpływać na wyniki estymacji modelu (Pagan i Ullah, 1999).

3. Pomiar ryzyka rynkowego z wykorzystaniem miar wrażliwości

Funkcjonowanie jednostki gospodarczej na rynku wiąże się z procesem podejmowania decyzji. Ze względu na posiadane informacje problemy decyzyjne można podzielić na trzy grupy (Jajuga, 2003):

- decyzje podejmowane w warunkach pewności, gdy każda decyzja pociąga za sobą określone, znane konsekwencje;
- decyzje podejmowane w warunkach ryzyka, gdy każda decyzja pociąga za sobą więcej niż jedną konsekwencję oraz znany jest zbiór możliwych konsekwencji i prawdopodobieństwo ich wystąpienia;
- decyzje podejmowane w warunkach niepewności, gdy prawdopodobieństwo wystąpienia konsekwencji decyzji nie jest znane.

W praktyce gospodarczej decyzje najczęściej podejmowane są w warunkach ryzyka i niepewności. Przez niepewność należy rozumieć obawę związaną z niewiedzą dotyczącą otaczającego świata i konsekwencjami tej niewiedzy. Może być ona wynikiem wielu czynników, np. niedostatecznej i nie w pełni transparentnej informacji, braku umiejętności interpretacji zjawisk gospodarczych – hipotetycznie związanych z rozważanym zakresem działalności rynkowej – lub indywidualnego podejścia każdego uczestnika rynku. Najczęstsza prawidłowość jest taka, że im bardziej złożone jest zjawisko rynkowe, tym większa niepewność i nieprzewidywalność dotycząca jego realizacji w przyszłości. Pojawia się bowiem ryzyko, że oczekiwania okażą się odmienne od nieznanej rzeczywistości (Krężołek, 2020).

Termin *ryzyko* najczęściej kojarzy się z sytuacją, w której realizacja podjętych działań przebiega na niższym poziomie, niż oczekiwano. Znane są dwa sposoby rozumienia tego terminu: neutralne i negatywne. W podejściu neutralnym przyjmuje się, że rzeczywista wielkość badanego zjawiska różni się od jej zakładanego poziomu (nie jest istotny znak tej różnicy), a w podejściu negatywnym rzeczywista wielkość badanego zjawiska jest mniejsza od jej oczekiwanego poziomu. Warto wskazać paradoks wynikający z neutralnego podejścia do rozumienia ryzyka – brak istotności

znaku różnicy poziomu badanego zjawiska zakłada, że w przypadku różnicy dodatniej (co z inwestycyjnego punktu widzenia oznacza zysk) uczestnik rynku nie postrzega takiej sytuacji jako ryzykownej. Stąd wniosek, że ryzyko jest zazwyczaj postrzegane negatywnie. W literaturze podano wiele definicji i kryteriów klasyfikacji ryzyka z uwzględnieniem jego możliwych konsekwencji (Alexander, 2008; Daníelsson, 2011; Dowd, 2005; Trzpiot, 2010).

Z punktu widzenia funkcjonowania gospodarki szczególnie ważną kategorią ryzyka jest ryzyko rynkowe (ang. *market risk*). Pojawia się ono wtedy, gdy na realizację transakcji kupna-sprzedaży pomiędzy stronami popytową i podażową wpływają różne czynniki rynkowe, najczęściej niezależne od realizowanego przedsięwzięcia. Najogólniej mówiąc, ryzyko rynkowe związane jest z prawdopodobieństwem wystąpienia straty, która stanowi efekt nieprzewidywalnego kierunku zmiany ceny przedmiotu transakcji. Stąd wniosek, że ryzyko rynkowe związane jest z każdym rynkiem, na którym prowadzi się wymianę handlową (materialną lub niematerialną), w tym z rynkiem kapitałowym (akcje, obligacje), instrumentów pochodnych (opcji, kontraktów terminowych na różnego rodzaju produkty bazowe), towarów i surowców, a także ze zmianami kursów walut (Holliwell, 2001; Jajuga, 2003).

Aby efektywne kontrolowanie ryzyka było możliwe, należy postrzegać to zagadnienie jako proces. Wyróżnia się następujące elementy procesu zarządzania ryzykiem (Christoffersen, 2012):

- identyfikacja ryzyka;
- pomiar i analiza ryzyka;
- opracowanie strategii ograniczającej lub minimalizującej ryzyko;
- implementacja strategii;
- kontrola, monitorowanie i korygowanie strategii ograniczającej lub minimalizującej ryzyko.

Każdy z powyższych elementów należy traktować równorzędnie.

W literaturze przedmiotu opisano wiele metod pomiaru ryzyka. Jedną z nich jest analiza wrażliwości, polegająca na dokonaniu oceny stopnia reakcji zmiennej ryzyka na zmienność określonych czynników, które to ryzyko determinują, za pomocą miary wrażliwości. Miary wrażliwości umożliwiają przeprowadzenie analizy przyczyn ryzyka, ponieważ obrazują wpływ czynników na zmienną ryzyka dzięki pewnej funkcji zależności od zmiennej ryzyka (Van Groenendaal, 1995). Matematyczna postać funkcji zależności zmiennej ryzyka od czynników ryzyka przyjmuje zatem postać

$$R = g(F_1, F_2, \dots, F_i, \varepsilon), \tag{9}$$

gdzie:

R – zmienna ryzyka (zmienna losowa),

F_i – i-ty czynnik ryzyka ($i = 1, 2, \dots, q$),

$g(\cdot)$ – funkcja określająca zależność pomiędzy zmienną ryzyka a czynnikami ryzyka,
ε – składnik losowy.

Funkcja $g(\cdot)$, określająca zależność między zmienną ryzyka a czynnikami ryzyka, może być liniowa lub nieliniowa. W przypadku liniowej postaci funkcji $g(\cdot)$ zmienna ryzyka przyjmuje postać

$$R = \beta_1 F_1 + \beta_2 F_2 + \cdots + \beta_i F_i + \varepsilon. \tag{10}$$

Miara wrażliwości powiązana z funkcją $g(\cdot)$ definiowana jest jako pierwsza pochodna tej funkcji względem i-tego czynnika ryzyka:

$$\beta_i = \frac{\partial R}{\partial F_i}. \tag{11}$$

W szczególności:

- jeśli $g(\cdot)$ jest funkcją liniową, to miara wrażliwości dla i-tego czynnika jest parametrem występującym przy i-tym czynniku ryzyka w modelu liniowym;
- jeśli $g(\cdot)$ jest funkcją nieliniową, to wykorzystując rozwinięcie ΔR w szereg Taylora, miarę wrażliwości dla i-tego czynnika wyznacza się jako $\Delta R \approx \frac{\partial R}{\partial F_i}\Delta F_i$.

Należy zwrócić uwagę na podobieństwo funkcji definiujących semiparametryczny model jednowskaźnikowy (1) i (2) z funkcją $g(\cdot)$, określającą zależność między zmienną ryzyka a czynnikami ryzyka (9), co w nomenklaturze modelu semiparametrycznego można zapisać jako

$$y_i = m\big(\varphi(\boldsymbol{x}_i, \boldsymbol{\beta})\big) + u_i \Rightarrow R = m\big(\varphi(\boldsymbol{F}, \boldsymbol{\beta})\big) + \varepsilon_i. \tag{12}$$

Stąd wniosek, że zmienna ryzyka może zostać zapisana przy wykorzystaniu nieznanej funkcji wygładzającej $m(\cdot)$, znanej funkcji parametrycznej $\varphi(\cdot,\cdot)$, q-wymiarowego wektora czynników $\boldsymbol{F}$ i p-wymiarowego wektora $\boldsymbol{\beta}$. Tę własność wykorzystano w przeprowadzonym badaniu.

4. Analiza ryzyka wybranych spółek notowanych na GPW

4.1. Metoda badania

W badaniu empirycznym przeprowadzono analizę ryzyka z wykorzystaniem semiparametrycznego modelu jednowskaźnikowego, uwzględniając koncepcję miar wrażliwości. Do analizy wybrano 10 spółek sektora informatycznego notowanych na GPW

od 4 stycznia 2018 r. do 30 grudnia 2021 r.: Ailleron, Atende, Betacom, Comarch, Ifirmę, LiveChat, LSI, NTT, Talex i Wasko. Dane zaczerpnięto z serwisu finansowego Bloomberg. Wyboru spółek dokonano na podstawie kryterium zbieżności liczby dni notowań danej spółki z liczbą notowań WIG20 w badanym okresie. Jako czynniki ryzyka wskazano dwa indeksy giełdowe: WIG20 i S&P 500. Na podstawie dziennych cen zamknięcia wyznaczono dzienne logarytmiczne stopy zwrotu zgodnie ze wzorem $R_t = 100 \ln \frac{p_t}{p_{t-1}}$ (gdzie p_t oznacza cenę zamknięcia w dniu t, a p_{t-1} – cenę zamknięcia w dniu $t-1$). Estymację semiparametrycznego modelu jednowskaźnikowego przeprowadzono, wykorzystując procedurę Ichimury (1993), w której zastosowano estymator typu *leave-one-out* średniej warunkowej R względem $\boldsymbol{F\beta}$ przy jednoczesnej minimalizacji sumy kwadratów reszt względem q-wymiarowego wektora $\boldsymbol{\beta}$.

4.2. Wyniki analizy empirycznej

Badanie rozpoczęto od przeprowadzenia analizy podstawowych charakterystyk stóp zwrotu badanych aktywów. Jej wyniki zawiera tabl. 1.

Tabl. 1. Statystyki opisowe dla szeregów stóp zwrotu analizowanych spółek notowanych na GWP w latach 2018–2021

Wyszczególnienie	*M*	S_d	W_{zm}	*S*	*K*	Min	Max	Test normalności	
								K-S	wartość *p**
WIG20	–0,008	1,410	–16924,324	–1,150	12,640	–14,246	6,335	0,069	<0,001
S&P 500	0,057	1,326	2334,836	–1,076	19,000	–12,765	8,968	0,127	<0,001
Ailleron	–0,038	3,322	–8685,027	0,308	5,604	–21,492	19,236	0,089	<0,001
Atende	0,036	2,392	6647,479	0,237	5,237	–14,540	13,353	0,151	<0,001
Betacom	–0,051	2,752	–5384,450	–0,426	1,707	–14,023	7,688	0,152	<0,001
Comarch	–0,003	2,073	–73124,217	–0,077	4,162	–12,851	9,345	0,074	<0,001
Ifirma	0,202	2,863	1418,183	0,101	2,920	–13,352	14,774	0,071	<0,001
LiveChat	0,120	2,581	2142,977	0,426	3,777	–11,118	14,228	0,089	<0,001
LSI	–0,009	2,573	–30197,148	0,203	9,833	–22,182	18,740	0,132	<0,001
NTT	0,096	2,906	3024,925	0,514	5,142	–19,416	15,207	0,130	<0,001
Talex	–0,009	2,654	–28883,459	–0,318	6,369	–18,469	14,898	0,157	<0,001
Wasko	–0,030	2,810	–9284,848	–0,381	7,090	–21,911	11,772	0,104	<0,001

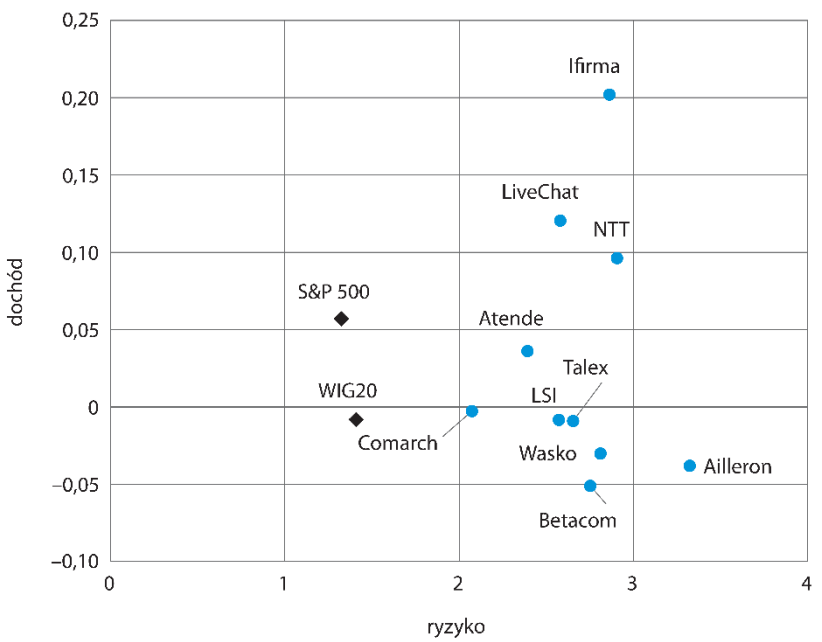
Uwaga. *M* – średnia, S_d – odchylenie standardowe, W_{zm} – współczynnik zmienności, *S* – skośność, *K* – kurtoza istotność, Min – minimum, Max – maksimum. * – istotność statystyczna na poziomie 0,01.

Źródło: obliczenia własne na podstawie danych z serwisu Bloomberg.

Zaobserwowano, że większość analizowanych walorów charakteryzowała się średnio ujemną stopą zwrotu. Ponadto inwestycje w WIG20 generowały w badanym okresie średniookresowe straty, a indeks S&P 500 uzyskał średnio dodatnią stopę zwrotu. Odnotowano także wysoką zmienność stóp zwrotu. Większość rozkładów wykazywała asymetrię prawostronną. Zauważono również silną leptokurtyczność i rozbieżność z rozkładem normalnym (odrzucono hipotezę zerową w teście zgodności K-S).

Z punktu widzenia potencjalnych inwestycji decydent jest zazwyczaj zainteresowany oczekiwanym poziomem zysku lub straty i związanym z tym ryzykiem. Na wykr. 1 zestawiono te dwie charakterystyki dla analizowanych walorów.

Wykr. 1. Mapa ryzyko – dochód dla analizowanych aktywów GPW w latach 2018–2021



Źródło: opracowanie własne na podstawie danych z serwisu Bloomberg.

Z informacji zamieszczonych na wykr. 1 wynika, że WIG20 i S&P 500 wykazywały niższy poziom ryzyka w porównaniu z pozostałymi walorami. Najbardziej ryzykowne spośród badanych aktywów były Ailleron, NTT i Ifirma, a najmniej – Comarch. Najbardziej dochodowe okazały się natomiast inwestycje w walory Ifirmy, a najmniej – Betacomu. Wszystkie spółki (poza Talexem) wykazywały także statystycznie istotną, dodatnią korelację z oboma indeksami (tabl. 2).

Tabl. 2. Współczynniki korelacji Pearsona pomiędzy analizowanymi aktywami GPW w latach 2018–2021

Wyszczególnienie	WIG20	S&P 500	Ailleron	Atende	Betacom	Comarch	Ifirma	LiveChat	LSI	NTT	Talex	Wasko
WIG20	1,000	0,497*	0,290*	0,177*	0,137*	0,280*	0,206*	0,251*	0,324*	0,102*	0,044	0,155*
S&P 500	0,497*	1,000	0,169*	0,109*	0,072*	0,207*	0,176*	0,139*	0,219*	0,071*	0,046	0,116*
Ailleron	0,290*	0,169*	1,000	0,111*	0,048	0,103*	0,087*	0,158*	0,172*	0,121*	–0,002	0,066*
Atende	0,177*	0,109*	0,111*	1,000	0,013	0,024	0,040	0,058	0,103*	0,121*	0,053	0,075*
Betacom	0,137*	0,072*	0,048	0,013	1,000	0,070*	0,022	0,033	0,110*	0,082*	0,019	0,053
Comarch	0,280*	0,207*	0,103*	0,024	0,070*	1,000	0,153*	0,123*	0,176*	–0,055	0,036	0,006
Ifirma	0,206*	0,176*	0,087*	0,040	0,022	0,153*	1,000	0,085*	0,154*	0,036	0,008	0,015
LiveChat	0,251*	0,139*	0,158*	0,058	0,033	0,123*	0,085*	1,000	0,097*	0,022	0,053	0,052
LSI	0,324*	0,219*	0,172*	0,103*	0,110*	0,176*	0,154*	0,097*	1,000	0,027	0,046	0,072*
NTT	0,102*	0,071*	0,121*	0,121*	0,082*	–0,055	0,036	0,022	0,027	1,000	0,010	0,005
Talex	0,044	0,046	–0,002	0,053	0,019	0,036	0,008	0,053	0,046	0,010	1,000	0,017
Wasko	0,155*	0,116*	0,066*	0,075*	0,053	0,006	0,015	0,052	0,072*	0,005	0,017	1,000

Uwaga. * – istotność statystyczna na poziomie 0,05.

Źródło: obliczenia własne na podstawie danych z serwisu Bloomberg.

Na kolejnym etapie badania oszacowano semiparametryczny model jednowskaźnikowy określany dla każdej analizowanej spółki sektora informatycznego GPW, gdzie czynnikiem ryzyka były stopy zwrotu indeksów WIG20 i S&P 500. Zastosowano procedurę estymacji Ichimury, wykorzystując gaussowską funkcję jądra. Wyniki oszacowań przedstawiono w tabl. 3.

Tabl. 3. Oszacowania parametrów semiparametrycznych modeli jednowskaźnikowych dla analizowanych aktywów GWP w latach 2018–2021

Spółki	Współczynniki	Wartości współczynników	Wartość p^* dla testu t	Szerokość pasma	Funkcja celu	R^2	MAE
Ailleron	$\hat{\beta}_{WIG20}$	0,64347	<0,001	0,23865	3,43276	0,176	2,181
	$\hat{\beta}_{S\&P\ 500}$	0,08356	0,003				
Atende	$\hat{\beta}_{WIG20}$	0,27754	<0,001	0,28711	4,66361	0,204	2,356
	$\hat{\beta}_{S\&P\ 500}$	0,05016	0,001				
Betacom	$\hat{\beta}_{WIG20}$	0,26387	<0,001	0,21953	3,21873	0,193	2,729
	$\hat{\beta}_{S\&P\ 500}$	0,00900	0,001				
Comarch	$\hat{\beta}_{WIG20}$	0,34570	<0,001	0,20779	3,98047	0,147	1,986
	$\hat{\beta}_{S\&P\ 500}$	0,14046	0,005				
Ifirma	$\hat{\beta}_{WIG20}$	0,31893	<0,001	0,31653	5,29474	0,164	2,794
	$\hat{\beta}_{S\&P\ 500}$	0,21110	0,002				
LiveChat	$\hat{\beta}_{WIG20}$	0,44187	<0,001	0,27931	4,31885	0,221	1,874
	$\hat{\beta}_{S\&P\ 500}$	0,03700	0,001				
LSI	$\hat{\beta}_{WIG20}$	0,51985	<0,001	0,36271	5,96080	0,178	2,501
	$\hat{\beta}_{S\&P\ 500}$	0,15067	0,002				
NTT	$\hat{\beta}_{WIG20}$	0,18278	<0,001	0,53862	6,32957	0,148	2,893
	$\hat{\beta}_{S\&P\ 500}$	0,05961	0,001				
Talex	$\hat{\beta}_{WIG20}$	0,05308	<0,001	0,61496	6,99463	0,093	2,653
	$\hat{\beta}_{S\&P\ 500}$	0,06349	0,001				
Wasko	$\hat{\beta}_{WIG20}$	0,25670	<0,001	0,56328	6,23192	0,113	2,776
	$\hat{\beta}_{S\&P\ 500}$	0,10992	0,001				

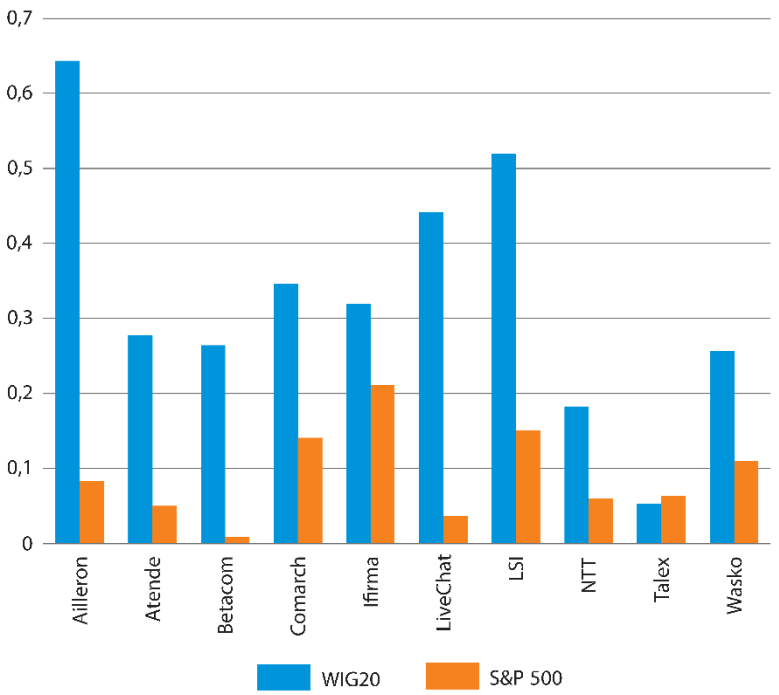
Uwaga. * – istotność statystyczna na poziomie 0,01.

Źródło: obliczenia własne na podstawie danych z serwisu Bloomberg.

Wyniki zamieszczone w tabl. 3 pokazują, że wybrane indeksy są statystycznie istotnymi czynnikami zmienności stóp zwrotu analizowanych spółek. Wartości oszacowań $\widehat{\boldsymbol{\beta}}$ wektora $\boldsymbol{\beta}$ dla WIG20 są wyższe niż dla S&P 500. To prawdopodobnie wynik powiązania analizowanych spółek z polskim rynkiem kapitałowym. W badanym okresie na zmienność WIG20 najwrażliwsze były stopy zwrotu Ailleronu, LSI i LiveChatu, a na zmienność S&P 500 – Ifirmy, LSI i Comarchu. Wszystkie oszacowane modele są istotne statystycznie, a wartości współczynnika R^2 wskazują, że jakość dopasowania jest niska (najmniejszą wartość współczynnika uzyskał Talex). Analizując wartości błędu MAE, zaobserwowano z kolei, że najmniejsze błędy wystąpiły w przypadku LiveChatu i Comarchu. Zauważono ponadto, że wartości funk-

cji celu wzrastają wraz ze zwyżką wartości parametru wygładzania (szerokości pasma). Oszacowania wartości wektora $\boldsymbol{\beta}$ zaprezentowano na wykr. 2.

Wykr. 2. Oszacowania wartości współczynnika wektora $\boldsymbol{\beta}$ semiparametrycznych modeli jednowskaźnikowych dla analizowanych aktywów GPW w latach 2018–2021



Źródło: opracowanie własne na podstawie danych z serwisu Bloomberg.

Jak widać na wykresie, największe różnice w oszacowaniach współczynników $\widehat{\boldsymbol{\beta}}$ w semiparametrycznym modelu jednowskaźnikowym dotyczyły Aillerona, LiveChatu i LSI, a najmniejsze – Talexu.

5. Podsumowanie

W artykule przedstawiono wyniki pomiaru ryzyka rynkowego z wykorzystaniem jednowskaźnikowego semiparametrycznego modelu ekonometrycznego. Badaniem objęto wybrane spółki sektora informatycznego notowane na GPW w latach 2018–2021. Analizowano dzienne logarytmiczne stopy zwrotu badanych aktywów, a jako czynniki ryzyka wskazano dwa indeksy giełdowe: polski WIG20 i amerykański S&P 500. Estymację semiparametrycznego modelu jednowskaźnikowego przepro-

wadzono z zastosowaniem procedury Ichimury bazującej na estymatorze typu *leave--one-out*. Ponadto jako funkcję jądra wykorzystano funkcję gaussowską.

Model semiparametryczny opiera się na estymacji elastycznej i powstaje dzięki połączeniu modeli parametrycznego i nieparametrycznego. Modele semiparametryczne są przydatne w sytuacjach, w których modele w pełni parametryczne i w pełni nieparametryczne nie dają właściwych oszacowań. Skutecznym sposobem redukcji zjawiska nadmiarowości, z którym można się spotkać podczas przeprowadzania analizy dużych zbiorów danych, jest wykorzystanie modelu jednowskaźnikowego. W takim modelu występuje jednowymiarowy wskaźnik, który stanowi element nieznanej funkcji nieparametrycznej. Modele jednowskaźnikowe działają podobnie jak wiele znanych modeli parametrycznych, co potwierdza ich użyteczność. Jednowskaźnikowy semiparametryczny model ekonometryczny ma tę przewagę nad modelem liniowym, że pozwala na uwzględnienie nieliniowych zależności między zmiennymi objaśniającymi a zmienną zależną, nie wymaga założeń dotyczących rozkładu i homoskedastyczności składnika losowego oraz umożliwia uwzględnienie efektów nieliniowych zmiennych w wariancji błędów. Ponadto szacunki składnika parametrycznego są asymptotycznie zbieżne z rzeczywistymi w tempie zbliżonym do modelu w pełni parametrycznego.

Na podstawie przeprowadzonego badania wykazano, że zarówno WIG20, jak i S&P 500 były istotnymi czynnikami zmienności stóp zwrotu badanych spółek sektora informatycznego w analizowanym okresie. Oszacowania współczynników $\widehat{\boldsymbol{\beta}}$ semiparametrycznych modeli jednowskaźnikowych dla indeksu polskiej giełdy były znacznie wyższe niż oszacowania dla S&P 500 (wyjątek stanowiła spółka Talex). Oba indeksy przyjmowały wartości dodatnie, poniżej progu 1. Świadczy to o tym, że stopy zwrotu badanych spółek zmieniały się w tym samym kierunku co stopy zwrotu indeksów, jednak ich reakcje na zmiany stopy zwrotu rynków polskiego i amerykańskiego były niewielkie. Nawet gdy spadek stóp zwrotu rynku był znaczny, stopy zwrotu badanych akcji zmniejszyły się tylko nieznacznie. Prowadzi to do wniosku, że akcje badanych spółek cechowało ryzyko niższe od rynkowego (są to akcje defensywne).

Z analizy oszacowanych semiparametrycznych modeli jednowskaźnikowych wynika, że w przypadku badanych spółek parametr wygładzania zawierał się w przedziale 0,20–0,62. Dodatkowo wraz ze wzrostem wartości tego parametru zwiększała się wartość funkcji celu wykorzystywanej do oszacowywania współczynników $\widehat{\boldsymbol{\beta}}$ modeli. Warto jednak zwrócić uwagę, że jakość dopasowania określona za pomocą wartości współczynnika R^2 jest niezadowalająca. Największą wartość uzyskano dla modelu spółek LiveChat i Atende, a najmniejszą – dla Talexu. Zaobserwowano także małe wartości błędu MAE (najmniejsze dla LiveChatu i Comarchu). Te wyniki są zbieżne z wynikami oceny korelacji pomiędzy badanymi spółkami i indeksami.

Semiparametryczne modele jednowskaźnikowe mają wiele zalet przydatnych w analizie ryzyka na giełdzie. Przede wszystkim są bardziej elastyczne – w przeciwieństwie do modeli parametrycznych (np. CAPM) nie narzucają szczegółowych założeń dotyczących formy funkcyjnej relacji między zmiennymi. Dzięki temu mogą lepiej dopasowywać się do rzeczywistych danych i trafniej uchwycić złożone zależności zachodzące między zmiennymi. Ponadto pozwalają na modelowanie wpływu wielu indeksów giełdowych na stopę zwrotu akcji, co może prowadzić do precyzyjniejszych oszacowań ryzyka. Za ich inną zaletę należy uznać uwzględnianie zmienności czasowej, ponieważ współczynnik $\widehat{\boldsymbol{\beta}}$ jest funkcją zmiennych niezależnych i może zmieniać się w czasie. To pozwala na uwzględnienie zmienności ryzyka systematycznego w czasie, co jest istotne w dynamicznie zmieniającym się środowisku giełdowym.

Modele jednoindeksowe mają również wady. Estymacja semiparametryczna może okazać się bardziej złożona i czasochłonna niż estymacja parametryczna, a uzyskane wyniki – wrażliwsze na wybór metody estymacji i parametrów modelu. Ważne zatem, aby przed zastosowaniem tych modeli dobrze zrozumieć ich założenia i związane z nimi ograniczenia.

Z punktu widzenia inwestora korzystanie z semiparametrycznego modelu jednowskaźnikowego w analizie ryzyka giełdowego może przynieść wiele korzyści. Modele tej klasy pozwalają na dokładniejszą ocenę ryzyka, ponieważ w ich przypadku nie zakłada się konkretnej formy funkcjonalnej zależności między zmiennymi, dzięki czemu mogą one lepiej pasować do rzeczywistych danych i zapewniać dokładniejsze oceny ryzyka, a tym samym prowadzić do właściwszych decyzji inwestycyjnych. Semiparametryczne modele jednowskaźnikowe pozwalają także trafniej uchwycić złożone zależności między stopami zwrotu a indeksami giełdowymi, co powinno pomóc inwestorom zrozumieć, jakie czynniki wpływają na ryzyko ich inwestycji. Łatwo je dostosować do zmieniających się warunków giełdowych, co w konsekwencji pozwala inwestorom na dokonywanie bardziej świadomych decyzji inwestycyjnych. Dokładniejsza ocena ryzyka przekłada się ponadto na możliwość lepszego alokowania kapitału. Inwestorzy mają możliwość przeznaczenia większego kapitału na aktywa o niższym ryzyku i wyższym potencjalnym zwrocie, czym zwiększają swoje szanse na osiągnięcie większych korzyści przy określonym poziomie ryzyka.

Należy przy tym pamiętać, że korzystanie z semiparametrycznych modeli ekonometrycznych nie gwarantuje sukcesu inwestycyjnego. Ważne, aby inwestorzy mieli świadomość ograniczeń tych modeli i traktowali je jako jedno z wielu narzędzi analizy inwestycyjnej.

Bibliografia

Alexander, C. (2008). *Market Risk Analysis* (vol. 1–4). John Wiley & Sons.

Cameron, A. C., Trivedi, P. K. (2005). *Microeconometrics. Methods and Applications*. Cambridge University Press.

Christoffersen, P. F. (2012). *Elements of Financial Risk Management* (2nd edition). Elsevier.

Daníelsson, J. (2011). *Financial Risk Forecasting. The Theory and Practice of Forecasting Market Risk with Implementation in R and MATLAB*. John Wiley & Sons.

Dowd, K. (2005). *Measuring Market Risk* (2nd edition). John Wiley & Sons.

Duan, N., Li, K. C. (1991). Slicing regression: A link-free regression method. *The Annals of Statistics*, *19*(2), 505–530. https://doi.org/10.1214/aos/1176348109.

Fix, E., Hodges, J. L. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. https://apps.dtic.mil/sti/pdfs/ADA800276.pdf.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, *21*(1), 157–178. https://doi.org/10.1214/aos/1176349020.

Härdle, W., Stoker, T. M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, *84*(408), 986–995. https://doi.org/10.2307/2290074.

Henderson, D. J., Parmeter, C. F. (2015). *Applied Nonparametric Econometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511845765.

Holliwell, J. (2001). *Ryzyko finansowe. Metody identyfikacji i zarządzania ryzykiem finansowym*. Wydawnictwo Liber.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics*, *58*(1–2), 71–120. https://doi.org/10.1016/0304-4076(93)90114-K.

Jajuga, K. (2003). *Metody statystyczne w finansach*. StatSoft Polska. https://media.statsoft.pl/_old_dnn/downloads/jajuga.pdf.

Krężołek, D. (2020). *Modelowanie ryzyka na rynku metali*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of American Statistical Association*, *86*(414), 316–327. https://doi.org/10.2307/2290563.

Li, G., Peng, H., Dong, K., Tong, T. (2014). Simultaneous confidence bands and hypothesis testing for single-index models. *Statistica Sinica*, *24*(2), 937–955. http://dx.doi.org/10.5705/ss.2012.127.

Li, Q., Racine, J. S. (2007). *Nonparametric Econometrics. Theory and Practice*. Princeton University Press.

Pagan, A., Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612503.

Powell, J. L., Stock, J. H., Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica, 57*, 1403–1430.

Racine J. S. (2008). Nonparametric Econometrics. *A Primer. Foundations and Trends in Econometrics, 3*(1), 1–88.

Ramanathan R. (1989). *Introductory Econometrics with Applications*. Harcourt Brace Jovanovich.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica, 54*(6), 1461–1481.

Trzpiot, G. (red.). (2010). *Wielowymiarowe metody statystyczne w analizie ryzyka inwestycyjnego*. Polskie Wydawnictwo Ekonomiczne.

Van Groenendaal, W. J. H. (1995). Measuring Risk: Risk Analysis or Sensitivity Analysis?. *IFAC Proceedings Volumes*, *28*(7), 443–450. https://doi.org/10.1016/S1474-6670(17)47145-X.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer. https://doi.org/10.1007/0-387-30623-4.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, *22*(6), 1112–1137. https://doi.org/10.1017/S0266466606060531.

Xia, Y., Tong, H., Li, W. K. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *64*(3), 363–410. https://doi.org/10.1111/1467-9868.03412.

Xue, L. (2013). Estimation and empirical likelihood for single-index models with missing data in the covariates. *Computational Statistics and Data Analysis*, *68*, 82–97. https://doi.org/10.1016/j.csda.2013.06.017.

Xue, L., Zhu, L. (2006). Empirical likelihood for single-index model. *Journal of Multivariate Analysis*, *97*(6), 1295–1312. https://doi.org/10.1016/j.jmva.2005.09.004.

Yang, Y., Song, Q. (2014). Jump detection in time series nonparametric regression models: a polynomial spline approach. *Annals of the Institute of Statistical Mathematics*, *66*(2), 325–344. https://doi.org/10.1007/s10463-013-0411-3.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press. https://doi.org/10.1017/CBO9780511615887.

Zhu, L., Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *68*(3), 549–570. https://doi.org/10.1111/j.1467-9868.2006.00556.x.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 24–48 https://doi.org/10.59139/ws.2023.10.2
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 20.09.2022, poprawiony/revised: 16.05.2023, zaakceptowany/accepted: 4.07.2023

Mass appraisal: a statistical approach to determining the impact of a property's attributes on its value

Anna Gdakowicz,[a] Ewa Putek-Szeląg[b]

Abstract. The mass valuation of real estate refers to the simultaneous estimation of the values of a large number of properties using the same method. This method should involve automation that would reduce the human element in the process. The algorithm that meets these requirements is the Szczecin Mass Valuation Algorithm for Real Estate (SAMWN), which was used to determine the values of selected land properties in Szczecin. The article presents a modification of the SAMWN which consists in an objective calculation of the influence of the attributes of a property on its value using dependency coefficients. Various approaches have been proposed to assigning weights to the attributes included in the model. The aim of the study was twofold. Firstly, to identify the coefficients of dependencies that can be used in property valuation based on SAMWN and in the analysis of a property using attributes measured on an ordinal scale. The second aim was to select such a combination of methods for taking into account the influence and weights of attributes which under the SAMWN procedure would produce results closest to property values determined by real estate appraisers. The study used data on 405 land properties located in Szczecin intended for residential purposes, valued individually by real estate appraisers for the purpose of the research.

The study proposes four types of dependency coefficients and their partial values and four ways of including these coefficients in the SAMWN procedure. Additionally, the study assesses six methods of weighing the proposed measures. As a result, 168 ways of measuring the influence of individual attributes on property value were obtained. In order to determine which variants produced values closest to the real values (estimated by real estate appraisers), appraisal error measures were calculated and linear ranking procedures were then adopted to identify the best combination of the applied variants.

The presented mass valuation algorithm may be applied to the estimation of values of various types of properties. However, it requires the procedure to be adapted to the specific characteristics of the appraised property, i.e. the attractiveness zones should be determined as well as the attributes that are relevant to the specific type of property.

Keywords: mass real estate valuation, statistical methods, mass valuation algorithm, dependency coefficients, linear ordering

JEL: C10, R30

[a] Uniwersytet Szczeciński, Instytut Ekonomii i Finansów, Polska / University of Szczecin, Institute of Economics and Finance, Poland. ORCID: https://orcid.org/0000-0002-4360-3755.
E-mail: anna.gdakowicz@usz.edu.pl.

[b] Uniwersytet Szczeciński, Instytut Ekonomii i Finansów, Polska / University of Szczecin, Institute of Economics and Finance, Poland. ORCID: https://orcid.org/0000-0003-0364-615X. Autor korespondencyjny / Corresponding author: ewa.putek-szelag@usz.edu.pl.

Masowa wycena nieruchomości. Statystyczny sposób określenia wpływu cech nieruchomości na jej wartość

Streszczenie. O masowej wycenie nieruchomości mówi się, gdy wartości dużej liczby nieruchomości są szacowane w tym samym czasie przy użyciu tej samej metody. Zastosowana metoda powinna charakteryzować się automatyzacją ograniczającą udział człowieka. Założenia te spełnia Szczeciński Algorytm Masowej Wyceny Nieruchomości (SAMWN), wykorzystywany do wyceny wartości wybranych nieruchomości gruntowych w Szczecinie. W artykule przedstawiono modyfikację SAMWN, polegającą na obiektywnym obliczaniu wpływu cech nieruchomości na jej wartość z wykorzystaniem współczynników zależności. Ponadto zaproponowano różne sposoby nadawania wag atrybutom użytym w modelu. Cel badania omawianego w artykule jest dwojaki. Po pierwsze polega na wskazaniu tych współczynników zależności, które mogą być użyte do wyceny nieruchomości opartej na SAMWN oraz do analizy nieruchomości wykorzystującej atrybuty mierzone na skali porządkowej. Po drugie celem jest wyłonienie takiej kombinacji metod uwzględniania wpływu i wag atrybutów, której zastosowanie w procedurze SAMWN dałoby wyniki najbliższe wartościom nieruchomości oszacowanym przez rzeczoznawców majątkowych. W badaniu wykorzystano dane dotyczące 405 nieruchomości gruntowych w Szczecinie przeznaczonych na cele mieszkaniowe, które na potrzeby badania zostały indywidualnie wycenione przez rzeczoznawców.

Zaproponowano cztery rodzaje współczynników zależności i ich wartości cząstkowych oraz cztery sposoby uwzględniania tych współczynników w procedurze SAMWN. Ponadto przeprowadzono ocenę zastosowania sześciu metod ważenia proponowanych miar. W efekcie uzyskano 168 sposobów pomiaru wpływu poszczególnych atrybutów na wartość nieruchomości. W celu określenia, w którym przypadku uzyskane wartości wyceny są najbardziej zbliżone do wartości rzeczywistych (oszacowanych przez rzeczoznawców majątkowych), obliczono miary błędów wyceny, a następnie wdrożono procedury porządkowania liniowego, aby wskazać najlepszą kombinację rozpatrywanych wariantów.

Przedstawiony algorytm wyceny masowej może być zastosowany do szacowania wartości różnego rodzaju nieruchomości. Wymaga to jednak dostosowania procedury do specyfiki wycenianej nieruchomości, tj. określenia strefy atrakcyjności i atrybutów istotnych dla danego typu nieruchomości.

Słowa kluczowe: masowa wycena nieruchomości, metody statystyczne, algorytm wyceny masowej, współczynniki zależności, porządkowanie liniowe

1. Introduction

One of the problems occurring when estimating the value of a property is determining how the individual attributes of a property affect its value. Until now, this was estimated either on the basis of analyses of prices and market features of similar properties, or by analogy to other properties in terms of the type and area of markets, by examining or observing the preferences of potential buyers, or in other ways (Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, 2008). This article refers to the last option and presents an alternative way of dealing with the aforementioned issue. An approach based on statistical methods, using dependency

coefficients is introduced as an objective proposal to determine the impact of individual attributes on the value of real estate. The idea of including statistical methods (dependency coefficients) in the procedure of mass real estate valuation is the basis of the statistical approach of the Szczecin Mass Real Estate Valuation Algorithm (Pol. *Szczeciński Algorytm Masowej Wyceny Nieruchomości* – SAMWN).

Considering the specificity of the features (attributes) describing real estate, which in most cases may be presented on an ordinal scale, four coefficients measuring the strength of the relationship between the features and the value of the property were proposed. Since previous studies, e.g. Dmytrów et al. (2020), indicated that the results obtained for partial coefficients reflect the relationships found on the market more efficiently, the analysis was extended to include partial variants of selected coefficients. Then, six procedures for calculating weights in SAMWN were proposed. The obtained results were subsequently ordered on the basis of linear ordering procedures, as described in the work by Strahl (1978) and Pawlukowicz (2010).

The aim of the article was twofold. Firstly, to identify the coefficients of dependencies that can be used in property valuation based on SAMWN, and also in the analysis of those properties whose attributes are often described on an ordinal scale. The second aim was to identify a combination of methods describing the influence and weights of the attributes which under SAMWN procedure produces results closest to the property values determined by real estate appraisers.

2. Mass appraisal methods

2.1. Literature review

According to the literature, mass property valuation (Jahanshiri et al., 2011) requires the following conditions to occur simultaneously (Hozer et al., 2002; Kuryj, 2007; Telega et al., 2002):

- a large number of properties are subject to valuation;
- the properties are valued using the same method (making the results comparable);
- all properties must be valued at the same time, using the same data and price levels.

Since the first definition of mass appraisal was formulated in 1984 (International Association of Assessing Officers, 2019), there have been many theoretical ideas for the simultaneous valuation of multiple properties, followed by attempts to apply these ideas in practice. Scientists unanimously agreed that it was necessary to develop a method or model whose application would show the relationship between the property value and the factors affecting it. While statistical, geographical or computer science-based models were used for mass valuations, the proposed solutions can be classified as the 3I-trend: the AI-based model, the GIS-based model and the MIX-based model (Wang & Li, 2019).

The AI-based model includes:

- multiple regression analysis (MRA) – a relatively simple and easy-to-use method, although problematic in terms of choosing the analytical form of the function that best describes the relationship between the value and the selected attributes; additionally, the estimation error is often unsatisfactory (Lin & Mohan, 2011);
- an expert system and a decision support system – systems based on the knowledge of property appraisers. These systems do not learn by themselves but use expert knowledge to develop adjustment factors (Amidu & Boyd, 2018; Kilpatrick, 2011; Lam et al., 2009);
- artificial neural networks (ANNs) – the use of neural networks in the mass valuation process does not require the development of an initial model or validation. Based on a learning sample, the network learns by itself and provides the final results. Despite the fact that the results are at a satisfactory level of matching, researchers pay attention to a part of the ANN procedure called the 'black box' (Abidoye & Chan, 2018; Yacim & Boshoff, 2018; Zhou et al., 2018);
- tree-based model – a group of models based on a decision tree, random forest and boosted tree. These methods perform well in both classification and regression. The results obtained by using these methods are more accurate compared to other models (Antipov & Pokryshevskaya, 2012) and the computational process is fast. However, interpreting the results is problematic (McCluskey et al., 2014);
- hierarchical model – a model which takes into account the hierarchical structure of data. The hierarchical Bayesian approach (Hui et al., 2010), the analytical Bayesian approach (Cervelló-Royo et al., 2016) and the analytical hierarchical process have been used for property valuation;
- cluster analysis – the use of data grouping methods is also applied in mass real estate valuation. The similarity of real estate allows for the isolation of clusters of similar properties; however, it is important to explain the resulting subgroups in practice. Emphasis should also be placed on adopting correct assumptions according to which the cluster analysis process runs, as changing the assumptions changes the clusters. There are several types of classifications: model-based clustering, partitioning clustering, density-based clustering, hierarchical clustering, grid-based clustering, and fuzzy-based clustering (Calka, 2019; Gabrielli et al., 2017; Napoli et al., 2017);
- rough set theory (RST) and fuzzy set theory – methods used for underdeveloped, emerging markets with a low level of computerisation. The use of RST in the real estate market enables mass valuation even when the relevant factors influencing the property value are unknown, and can be used to adjust data weights even if the similarity is at a low level (Alcantud et al., 2017; Guan et al., 2014; Lasota et al., 2011; Ma et al., 2018);
- other models – a group of various other classic models, e.g. the genetic algorithm (GA), the support vector machine (SVM), data envelopment analysis (DEA), and

conformal predictors (CP). The use of these models in the real estate market has only just begun, but their application has already yielded promising results (Ahn et al., 2012; Bellotti, 2017; Chen, Z. et al., 2017; Morano et al., 2018; Zurada et al., 2011).

The GIS-based models involve the use of the geoinformation system/science (GIS). Since an attribute related to its spatial position can be created for each property, scientists take advantage of this possibility and combine the information with other attributes that can affect the property's value. Among these models, the following stand out:

- geographically weighted regression (GWR) – the most commonly used GIS model; it is an extension of the MRA model with spatial analysis. It enables regression analysis for each location. Models of this type are easy to apply and the obtained results allow for an interpretation (Dimopoulos & Moulas, 2016; Lockwood & Rossini, 2011; McCluskey & Borst, 2011);
- geographically weighted principal component analysis – models that refer to principal component analysis (PCA) methods, extended to include spatial elements, allowing the analysis to take into account e.g. the presence of submarkets (Wu et al., 2018);
- spatial error model (SEM) and spatial lag model (SLM) – the most frequently used extension of MRA models with spatial dependence. The SEM assumes that the occurrence of an error in the property valuation depends on the error of its surroundings (Zhang et al., 2015), while the SLM uses the spatially lagged dependent variable of the regression model: the price of a property depends on the prices of the neighbouring properties (Anselin, 2002).

The group of MIX-based models includes a wide range of models (constantly developed) that use (mix) simultaneously different approaches. Researchers propose e.g. forecasting according to different methods and, based on the obtained results, determining the weighted average of the component forecasts (Glennon et al., 2018), combining traditional estimation models with AI and GIS methods (Calka & Bielecka, 2016), or using information from other innovative sources in known models (Chen, J.-H. et al., 2017).

2.2. The SAMWN

As mentioned above, mass valuation involves appraising multiple properties at the same time using the same tool (algorithm). To ensure the comparability of the results and that the generalisation of the calculations is possible, the tool which is used should be automated and the impact of the human factor should be limited to a minimum. SAMWN is such a tool. The algorithm is classified as a MIX-based model, as it represents a non-classical approach. On the one hand, the knowledge of

real estate appraisers is used (expert system and decision support system), on the other hand, however, the structure of the model refers to hierarchical solutions, and the entire valuation process is automated. SAMWN can be presented as follows:

$$w_{ji} = wwr_j \cdot pow_i \cdot w_{\text{baz}} \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} \left(1 + A_{kpi}\right), \tag{1}$$

where:
w_{ji} is the i-th market value (or cadastral value) of the property in the j-th zone of location attractiveness ($j = 1, 2, \ldots, J$),
wwr_j is the market value coefficient in the j-th zone of location attractiveness,
J is the number of the location attractiveness zone,
pow_i is the area of the i-th real estate,
w_{baz} is the estimated value of 1 sqm of the property with the worst attribute states in the worst location attractiveness zone,
A_{kpi} is the impact of the p-th category of the k-th attribute for the i-th real estate ($k = 1, 2, \ldots, K$; $p = 1, 2, \ldots, k_p$),
K is the number of attributes,
k_p is the number of categories of the k-th attribute,
N is the number of valued properties ($i = 1, 2, \ldots, N$).

The algorithm determines the unit market or cadastral value of the property, not the price. Since the algorithm design (formula 1) requires the multiplication of individual factors, the reference point when determining the value of the property is the base value, i.e. the value of 1 sqm of the property with the worst attribute states, in the worst location attractiveness zone, theoretically of the lowest value. The base value is multiplied by the area of the property, the market value coefficient and the effect of the attribute states of the valued property.

The impact of attribute states (A_{kpi}) can be determined e.g. on the basis of quantitative methods – using statistical methods (Gdakowicz & Putek-Szeląg, 2020a, 2020b). Thus, the objectivity of the procedure is retained and the human factor is limited to a minimum.

The value of a property depends not only on its attributes but also on external factors presented by the demand side. Two properties, very similar in terms of attributes, can have different values if they are located in different zones of location attractiveness (Pol. *strefa atrakcyjności lokalizacji* – SAL). The algorithm influences this type of factor through market value coefficients (wwr_j). These coefficients are determined for each location's attractiveness zone and they show the impact of the

location. The market value ratio for the j-th zone of location attractiveness can be determined as

$$wwr_j = \sqrt[n_j]{\prod_{i=1}^{n_j} \frac{w_{ji}^{\mathrm{rz}}}{w_{ji}^{\mathrm{h}}}}, \tag{2}$$

where:
w_{ji}^{rz} is the value of the i-th property in the j-th zone of location attractiveness determined by a property appraiser,
w_{ji}^{h} is the hypothetical value of the i-th property in the j-th zone of location attractiveness,
n_j is the number of representative properties in the j-th zone of location attractiveness.

The market value ratios are calculated based on representative real estate valuations carried out by property appraisers on an individual basis, providing real property values (w_{ji}^{rz}). On the other hand, hypothetical values (w_{ji}^{h}) are calculated based on formula 1, but excluding market value coefficients:

$$w_{ji}^{\mathrm{h}} = pow_i \cdot w_{\mathrm{baz}} \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + A_{kpi}). \tag{3}$$

When the values of representative real estate (w_{ji}^{rz}) are known, as are their attribute states and their impact, their base value (w_{baz}) and the surfaces, the market value coefficients are estimated for each zone of location attractiveness. This involves calculating the geometric mean of the quotients of the real and hypothetical values of the real estate. As the market value coefficients for individual zones of location attractiveness become known, it is possible to estimate the market (cadastral) value of each property within the SAL taking into account the attribute states of the valued property.

The impact of the attributes on the value of a real estate can be calculated according to a statistical approach using dependency coefficients. It should, however, be noted that the use of statistical methods is always limited by certain formal requirements. Therefore, when examining the correlations between attributes in the real estate market, the following problems may occur (Dmytrów et al., 2020):

- the attributes describing real estate subject to valuation are very often qualitative, i.e. measured on an ordinal scale;

- the differences between the successive values of the attributes measured on the ordinal scale are not necessarily constant (changes are not linear);
- when examining the correlations between the value of 1 sqm and the individual attributes, it is necessary to eliminate the influence of other attributes which can significantly disrupt the examined interdependence.

Real estate attributes are most often measured on an ordinal scale, which is why rank-based ratios are the most natural to use, i.e. Spearman's rank correlation coefficient (Kendall, 1948), the Kendall coefficient (Han & Zhu, 2008; Parker et al., 2011) or generalised correlation coefficient Gamma (Siegel & Castellan, 1988).

In practice, however, property appraisers most often use Pearson's linear correlation coefficients[1] in the property valuation process, although they should not be used at all for such features. Partial coefficients have also been calculated for the above-mentioned coefficients to eliminate the influence of other variables on the attribute–property value relationship. If the value of the coefficient is lower than 0, then it is assumed that this attribute has an insignificant influence on the value of the property and, in the further calculations, 0 is assumed.

In addition, four variants were introduced to take into account the impact of the factors on the value of the property. Only statistically significant coefficients or the square of these coefficients and all coefficients (greater than 0), or their square have been considered in the further calculations. Detailed variants of the use of dependence coefficients are presented in Table 1.

Table 1. Dependence coefficients – calculation variants

Dependence coefficients	Significant coefficients	Square of significant coefficients	All coefficients	Square of all coefficients
Spearman's	Si.	Si2.	Sw.	Sw2.
Partial Spearman's	Sci.	Sci2.	Scw.	Scw2.
Kendall's	Ti.	Ti2.	Tw.	Tw2.
Partial Kendall's	Tci.	Tci2.	Tcw.	Tcw2.
Gamma	Gi.	Gi2.	Gw.	Gw2.
Pearson's	ri.	ri2.	rw.	rw2.
Partial Pearson's	rci.	rci2.	rcw.	rcw2.

Source: authors' work.

[1] It should be emphasised that the use of this coefficient is methodologically incorrect (property features / attributes are presented on an ordinal scale, while changes between feature variants are not linear); however, due to the widespread use of this measure by property appraisers, it has been included in the further calculations.

For each of the 28 methods of calculating the coefficients (Table 1), six methods of calculating the weights necessary in the $1 + A_{kpi}$ (formula 1) determination process were proposed (the number of the used method of calculating the weights was added to each coefficient, after the dot):

1. the average value ratio of the real estate with the best attribute states to the average value of the real estate with the weakest attribute states:

$$1 + a_{kp} = \mathrm{e}^{\frac{w_{\mathrm{Amax}k}}{w_{\mathrm{Amin}k}} u_{kp}}, \tag{4}$$

where:

$$u_{kp} = \frac{\rho}{\sum_{k=1}^{K} \rho} \cdot \frac{(p-1)}{(k_p - 1)}, \tag{5}$$

where:

u_{kp} is the weight of categories p of the k-th attribute (used also in methods described in items 2–6),

ρ is the dependence coefficient,

p is the number of states of the k-th attribute ($p = 1, \ldots, k_p$);

2. the median value ratio of the property with the best attribute states to the median value of the property with the weakest attribute states:

$$1 + \alpha_{kp} = \mathrm{e}^{\frac{w_{\mathrm{Mmax}k}}{w_{\mathrm{Mmin}k}} u_{kp}}; \tag{6}$$

3. the average value ratio of the property with the best attribute states to the average value of the property with the weakest attribute states, corrected by factor $\ln\left(\frac{w_{\max}}{w_{\min}}\right)$:

$$1 + \alpha_{kp} = \mathrm{e}^{\ln\left(\frac{w_{\max}}{w_{\min}}\right) \cdot \frac{w_{\mathrm{Amax}k}}{w_{\mathrm{Amin}k}} u_{kp}}; \tag{7}$$

4. the median value ratio of the property with the best attribute states to the median value of the weakest attribute states, corrected by factor $\ln\left(\frac{w_{\max}}{w_{\min}}\right)$:

$$1 + \alpha_{kp} = \mathrm{e}^{\ln\left(\frac{w_{\max}}{w_{\min}}\right) \cdot \frac{w_{\mathrm{Mmax}k}}{w_{\mathrm{Mmin}k}} u_{kp}}; \tag{8}$$

5. the coefficient ratio of the property value variation with the best attribute states to the coefficient of the property value variation with the weakest attribute states:

$$1 + \alpha_{kp} = \mathrm{e}^{\frac{w_{\mathrm{Vmax}k}}{w_{\mathrm{Vmin}k}} u_{kp}}; \tag{9}$$

6. the coefficient ratio of the property value variation with the best attribute states to the coefficient of the property value variation with the weakest attribute states, corrected by coefficient $\ln\left(\frac{w_{\max}}{w_{\min}}\right)$:

$$1 + \alpha_{kp} = \mathrm{e}^{\ln\left(\frac{w_{\max}}{w_{\min}}\right)\cdot\frac{w_{\mathrm{V}\max k}}{w_{\mathrm{V}\min k}}u_{kp}}, \tag{10}$$

where:
$w_{\mathrm{A}\max k}$, $w_{\mathrm{M}\max k}$, $w_{\mathrm{V}\max k}$ are: the average, the median, the variation coefficient of the property value for the highest category of the k-th attribute, respectively,
$w_{\mathrm{A}\min k}$, $w_{\mathrm{M}\min k}$, $w_{\mathrm{V}\min k}$ are: the average, the median, the variation coefficient of the property value for the lowest category of the k-th attribute, respectively,
$w_{\max}$ is the maximum value of the property in the data set,
$w_{\min}$ is the minimum value of the property in the data set.

Values $1 + \alpha_{kp}$ are determined for all attribute states. The value of $1 + A_{kpi}$ corresponds to a specific attribute state of the appraised property.

3. Methodology for selecting the best variant

Based on the applied variants of determining the impact of attributes in real estate valuation using SAMWN, 168 cases of real estate value estimations were received. The obtained real estate values were compared with the real values, valued individually by property appraisers. The measures used were as follows:
1. root means square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(w_{\mathrm{r}i} - w_{\mathrm{t}i})^2}{n}}, \tag{11}$$

where:
$w_{\mathrm{r}i}$ is the real unit value of the real estate determined by a real estate appraiser,
$w_{\mathrm{t}i}$ is the theoretical unit value of the property determined by the SAMWN,
n is the number of observations;
2. RMSE variation coefficient:

$$V_{\mathrm{RMSE}} = \frac{\mathrm{RMSE}}{\bar{w}_{\mathrm{r}i}} \cdot 100\%, \tag{12}$$

where:
$\bar{w}_{\mathrm{r}i}$ is the average value of the real unit values of the real estate designated by a property appraiser;

3. the mean absolute percentage error (MAPE):

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|w_{\mathrm{r}i} - w_{\mathrm{t}i}|}{w_{\mathrm{r}i}}; \tag{13}$$

4. based on the percentage error of PE:

$$\mathrm{PE}_i = \frac{w_{\mathrm{r}i} - w_{\mathrm{t}i}}{w_{\mathrm{r}i}} \cdot 100\%; \tag{14}$$

5. share of B^+ valuations for which $\mathrm{PE}_i > 0$:

$$\mathrm{B}^+ = \frac{\sum_{i=1}^{n^+}\mathrm{PE}_i^+}{n^+} \cdot 100\%; \tag{15}$$

6. share of B^- valuations for which $\mathrm{PE}_i < 0$:

$$\mathrm{B}^- = \frac{\sum_{i=1}^{n^-}\mathrm{PE}_i^-}{n^-} \cdot 100\%, \tag{16}$$

where:
n^+ is the number of observations for which $\mathrm{PE}_i > 0$,
n^- is the number of observations for which $\mathrm{PE}_i < 0$.

The linear ordering of the proposed variants for determining the impact of the attributes on the value of the real estate estimated using the SAMWN was carried out according to the valuation error measures (formulae 11–13 and 15–16). The next steps of the linear ordering procedure included:

1. preparing a data matrix based on the collected data (valuation errors calculated for the proposed variants);
2. the standardisation transformation of the variables (Gatnar & Walesiak, 2011; Zeliaś, 2002);
3. defining the variables as stimulants or destimulants (Hellwig, 1968);
4. applying the upper development pattern (ordering the elements of the set of objects according to the increasing values of the distance measure);
5. calculating the distances between objects by means of the GDM1 measure for variables measured on an interval scale (Walesiak, 2016):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m}(x_{ij} - x_{kj})(x_{kj} - x_{ij}) + \sum_{j=1}^{m}\sum_{\substack{l=1 \\ l\neq i,k}}^{n}(x_{ij} - x_{lj})(x_{kj} - x_{lj})}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n}(x_{ij} - x_{lj})^2 \cdot \sum_{j=1}^{m}\sum_{l=1}^{n}(x_{kj} - x_{lj})^2\right]^{\frac{1}{2}}}, \tag{17}$$

where: x_{ij}, x_{kj}, x_{lj} are the i-th, k-th and l-th observation of the j-th variable;

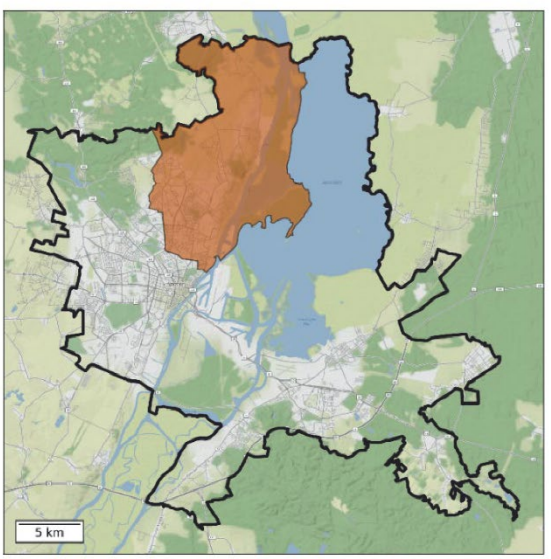
6. a graphical presentation of the results obtained.

All calculations related to the linear ordering were made using the cluster SIM package in the R programme as developed and presented in the work by Walesiak and Dudek (2020).

4. Empirical data

The study used data relating to 405 plots of land in Szczecin, intended for housing purposes, which were located in the northern part of the city (Figure 1). The real estate has been grouped into 17 location attractiveness zones – SALs (Figure 2). The work by Hozer et al. (2019) describes the methodology applied for setting SALs. All properties have been individually valued by property appraisers. The base which consisted of all 405 plots of land was divided into two groups. The first group included 117 properties, which was the learning base. Its composition included real estate representatives from all SALs. The second group, with 288 properties, formed the testing base. The properties on which the algorithm was tested were located in the 13th, 14th and 15th SAL.

Figure 1. The northern region of Szczecin



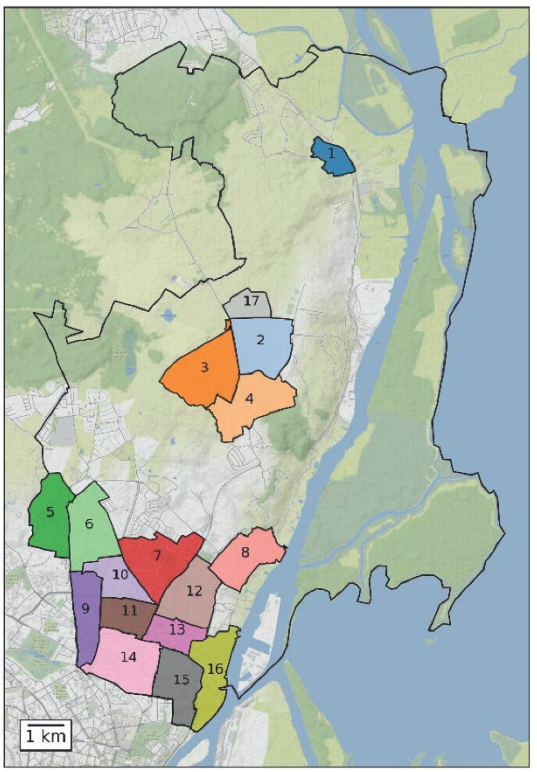
Source: authors' work.

Each real estate was described using the following set of attributes:

- $x1$ is the area: large, average, small;
- $x2$ are the utilities: none, incomplete, full;
- $x3$ is the neighbourhood: troublesome, unfavourable, average, favourable;

- $x4$ is the communication accessibility: unfavourable, average, good;
- $x5$ are the physical features: unfavourable, average, favourable.

Figure 2. The northern region of Szczecin with location attractiveness zones marked

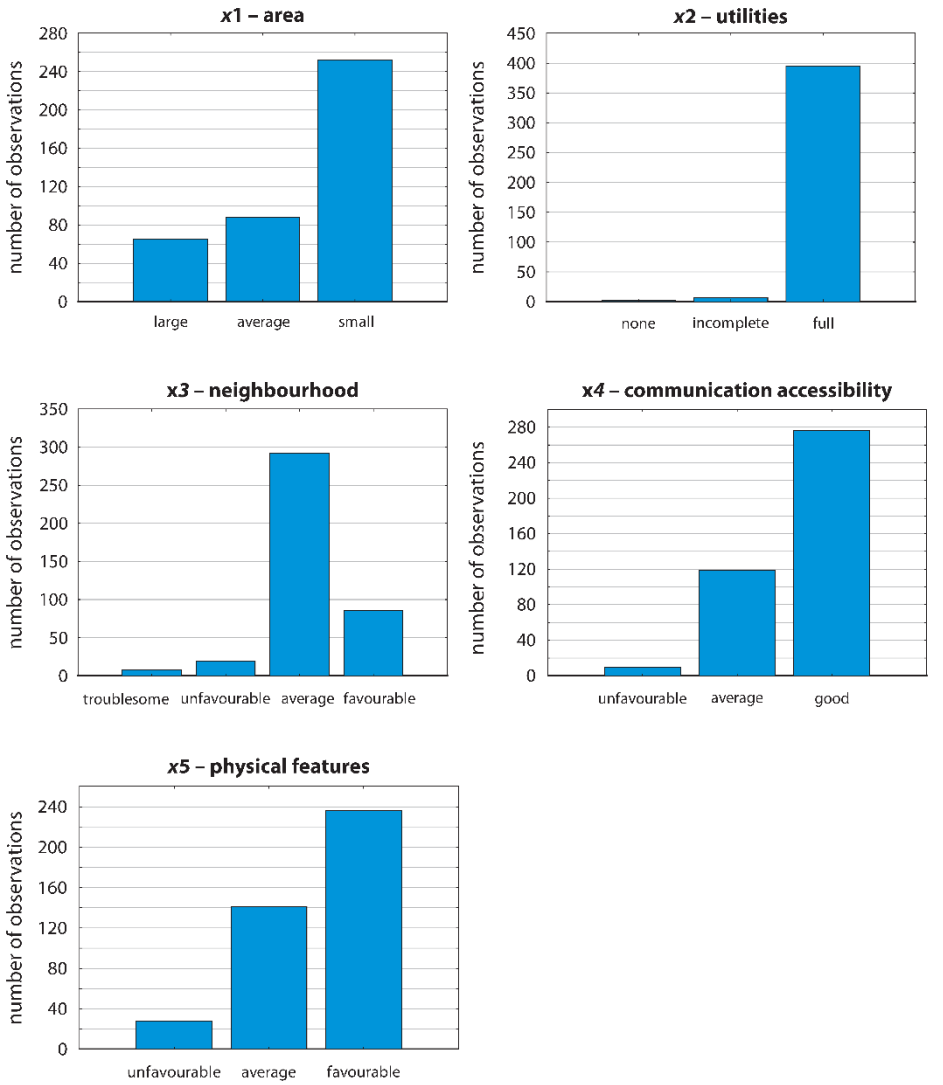


Source: authors' work.

The attributes were presented on ordinal scales, with the value 1 designating the weakest attribute level and subsequent numbers more favourable levels (Figure 3). This method of data encoding should result in a positive correlation between the consecutive attributes and the property value.

Among the properties analysed, those with attributes at the highest levels dominated (Figure 3), except for the neighbourhood attribute, as most properties were situated in average surroundings. Since the analysis concerned land properties located in the city, only a few of them were characterised by none or incomplete infrastructure.

Figure 3. Distributions of individual attribute variants



Source: authors' work.

5. Research results

For the real estate designated as the learning base, seven dependency coefficients between the real estate attributes and their values were calculated. Then, the impact of individual variables on the value of the property $(1 + A_{kpi})$ was determined according to the assumed variants (Table 1) and methods of determining weights

(formulae 4 and 6–10). 168 options were received. For each variant based on formula 3, the hypothetical values of the analysed properties were calculated and the market coefficient values were determined (formula 2). Then, the values calculated in wwr_j were used and based on formula 1, the values of properties located within SAL 13, 14, and 15 (the test base) were estimated. In each case, valuation matching measures were calculated (formulae 11–13 and 15–16). A list of fit measures for the first method of determining weights is presented in Table 2.

Table 2. Measurement errors of real estate valuations according to dependency coefficients for the first method of calculating weights (formula 4)

Dependency coefficients	RMSE	V_{RMSE}	MAPE	B^+	B^-
		in %			
Si.1	495.35	82.53	54.87	3.37	–51.50
Si2.1	495.35	82.53	54.87	3.37	–51.50
Sw.1	303.47	50.56	34.97	3.64	–31.33
Sw2.1	303.47	50.56	34.97	3.64	–31.33
Sci.1	106.30	17.71	14.12	13.62	–0.51
Sci2.1	106.30	17.71	14.12	13.62	–0.51
Scw.1	106.30	17.71	14.12	13.62	–0.51
Scw2.1	106.30	17.71	14.12	13.62	–0.51
Ti.1	495.35	82.53	54.87	3.37	–51.50
Ti2.1	495.35	82.53	54.87	3.37	–51.50
Tw.1	311.42	51.89	35.82	3.62	–32.20
Tw2.1	311.42	51.89	35.82	3.62	–32.20
Tci.1	106.33	17.72	14.12	13.61	–0.51
Tci2.1	106.33	17.72	14.12	13.61	–0.51
Tcw.1	106.33	17.72	14.12	13.61	–0.51
Tcw2.1	106.33	17.72	14.12	13.61	–0.51
Gi.1	495.35	82.53	54.87	3.37	–51.50
Gi2.1	495.35	82.53	54.87	3.37	–51.50
Gw.1	313.49	52.23	36.05	3.62	–32.43
Gw2.1	313.49	52.23	36.05	3.62	–32.43
ri.1	193.63	32.26	21.89	3.47	–18.42
ri2.1	193.63	32.26	21.89	3.47	–18.42
rw.1	174.77	29.12	20.33	3.36	–16.97
rw2.1	174.77	29.12	20.33	3.36	–16.97
rci.1	495.35	82.53	54.87	3.37	–51.50
rci2.1	495.35	82.53	54.87	3.37	–51.50
rcw.1	98.09	16.34	13.90	6.42	–7.48
rcw2.1	98.09	16.34	13.90	6.42	–7.48

Source: authors' work.

In many cases, the variants used have produced the same results, which is a direct result of the number of dependency factors taken into account. The adopted assumption that dependencies on the negative direction are ignored narrowed the number of coefficients used in the subsequent steps. Only one relationship was statistically significant (for all coefficients): between communication accessibility

and the value of the property, and in the case of partial coefficients (for Spearman's, Kendall's and the Gamma coefficient) two relationships: between the utilities and surroundings and the value of the property. According to the Pearson coefficient, the relationship between utilities, transport accessibility, the environment and the value of the property was statistically significant (and for the partial coefficient, all relationships were statistically significant).

From the point of view of the results obtained, it was irrelevant to include the squares of the relationship between the attributes and the value of the property in the further calculations, as the results obtained for both variants were the same. This was the case for Kendall and Spearman's partial coefficients: the fit measures were the same for all four options. The further calculations omitted 54 variants that generated the same results.

The proposed measures of errors between the real values of the real estate and the values estimated based on SAMWN showed a wide variation resulting from the approach used. The basic descriptive statistics of the received errors are summarised in Table 3.

Table 3. Descriptive statistics of the valuation errors

Selected descriptive statistics	RMSE	V_{RMSE}	MAPE	B^+	B^-
		in %			
Minimum	46.56	7.76	6.23	1.99	–161.07
Maximum	1753.73	292.19	165.32	17.50	–0.51
Arithmetic mean	313.35	52.21	33.79	5.70	–28.09
Standard deviation	378.77	63.11	36.16	4.37	37.08
Variation coefficient	120.88	120.88	107.01	76.64	132.00
Median	133.53	22.25	18.07	4.20	–9.07
Semi-interquartile range	187.38	31.22	18.69	1.42	18.92
Positional variation coefficient	140.33	140.33	103.43	33.70	208.72

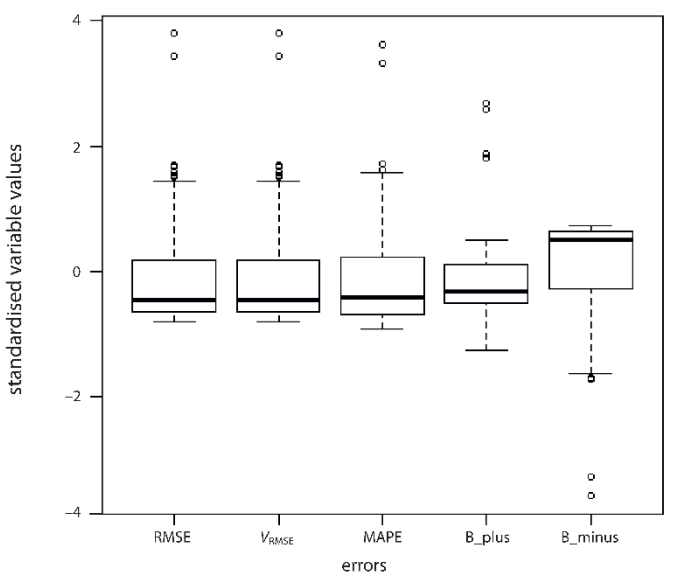
Source: authors' work.

The most decent level of the analysed errors was as close as possible to 0, which would indicate no differences between the value estimated by property appraisers and the value estimated by SAMWN. The range of the obtained results was very large (Table 3). The smallest observed MAPE value was 6.23%, while the highest was 165.32%. The coefficient of variation for this measure was over 100% (also in a narrowed area of variation). Moreover, RMSE and V_{RMSE} errors were characterised by very high variability (over 120% and 140% in the narrowed area of variability). The B^+ error indicates in what percentage the values estimated by the model are higher than the actual values. Similarly, the B^- error provides information about the values estimated by the model below the actual values. It turns out that the used variants caused a significant shift in the obtained results in plus or in minus (the

range between the errors for individual variants extended from 6% to even 165%). This is not a desirable occurrence, because after all, the use of SAMWN in mass valuation aims to achieve results close to the real values, calculated in individual appraisals. Therefore, when selecting the best variant of the algorithm, not only the valuation errors related to the differences between the actual and estimated value (formulae 11–13) should be taken into account, but also the errors indicating how much the model underestimates or overestimates the value of the property (formulae 15–16).

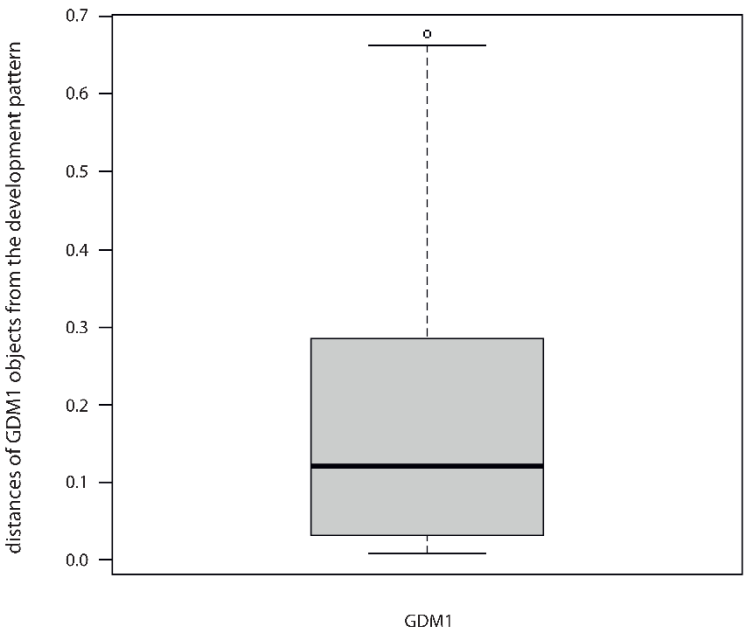
The procedure of linear ordering of the used dependency coefficients according to the received valuation errors was carried out. All variables were characterised by very high variability (in most cases the variation coefficient exceeded 100%). To achieve comparable variables, standardised variables were derived. The first four variables were designated as destimulants whose smaller value is desirable, and B^- as a stimulant. The reference level for all variables was 0, i.e. the most desirable level of the selected variables. Then distance measure GDM1 (formula 17) and the upper development pattern were used. Figure 4 presents the evolution of the variables after standardisation, while Figure 5 shows the diversity of the synthetic variable calculated using the GDM1 distance measure.

Figure 4. Error differentiation after standardisation



Source: authors' calculations.

Figure 5. Differentiation of a synthetic (aggregate) variable covering GDM1 distances from the development pattern for all SAMWN estimation variants

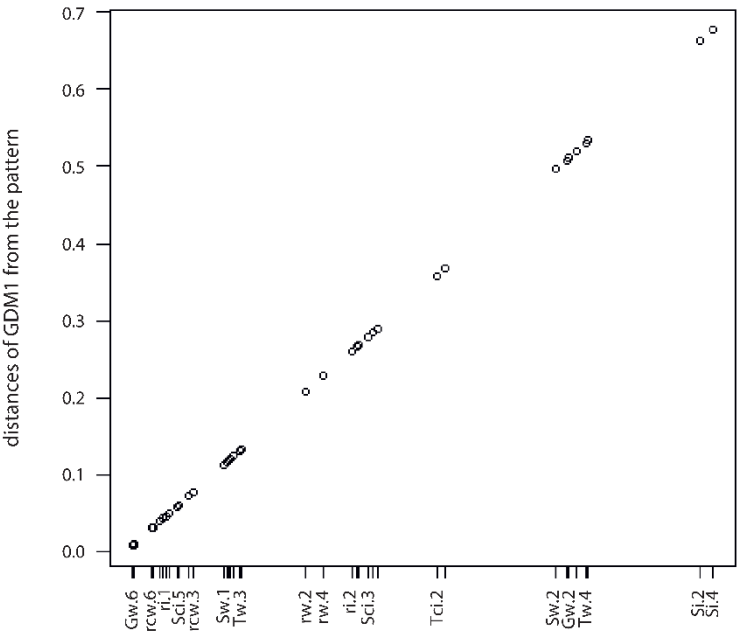


Source: authors' calculations.

Distributions of valuation errors indicated strong right-hand asymmetrical distributions for RMSE, V_{RMSE}, and the MAPE variables (Figure 4). Atypical values occurred, which indicated that in some of the proposed variants the obtained valuations significantly differed from the actual values. The valuation error deviating in plus was also distinguished by right-hand asymmetry, but proved weaker than in the case of the previous errors. Unusual values were also observed in this distribution. The left-hand asymmetry was characterised by the B^- error distribution, which is related to the nature of this variable being a stimulant. The aggregate variable calculated based on GDM1 distances showed right-hand asymmetrical distribution, but with no atypical values (Figure 5).

Figure 6 presents the results of linear ordering and Table 4 lists the three best and three worst (according to GDM1) valuation options using appropriate relationship factors.

Figure 6. Ordering of the dependence coefficients from the best to the worst due to error valuations obtained using SAMWN



Source: authors' calculations.

Table 4. Ordering of the dependence coefficients from the best to the worst according to the GDM1 measure value

Dependence coefficients	GDM1 measure	Dependence coefficients	GDM1 measure
Best		**Worst**	
Gw.6	0.008	Gw.4	0.534
Tw.6	0.008	Si.2	0.662
Sw.6	0.008	Si.4	0.677

Source: authors' calculations.

The least distant from the standard (representing the best results) were cases calculated based on the Gamma, Kendall and Spearman coefficients and taking into account all measures of dependence at the 6th weighing method (results differed from each other at the 4th and 5th decimal place). The furthest from the standard (representing the worst results) were obtained for cases where the weighing was based on the median, for the Spearman coefficients and taking into account only significant dependencies (only one relationship), or the Gamma coefficient (all dependencies, weighing based on the median).

In the presented study, the best results were obtained for weighing methods based on a coefficient of variation (formulae 9–10), and then for balances constructed with an arithmetic mean (formula 4 and 7). The worst results concerned the 3rd and 4th weighing method (formula 6 and 8), which were associated with the use of the median.

6. Discussion

The use of automated algorithms in the process of property valuation still raises controversy in Poland. According to the Polish law, only real estate appraisers are authorised to issue opinions on property values. However, there are attempts to apply advanced computational procedures in the valuation of real estate, especially in the context of the held discussions on changing the basis for calculating property tax from surface area (as it currently is) to its value. If such a change were to occur, proper tools (computational algorithms) would be required to allow for the mass valuation of various types of properties. The Szczecin algorithm for mass valuation is such a tool.

The SAMWN has already been successfully applied in business practice. During the process of valuating properties based on SAMWN, certain problems were identified which required improvement. One of them was the automation of the process of determining market value coefficients *wwr*, which was presented in the article. The use of SAMWN in real estate valuation is, of course, limited by certain conditions.

Obtaining accurate valuation results requires a well-prepared database containing information about the valued properties (including their attributes). An important aspect of creating such a database is to ensure the comparability of the levels of attributes describing individual properties so that e.g. an 'average' neighbourhood is understood in the same way for each property. Next, the elementary units (SALs) for which market value coefficients *wwr* are calculated must be determined. The proper selection of representative properties for the training sample is a very important stage where the representative properties must have all attribute states. In addition, the sample should be sufficiently large to ensure the stability of the calculated dependencies. The representatives should be valued individually by authorised specialists, i.e. property appraisers.

The article systematises the coefficients of dependencies used to determine the impact of individual attributes on the value of a real estate. Additionally, the best method for weighing the impact of individual attribute states was proposed. The advantage of the presented method is that the algorithm does not require additional assumptions regarding the attributes, such as an appropriate type of distribution.

Additionally, SAMWN takes into account various attributes of real estate (which can be adjusted depending on the type of the real estate), including the mode. The proposed algorithm is universal and can be used to value different types of real estate.

7. Conclusions

Estimating the value of real estate in mass can be based on calculation algorithms. In addition, statistical methods are used in the mass valuation process. One of the proposals for mass valuation is the Szczecin Mass Valuation Algorithm. One of the aims of the article was to present the possibility of using various dependency coefficients to determine the impact of the attributes on the value of the real estate. The use of dependency coefficients was proposed, dedicated to ordinal variables (real estate attributes were presented on ordinal scales) as was Pearson's correlation coefficient due to its high popularity among property appraisers.

Various methods were used to consider dependency coefficients in the subsequent property valuation procedure: only significant dependencies or their squares were taken into account, as were all dependencies or their squares. The study showed that the consideration of the coefficients or their squares did not change the scale of the valuation errors.

Finally, 54 variants of property valuation using SAMWN were further analysed. Linear ordering methods were used to receive valuation errors in order to select the best option, which was the second aim of the article. As a result of the used ordering procedure, it turned out that the choice of the coefficient was not as significant (although the use of the Person coefficient did not produce the best results, nor did it give the worst). The valuation results were mainly influenced by the formula used to calculate the impact weights of individual attributes on the value of the real estate in SAMWN.

The analysis shows that the best combination is the use of one of the coefficients: Gamma, Kendall, or Spearman. Additionally, all coefficients greater than 0 between the attributes and the value of the property should be taken into account, as well as the use of the formula no. 6 (designated in formula 10) for calculating the weights: the ratio of the variation coefficient of the property value with the best states of the attributes to the variation coefficient of the property value with the weakest attribute states, corrected by the $\ln\left(\frac{w_{\max}}{w_{\min}}\right)$ coefficient. The weakest results occurred when the median was considered in the method of calculating the weights.

References

Abidoye, R. B., & Chan, A. P. C. (2018). Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal*, *24*(1), 71–83. https://doi.org/10.1080/14445921.2018.1436306.

Ahn, J. J., Byun, H. W., Oh, K. J., & Kim, T. Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, *39*(9), 8369–8379. https://doi.org/10.1016/J.ESWA.2012.01.183.

Alcantud, J. C. R., Rambaud, S. C., & Torrecillas, M. J. M. (2017). Valuation Fuzzy Soft Sets: A Flexible Fuzzy Soft Set Based Decision Making Procedure for the Valuation of Assets. *Symmetry*, *9*(11), 1–19. https://doi.org/10.3390/SYM9110253.

Amidu, A.-R., & Boyd, D. (2018). Expert problem solving practice in commercial property valuation: an exploratory study. *Journal of Property Investment & Finance*, *36*(4), 366–382. https://doi.org/10.1108/JPIF-05-2017-0037.

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*(3), 247–267. https://doi.org/10.1111/J.1574-0862.2002.TB00120.X.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778. https://doi.org/10.1016/J.ESWA.2011.08.077.

Bellotti, A. (2017). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, *81*(1–2), 71–84. https://doi.org/10.1007/S10472-016-9534-6.

Calka, B. (2019). Estimating Residential Property Values on the Basis of Clustering and Geostatistics. *Geosciences*, *9*(3), 1–14. https://doi.org/10.3390/GEOSCIENCES9030143.

Calka, B., & Bielecka, E. (2016). The Application of Geoinformation Theory in Housing Mass Appraisal. In *2016 Baltic Geodetic Congress (BGC Geomatics)* (pp. 239–243). IEEE. https://doi.org/10.1109/BGC.GEOMATICS.2016.50.

Cervelló-Royo, R., Guijarro, F., Pfahler, T., & Preuss, M. (2016). An Analytic Hierarchy Process (AHP) framework for property valuation to identify the ideal 2050 portfolio mixes in EU-27 countries with shrinking populations. *Quality & Quantity*, *50*(5), 2313–2329. https://doi.org/10.1007/s11135-015-0264-3.

Chen, J.-H., Ong, C. F., Zheng, L., & Hsu, S.-C. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management*, *21*(3), 273–283. https://doi.org/10.3846/1648715X.2016.1259190.

Chen, Z., Hu, Y., Zhang, C. J., & Liu, Y. (2017). An Optimal Rubrics-Based Approach to Real Estate Appraisal. *Sustainability*, *9*(6), 1–19. https://doi.org/10.3390/SU9060909.

Dimopoulos, T., & Moulas, A. (2016). A Proposal of a Mass Appraisal System in Greece with CAMA System: Evaluating GWR and MRA techniques in Thessaloniki Municipality. *Open Geosciences*, *8*(1), 675–693. https://doi.org/10.1515/GEO-2016-0064.

Dmytrów, K., Gdakowicz, A., & Putek-Szeląg, E. (2020). Methods of analyzing qualitative variable correlation on the real estate market. *Real Estate Management and Valuation*, *28*(1), 80–90. https://doi.org/10.2478/REMAV-2020-0007.

Doszyń, M. (Ed.). (2020). *System kalibracji macierzy wpływu atrybutów w szczecińskim algorytmie masowej wyceny nieruchomości.* Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.

Gabrielli, L., Giuffrida, S., & Trovato, M. R. (2017). Gaps and overlaps of urban housing sub-market: Hard clustering and fuzzy clustering approaches. In S. Stanghellini, P. Morano, M. Bottero & A. Oppio (Eds.), *Appraisal: From Theory to Practice* (pp. 203–219). Springer. https://doi.org/10.1007/978-3-319-49676-4_15.

Gatnar, E., & Walesiak, M. (Eds.). (2011). *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Wydawnictwo C.H. Beck.

Gdakowicz, A., & Putek-Szelag, E. (2020a). Is It Possible to Overfit the Algorithm? Case Study of Mass Valuation of Land Properties in Szczecin. *European Research Studies Journal*, *23*(SI2), 110–122. https://doi.org/10.35808/ERSJ/1812.

Gdakowicz, A., & Putek-Szeląg, E. (2020b). The Use of Statistical Methods for Determining Attribute Weights and the Influence of Attributes on Property Value. *Real Estate Management and Valuation*, *28*(4), 33–47. https://doi.org/10.1515/REMAV-2020-0030.

Glennon, D., Kiefer, H., & Mayock, T. (2018). Measurement error in residential property valuation: An application of forecast combination. *Journal of Housing Economics*, *41*, 1–29. https://doi.org/10.1016/J.JHE.2018.02.002.

Guan, J., Shi, D., Zurada, J. M., & Levitan, A. S. (2014). Analyzing Massive Data Sets: An Adaptive Fuzzy Neural Approach for Prediction, with a Real Estate Illustration. *Journal of Organizational Computing and Electronic Commerce*, *24*(1), 94–112. https://doi.org/10.1080/10919392.2014.866505.

Han, L., & Zhu, J. (2008). Using matrix of thresholding partial correlation coefficients to infer regulatory network. *Biosystems*, *91*(1), 158–165. https://doi.org/10.1016/J.BIOSYSTEMS.2007.08.008.

Hellwig, Z. (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. *Przegląd Statystyczny*, *15*(4), 307–327.

Hozer, J., Gnat, S., Kokot, S., & Kuźmiński, W. (2019). The Problem of Designating Elementary Terrains for the Purpose of Szczecin Algorithm of Real Estate Mass Appraisal. *Real Estate Management and Valuation*, *27*(3), 42–58. https://doi.org/10.2478/REMAV-2019-0024.

Hozer, J., Kokot, S., & Kuźmiński, W. (2002). *Metody analizy statystycznej rynku w wycenie nieruchomości*. Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych.

Hui, S. K., Cheung, A., & Pang, J. (2010). A Hierarchical Bayesian Approach for Residential Property Valuation: Application to Hong Kong Housing Market. *International Real Estate Review*, *13*(1), 1–29. https://ideas.repec.org/a/ire/issued/v13n012010p1-29.html.

International Association of Assessing Officers. (2019). *Standard on Mass Appraisal of Real Property. A criterion for measuring fairness, quality, equity and accuracy*. https://www.iaao.org/media/standards/StandardOnMassAppraisal.pdf.

Jahanshiri, E., Buyong, T., & Shariff, A. R. M. (2011). A Review of Property Mass Valuation Models. *Pertanika Journal of Science & Technology*, *19*(*special issue*), 23–30. https://core.ac.uk/download/pdf/153832642.pdf#page=34.

Kendall, M. G. (1948). *Rank correlation methods* (4th edition). Charles Griffin & Company Limited.

Kilpatrick, J. (2011). Expert systems and mass appraisal. *Journal of Property Investment and Finance*, *29*(4/5), 529–550. https://doi.org/10.1108/14635781111150385.

Kuryj, J. (2007). Metodyka wyceny masowej nieruchomości na bazie aktualnych przepisów prawnych. *Wycena. Wartość, Obrót, Zarządzanie Nieruchomościami*, (4), 50–58.

Lam, K. C., Yu, C. Y., & Lam, C. K. (2009). Support vector machine and entropy based decision support system for property valuation. *Journal of Property Research*, *26*(3), 213–233. https://doi.org/10.1080/09599911003669674.

Lasota, T., Telec, Z., Trawiński, B., & Trawiński, K. (2011). Investigation of the eTS Evolving Fuzzy Systems Applied to Real Estate Appraisal. *Journal of Multiple-Valued Logic and Soft Computing*, *17*(2), 229–253.

Lin, C. C., & Mohan, S. B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis*, *4*(3), 224–243. https://doi.org/10.1108/17538271111153013.

Lockwood, T., & Rossini, P. (2011). Efficacy in modelling location within the mass appraisal process. *Pacific Rim Property Research Journal*, *17*(3), 418–442. https://doi.org/10.1080/14445921.2011.11104335.

Ma, X., Zeng, D., Wang, R., Gao, J., & Qin, B. (2018). An intelligent price-appraisal algorithm based on grey correlation and fuzzy mathematics. *Journal of Intelligent and Fuzzy Systems*, *35*(3), 2943–2950. https://doi.org/10.3233/JIFS-169650.

McCluskey, W. J., & Borst, R. A. (2011). Detecting and validating residential housing submarkets: A geostatistical approach for use in mass appraisal. *International Journal of Housing Markets and Analysis*, *4*(3), 290–318. https://doi.org/10.1108/17538271111153040.

McCluskey, W. J., Daud, D. Z., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction*, *19*(2), 152–167. https://doi.org/10.1108/JFMPC-06-2013-0022.

Morano, P., Tajani, F., & Locurcio, M. (2018). Multicriteria analysis and genetic algorithms for mass appraisals in the Italian property market. *International Journal of Housing Markets and Analysis*, *11*(2), 229–262. https://doi.org/10.1108/IJHMA-04-2017-0034.

Napoli, G., Giuffrida, S., & Valenti, A. (2017). Forms and Functions of the Real Estate Market of Palermo (Italy). Science and Knowledge in the Cluster Analysis Approach. In S. Stanghellini, P. Morano, M. Bottero & A. Oppio (Eds.), *Appraisal: From Theory to Practice* (pp. 191–202). Springer. https://doi.org/10.1007/978-3-319-49676-4_14.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining Nonoverlap and Trend for Single-Case Research: Tau-U. *Behavior Therapy*, *42*(2), 284–299. https://doi.org/10.1016/J.BETH.2010.08.006.

Pawlukowicz, R. (2010). Wykorzystanie metodyki porządkowania liniowego do określania wartości rynkowej nieruchomości. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, *17*(107), 377–385.

Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych. (2008). *Powszechne Krajowe Zasady Wyceny (PKZW) – nota interpretacyjna. Zastosowanie podejścia porównawczego w wycenie nieruchomości*. http://wycena.net.pl/standardy/NI-Zastosowanie.podejscia.porownawczego.pdf.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd edition). McGraw-Hill. http://www.sciepub.com/reference/20400.

Strahl, D. (1978). Propozycja konstrukcji miary syntetycznej. *Przegląd Statystyczny*, *25*(2), 205–215.

Telega, T., Bojar, Z., & Adamczewski, Z. (2002). Wytyczne przeprowadzania powszechnej taksacji nieruchomości (projekt). *Przegląd Geodezyjny*, *74*(6), 6–11.

Walesiak, M. (2016). *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R* (2nd edition). Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Walesiak, M., & Dudek, A. (2020). The Choice of Variable Normalization Method in Cluster Analysis. In K. S. Soliman (Ed.), *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges* (pp. 325–340). International Business Information Management Association.

Wang, D., & Li, V. J. (2019). Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. *Sustainability*, *11*(24), 1–14. https://doi.org/10.3390/SU11247006.

Wu, C., Ye, X., Ren, F., & Du, Q. (2018). Modified Data-Driven Framework for Housing Market Segmentation. *Journal of Urban Planning and Development*, *144*(4), 1–15. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000473.

Yacim, J. A., & Boshoff, D. G. B. (2018). Combining BP with PSO algorithms in weights optimisation and ANNs training for mass appraisal of properties. *International Journal of Housing Markets and Analysis*, *11*(2), 290–314. https://doi.org/10.1108/IJHMA-02-2017-0021.

Zeliaś, A. (2002). Some Notes on the Selection of Normalization of Diagnostic Variables. *Statistics in Transition*, *5*(5), 787–802.

Zhang, R., Du, Q., Geng, J., Liu, B., & Huang, Y. (2015). An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. *Habitat International*, *46*, 196–205. https://doi.org/10.1016/J.HABITATINT.2014.12.001.

Zhou, G., Ji, Y., Chen, X., & Zhang, F. (2018). Artificial Neural Networks and the Mass Appraisal of Real Estate. *International Journal of Online and Biomedical Engineering*, *14*(3), 180–187. https://doi.org/10.3991/IJOE.V14I03.8420.

Zurada, J., Levitan, A. S., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, *33*(3), 349–387. https://doi.org/10.1080/10835547.2011.12091311.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 49–70 https://doi.org/10.59139/ws.2023.10.3
Spisy powszechne – problemy i wyzwania / Issues and challenges in census taking
Zgłoszony/submitted: 5.04.2023, poprawiony/revised: 3.08.2023, zaakceptowany/accepted: 16.08.2023

Digital population and housing census – the experience of Serbia

Miladin Kovačević,[a] Mira Nikić,[b] Branko Josipović,[c] Snežana Lakčević,[d] Vesna Pantelić,[e] Nevena Mitrović,[f] Adil Kolaković,[g] Petar Korović[h]

Abstract. The aim of the paper is to present the experience of the Republic of Serbia in conducting the 2022 Census of Population, Households and Dwellings, focusing on the employment, legal framework and financing of the census as well as on its successful implementation. It discusses strategic decisions on data collection and the integration of information technology – including geospatial data, data collection techniques, machine learning, record linkage and monitoring system – to overcome the challenges posed by the census. The paper addresses the census undercoverage, explores the use of administrative data for item imputation, and examines the development of a statistical population register. The study demonstrates the benefits of adopting a digital-census approach: significant improvement of accuracy, cost reduction and acquired expeditiousness.

The Statistical Office of the Republic of Serbia conducted a digital census combined with traditional methods, excluding self-enumeration, along with the use of administrative data for item imputation, and recommends this approach as the most effective way to obtain precise and comprehensive information about a population, including its demographic characteristics, geographic distribution and overall size.

Keywords: 2022 Census of Population, Households and Dwellings, digital census, geospatial data, monitoring system, machine learning, administrative data, record linkage, imputation, statistical population register, Serbia

JEL: J18, M15, N34

[a] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-2398-7955. E-mail: miladin.kovacevic@stat.gov.rs.
[b] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0003-2340-8789. Autor korespondencyjny / Corresponding author, e-mail: mira.nikic@stat.gov.rs.
[c] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0009-3191-5784. E-mail: branko.josipovic@stat.gov.rs.
[d] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-8106-8412. E-mail: snezana.lakcevic@stat.gov.rs.
[e] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0002-6430-9288. E-mail: vesna.pantelic@stat.gov.rs.
[f] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0000-2730-8011. E-mail: nevena.mitrovic@stat.gov.rs.
[g] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0008-2553-3707. E-mail: adil.kolakovic@stat.gov.rs.
[h] Statistical Office of the Republic of Serbia, Serbia. ORCID: https://orcid.org/0009-0004-5761-8554. E-mail: petar.korovic@stat.gov.rs.

Cyfrowy powszechny spis ludności i mieszkań – przykład Serbii

Streszczenie. Celem artykułu jest przedstawienie doświadczeń Republiki Serbii w zakresie organizacji Powszechnego Spisu Ludności, Gospodarstw Domowych i Mieszkań 2022, ze szczególnym uwzględnieniem zagadnień dotyczących zatrudnienia personelu, ram prawnych i finansowania tego badania oraz warunków jego udanej realizacji. Praca skupia się na strategicznych decyzjach w sprawie zbierania danych oraz zastosowania technik informatycznych, takich jak: wykorzystanie danych przestrzennych, cyfrowe metody uzyskiwania danych, uczenie maszynowe, łączenie rekordów czy system monitorujący, mających na celu sprostanie wyzwaniom związanym ze spisem. Autorzy poruszają także kwestie niedostatecznego pokrycia spisu oraz wykorzystania rejestrów administracyjnych do imputacji danych. Ponadto poświęcają uwagę opracowaniu i udoskonalaniu statystycznej ewidencji ludności, dokładności danych, obniżeniu kosztów i zwiększeniu efektywności badania.

Główny Urząd Statystyczny Republiki Serbii przeprowadził spis powszechny w sposób cyfrowy, łącząc ten mechanizm z metodami tradycyjnymi (z wyłączeniem samospisu) i posiłkując się rejestrami administracyjnymi w celu imputacji danych. Metoda ta jest w artykule rekomendowana jako najefektywniejszy sposób uzyskania precyzyjnych i wyczerpujących informacji na temat populacji, w tym jej charakterystyki demograficznej, rozmieszczenia przestrzennego i liczebności.

Słowa kluczowe: Powszechny Spis Ludności, Gospodarstw Domowych i Mieszkań 2022, spis cyfrowy, dane geoprzestrzenne, system monitorujący, uczenie maszynowe, dane administracyjne, łączenie rekordów, imputacja, statystyczna ewidencja ludności, Serbia

1. Introduction

The Serbian census has a long history dating back to 1834 with nearly five-year intervals until the First World War. The 1866 Census was the first modern and complete survey of this kind. Between the world wars, only two censuses were conducted. After the Second World War, a census was urgently conducted in 1948 to collect data on damages caused by the war. Starting from 1961, a ten-year periodicity for censuses was established. Thus, the subsequent censuses were conducted in 1961, 1971, 1981 and 1991. The census of 2002 was delayed due to financial constraints and the authorities in Montenegro requesting postponement. The 2011 Census returned to the established periodicity, and the latest survey of this kind was conducted in October 2022, having been postponed due to the COVID-19 pandemic. According to the most recent results, the population of the Republic of Serbia totalled 6,647,003 usual residents, 2,589,344 households and 3,613,352 dwellings.

Serbia's censuses have been regulated by special legal provisions since 1884. Nowadays, as an EU candidate, the country follows the EU census legislation, which

upholds diversity and stipulates that countries should define the content of their census legislation in such a way that it is compliant with the national legal practices and procedures. The Law on the 2021 Census of Population, Households and Dwellings regulates the preparation, financing, organisation and implementation of the Census in the Republic of Serbia. It primarily determines the period and method of the census implementation, with the Statistical Office of the Republic of Serbia (SORS) playing the key role in the process. It also specifies the tasks of census commissions, formed for the purpose of conducting the survey, as well as the competences and obligations of ministries and other governmental bodies and organisations involved in the preparation and implementation of the survey. The law moreover stipulates the obligations of the participants, persons who directly perform duties related to the census, as well as the use and protection of the census data, the publication of the results and the financing of the survey.

The SORS went through all the procedural steps for the adoption of the census law (the public debate, the analysis of the probable effects of the regulation, collecting opinions of the relevant state authorities, etc.). Also, the census law was enhanced and aligned with the Law on Personal Data Protection, which ensured that all the aspects of the preparation and implementation of the census were acceptable to the public.

For the first time in history, the text of the law together with the Table of Concordance (where the concrete articles were compared to the relevant articles from the EU legislation) were submitted to the European Commission. In response, the European Commission suggested certain methodological improvements, but in general, the law was assessed as compliant with European legislation.

In April 2021, the above-mentioned law was amended by the Law on Amendments to the Law on the Census of Population, Households and Dwellings, which sanctioned the postponement of the census due to the COVID-19 pandemic.

The aim of the paper is to present the experience of the Republic of Serbia in conducting the 2022 Census of Population, Households and Dwellings, focusing on the employment, legal framework and financing of the census as well as on its successful implementation. It examines the undertaken strategic decisions regarding data collection and the integration of information technology (such as geospatial data, data-collection techniques, machine learning and the monitoring system), emphasising their role in addressing the challenges posed by the census. The paper also addresses the issue of undercoverage and explores the use of administrative data for item imputation. Future-oriented attitude to the topic in question adopted in the paper invites discussion on the development of a statistical population register.

The overarching objective of the paper is to highlight how the digital-census approach, in combination with the use of administrative data for item imputation (without self-enumeration), can lead to the improved accuracy, reduced costs and a faster process of data collection, processing, and dissemination.

2. Financing and implementation of the Population Census in Serbia

The population census was a part of the 'EU for Development Statistics in Serbia' project, funded by the EU and the Republic of Serbia through the Pre-Accession Assistance (IPA) 2018 national programme. The grant contract was signed on 29th July 2019. The cost of the census was estimated at approximately 21.2 million euro excluding IT equipment, and 29.6 million euro including it. The government of Serbia committed itself to providing co-financing (32.77%) and funds for the purchase of the IT equipment. The cost of the census, compared with 2011 census, increased due to changes in tax laws. They resulted in the additional cost of 4.6 million euro for taxes and contributions for enumerators and trainers, and 5.5 million euro for all externally-engaged field workers.

The budget for the Population Census was planned in 2017, with an average enumerator salary of 45,000 RSD (370 EUR), equivalent to the average monthly earnings in Serbia. The inflation rate grew significantly between the previous census and 2021, i.e. by 139.2% in the period from 2012 to 2021.

The SORS decided to postpone the Population Census due to the COVID-19 pandemic which posed a risk to the preliminary activities and fieldwork operations. Field activities were initially postponed for six months and then for an additional year, so finally the census was scheduled for October 2022 (instead of April 2021). This delay necessitated additional expenses that could not have been initially planned for.

Due to the COVID-19 pandemic, additional funds were required for the census-related activities, as well as for the delayed field work. The European Commission rejected the request for the increase of the funds, and consequently, the Serbian Ministry of Finance was asked to provide additional resources for remuneration of the enumerators and instructors. The project budget did not account for inflation, which resulted in lower remuneration for people who directly served the census. This could have led to difficulties in finding qualified staff, as 18,000 people to be recruited needed to be computer-literate to participate in the census.

The SORS received confirmation from the Ministry of Finance in March 2022 that they would provide the necessary funds, meeting an important prerequisite for the successful census. The additional funds increased the share of the Republic of Serbia to 44.83%, while the EU grant amount remained the same, which effectively

decreased the EU share to 55.17%. A detailed Financial Instructions document set out the principles for the use of funds, and the planned financial activities proceeded without any significant problems.

3. Strategic decisions regarding data collection in the 2022 Census

After the completion of the 2011 Census, conducted by means of paper questionnaire and OCR (optical character recognition) technology used to scan paper questionnaires and input the data into a database, the SORS began examining new approaches to a traditional census. The basic aim was to explore the possibilities for reducing costs, minimising burden for respondents, and improving the availability of more up-to-date estimates and surveys in the interim period.

Following international trends and examples of good practices, the possibility and suitability of using administrative sources for the census purposes were analysed in detail. It was concluded that there were still no conditions for the implementation of a fully register-based census. Considering that different registers are continuously developing, some data can be used in different phases of the preparation and implementation of the census (for defining boundaries and creating maps of enumeration areas, controlling the coverage of individual contingents, imputations, etc.).

Activities connected to the implementation of the 2019 Pilot Census were aimed at testing all methodological, technical and IT solutions. Apart from using two new data collection techniques, i.e. computer-assisted interviews (CAPI) and self-enumeration via Internet (CAWI), the monitoring system as well as the methods for the collection and linkage of census and geospatial data were examined. Although the phases of preparation and implementation are considered to have been successfully done, the response rate for self-enumeration via Internet was very low and the quality of the obtained data was questionable. Therefore, it was decided that self-enumeration via Internet should not be used as a method of collecting data in the 2022 Census.

After an additional analysis of all the pros and cons of certain methods of the census implementation, it was concluded that the transition to a census based on administrative sources would be a complex and long process. For this reason, the 2022 Census had to be a traditional one, yet it was modernised by using the CAPI method, i.e. laptops for face-to-face interviews. The application of the above-mentioned method is considered a significant modernisation of the census processes; it increases data accuracy and speeds up data processing. It also significantly reduces the number of the engaged enumerators and makes it possible

to monitor the census activities in real time, as well as starting the dissemination of the final results faster than while using the traditional method of data collection.

Following the announcement of the COVID-19 pandemic, the SORS faced new challenges related to conducting the census and developing actions or options for reducing the impact of the epidemic. Primarily, the census field implementation was postponed from April 2021 to October 2022. Afterwards, the SORS focused on developing various scenarios, such as: the potential adjustments of census questionnaires, the adoption of additional data collection method, and changing methods and plans for the enumeration of special population groups.

Given that some groups of the population, e.g. those living in special institutions[1], were identified as potentially exposed to a higher risk of the COVID-19 infection, their enumeration was carried out by the employees of this institutions through a web application (CAWI).

Apart from CAPI, in the cases where face-to-face interview was not possible due to the unfavourable epidemiological situation, enumeration was performed via telephone (CATI). Likewise, the census activities were redesigned to respond to the effects of the pandemic and its impact on the quality of the census output.

4. Geospatial technology in support of the census operations

Geospatial technology during the enumeration was a fundamental component of the Serbian Census 2022. The SORS implemented recommendations from the UN Guidelines concerning the use of electronic data collection technologies in population and housing censuses (United Nations, 2019). In recent years, the Geodetic Authority of the Republic of Serbia has developed and maintained a sophisticated national geospatial database that consists of locations of housing units, addresses, administrative boundaries, statistical boundaries, and a digital map of all the streets in the country. In preparation for the census, the Geodetic Authority provided the SORS with a digital orthophoto produced from the aerial photogrammetric data, and a national geospatial database with the aforementioned variables.

During the census, geospatial technology was used to support field operations by providing field maps and address lists. Prior to the census, it was used to design and digitise enumeration areas, establish an address-coding system, predetermine coverage areas and prepare maps. During the enumeration, geospatial technology was also used to evaluate data coverage and quality.

[1] Institutions for the execution of criminal sanctions (prisons, correctional institutions), social welfare institutions (homes for the elderly, children with disabilities, etc.), special hospitals for psychiatric diseases.

5. Data collection and transfer

5.1. Considerations for selecting hand-held devices

The selection of hand-held devices for data collection in the Census 2022 was based on three key parameters: operating system selection, device characteristics and price. The decision to use a specific operating system (OS) was made after considering various factors such as the IT infrastructure, business applications, security, human resources and the solution already developed and implemented at the SORS for the CAPI surveys. Under the chosen OS, three types of devices could be used: a laptop, a '2 in 1' device (a portable computer that had features of both a tablet and a laptop), and a tablet. The choice of device was based on several factors such as the availability, usability in 2021 and beyond, price within the budget, resistance to physical damage, screen size, weight, resistance to market fluctuations, storage, keyboard, portability, processing power and external ports. A laptop with the chosen OS was selected, as it fitted the procurement budget and met all the necessary requirements. The SORS acquired 15,500 laptops with similar configurations and weight less than 1.5 kg a piece. In addition, the SORS insisted on a service-contracted agreement with the vendors for defective laptop replacements within 24 hours, and leased warehouses for storing and preparing equipment for use and distribution.

5.2. Data collection using laptops (CAPI)

The SORS developed its own metadata-based data integration platform (IST), which offered various data collection methods, including CATI, CAPI and CAWI. The IST platform used a relational database format for data storage (LocalDB for CAPI and MS SQL Server or SQL as service for other data-collection modules) and encryption to ensure data security. The IST platform allowed enumerators to work offline on laptops and to synchronise data when online.

During the census, the IST platform organised the questionnaire into four segments: building, dwelling, household and personal information. Enumerators received through their laptops lists of addresses that they had to visit. The application directed the enumerator through each part of the questionnaire during data entry and prevented them from skipping any segments. However, in the data-editing phase, they were allowed to skip segments to immediately approach preferred fields rather than navigating through all the questionnaires. The IST platform was non-linear, which allowed enumerators to directly enter the dwelling or personal information sections of the questionnaire and make any necessary changes.

To ensure accurate data entry, the IST platform sent warning and stop messages when users entered invalid data or skipped the required questions. This improved data quality and reduced manual cleaning. Skip patterns were used in order to allow certain questions or sections of a form to be skipped based on previous answers. That made the survey more efficient. Data validation rules alerted enumerators to errors in the entered data, helping to ensure the sufficient quality of data.

To guarantee data security, the information stored on laptops and sent to the server was encrypted with a key or password. Encryption obfuscates data, rendering it useless unless one has a decryption key or password. This was critical for protecting data both in the local database on the laptop and on the server.

5.3. Use of enumerator map application during CAPI data collection

The IST platform also used QGIS, a free and open-source geographic information system software, to display addresses on a map. The QGIS software was adapted to make it more user-friendly and integrated into the IST platform for easier use by enumerators. Python code behind QGIS (PyQGIS) was used to automate tasks and create custom plugins within the software. The QGIS Python API provided access to a range of functions and tools, enabling users to manipulate data, perform analysis and create custom user interfaces within the software. With just one click, enumerators could open the map, which contained addresses assigned to them. The map was directly connected to the enumerator's address book and any changes in the address book were reflected on the map, and *vice versa*.

5.4. Data transmission system – quality and security

To maintain data consistency and accuracy, it was necessary to synchronise data between a local database and a server. This was done through a web service that enabled secure and efficient data transfer. The synchronisation process, through regular data updates, made the latest data always available. Enumerators initiated two-way data synchronisation manually as many times as necessary, although twice a day (morning and evening) synchronisation was recommended. This allowed close monitoring of fieldwork and timely responses to any challenges. Synchronisation operations were treated as a single unit of work, like transactions, to maintain data integrity and consistency. Partial updates were prevented, and all operations were either completed successfully or not at all. Internet connectivity was required for the synchronisation, and data had to be complete and marked as 'ready to send' before it could be transmitted to the server.

5.5. Data centre

The SORS used centralised server infrastructure in one data center for the census. Two virtual servers with 64 GB RAM, Intel Processor 2.8 GHz, 6 TB HDD and a robust database system were installed in the data center. High availability disaster recovery was provided by replicating data in real time between the two servers. Transactional replication in real time was used. Primary data server was dedicated to transactional processing, while the secondary data server was used for analytical purposes. This solution guaranteed that the performance of the census would be unhindered, and opened the possibility for all the necessary analyses and reporting.

Database and log backup was continuously taken on both SQL servers and there was no possibility of data loss. To ensure the highest possible level of data security and protection, census data were encrypted.

Hardware metrics (including the use of CPU and the usage of memory and disk space) were monitored to assess the overall server performance and detect any performance or resource issues.

5.6. Accuracy and security

To ensure the accuracy and security of the database, log tables (on servers and laptops) were used to track all the changes made, including data inputted by enumerators during daily data entry. These tables were used to monitor enumerators' work, identify errors, and make necessary corrections. Log tables also served as an essential tool for database security, as they allowed the tracking of unauthorised access or failed login attempts. In short, log tables were an effective means of enhancing the census performance, protecting the database from malicious actors, and ensuring the accuracy and integrity of collected data. By means of these tables, administrators could efficiently monitor user activity, identify potential threats, and take appropriate measures to safeguard the database. Overall, the implementation of log tables was crucial for achieving the goals of the survey and maximising data quality.

5.7. Web data collection (CAWI)

To address the challenge of conducting surveys in person, three web applications were developed. One of them was specifically designed for enumerating people residing in social institutions, another was intended for those living in correctional institutions, and the third one was dedicated to individuals working in diplomatic or consular outposts. These applications were developed using the aspx.net framework, and the data collected from them were stored on a separate server for later analysis.

Furthermore, various user-friendly software applications, such as those for equipment distribution and enumerator selection, were also developed to ensure the smooth performance of the census. However, for the purpose of this paper, we focused on the main applications that were described above.

6. Central Operation Control System (monitoring)

To facilitate fieldwork monitoring, a role-based web application was developed using the aspx.net framework. It provided various levels of access and permissions to different types of users based on their roles and responsibilities within the population census organisational system. The application had four different interfaces for four different groups of users: instructors, municipal coordinators, regional coordinators and supervisors. Upon login, the application provided different options depending on the user's name and password.

The primary purpose of the application was to enable real-time monitoring of the fieldwork of enumerators, instructors and municipal coordinators, through tabular views and summary reports, based on the role of the logged-in user. The application also allowed the redistribution of materials from one enumerator to another in the event of an enumerator's resignation or inability to complete their work on time. Instructors could return incorrect or incomplete data to the enumerator for revision or approve a request for data revision. Through the application, users could mark which part of the material they reviewed and access data visualisation platform reports.

An essential feature of the application was generating final worklists for enumerators and instructors based on their performance. The application had a direct connection to the server and the tables that enumerators filled by synchronising data, ensuring the accuracy and consistency of the data recorded.

6.1. Census monitoring tool

To enhance the monitoring and control of the population census process, the SORS implemented data visualisation platform for business analytics. Interactive dashboards and reports were integrated into a web-monitoring application used by the census staff responsible for the project supervision and control.

The main objective of using the data visualisation platform component was to monitor enumerators' work in near real-time and identify any anomalies promptly, to be able to make timely decisions and take appropriate measures wherever necessary. Enumerators synchronised data twice a day, and the collected information was displayed in near real-time using dashboards. Different reports were created and customised for supervisors, regional coordinators, municipal

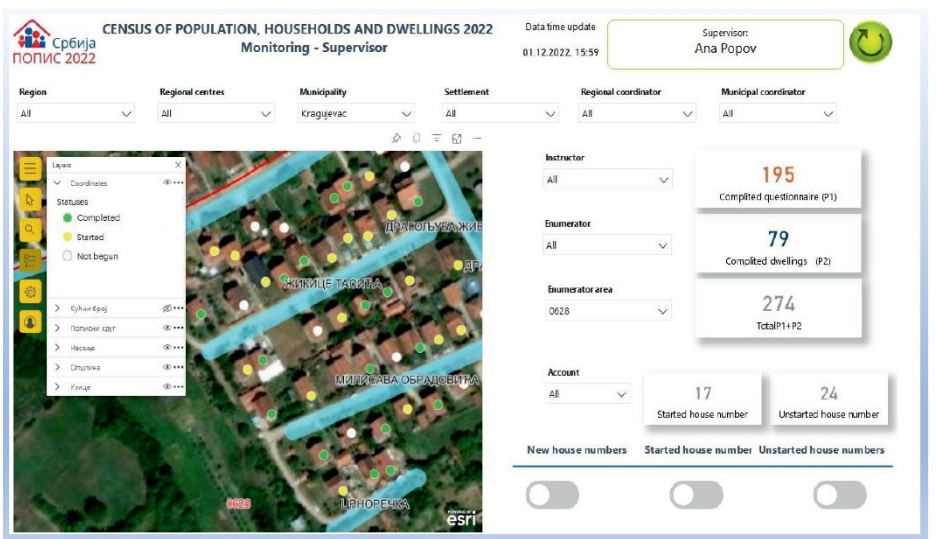
coordinators, municipal instructors, municipal census commissions and info-centres. In data visualisation platform, as in the monitoring web application, row-level security (RLS) approach was adopted to provide each user with access to only this part of work he/she was responsible for, while municipal and regional coordinators had insight into the work of their instructors and enumerators. The reports varied in size and type of data displayed depending on the user's role.

Each report provided comprehensive information on monitoring and controlling the work of all participants involved in the census. One of the key features of the reports was a map showing clear boundaries of census districts and the buildings that were surveyed.

These objects could have three different statuses (indicators):

- *not started* – all objects that were the subject of the census and were on the enumerator's list. Such objects were marked with a grey circle. At the very beginning of the census, only grey indicators could be seen on the map;
- *started* – objects in which the enumerator found a certain number of people but did not find them all. The yellow circle on the map indicated such dwellings;
- *completed* – all persons inside that building were enumerated. A green circle marked such buildings (Figure).

Figure. Example of monitoring Power BI report page on the 2022 Census of Population, Households and Dwellings



Source: authors' work.

It was possible to filter the desirable data within each report. By selecting a specific census participant, the report filtered information related only to her/him.

All important information, such as full name, telephone number and e-mail address of each participant assigned to the current report user could be seen, which greatly facilitated the communication process.

In addition, the SORS created a special Power BI report, intended for municipal census commissions. During fieldwork, the municipal census commissions were faced with large fluctuations of data. A large number of participants had changed their original status, so it was necessary to provide the commissions with a direct view of the central database, in real time. Authentication at the entry of the report ensured that each commission could see only the data for which it was responsible.

Customised Power BI report pages were created exclusively for top-level supervisors to facilitate their strategic and operational decision-making. These pages provided information necessary to determine whether it was necessary to recruit more enumerators and to identify critical territories where enumeration might have been carried out beyond the given timeframe.

Table 1. Data visualisation platform reports content on the 2022 Census of Population, Households and Dwellings

Row-level security	Specific report content	All reports content
Supervisors	Detailed side-by-side overview, and comparison of the present census with the one conducted in 2011. Insights to determine whether it was necessary to recruit more enumerators and to identify critical territories where enumeration might not be completed within the given timeframe	• Full name and role of the logged user • Time of the last data update • List of census participants and districts • Information about the enumerated population, households and dwellings • Number of completed questionnaires and rejections • Table with detailed information about the enumerated dwellings and population • Map with clearly defined boundaries of the census district and objects
Regional coordinators	The list of additional and on-duty enumerators, a side-by-side overview and comparison of the present census with the one conducted in 2011	
Municipal coordinators	Instructors and their enumerators' performance by day. Page with information which house numbers are assigned to a specific enumerator	
Municipal instructors	Enumerators' performance by day, their activity status (active/inactive), as well as the status of census districts, whether they are active or idle	
Municipal census commissions	Direct view of all participants' statuses and census staff statuses in a central database, in real time	
Info-centres	Contact information to enumerators to whom a certain address was assigned	

Source: authors' work.

6.2. Data visualisation platform metrics

The timely completion of the population census in Serbia was facilitated by the implementation of two key decisions, namely the recruitment of additional

enumerators and the reorganisation of existing enumerators based on their expertise. The identification of potential issues using the census monitoring tool, and the prompt response of the authorities to the changing situation were crucial.

Table 2. Data visualisation platform – number of report views

Power BI report	Views	Unique viewers
Info-centres	580	5
Municipal instructors	394 435	1 998
Municipal coordinators	46 031	255
Municipal census commissions	18 296	158
Regional coordinators	3 375	24
Supervisors	2 912	15

Source: authors' work.

The data collected from this metrics indicates that approximately 2,500 different viewers viewed the reports nearly half a million times. Notably, the viewers accessed the reports at least five times a day, which suggests a significant demand for the information provided.

6.3. The monitoring system and planning for enumerator reserves as key measures in overcoming population census challenges in Belgrade

At a certain point, the population census in Belgrade was facing a significant challenge. The working dynamics of the enumeration process were not aligned with the objectives and the census activities were in danger of not being completed on time. It was a critical situation that required immediate action.

With the help of the Power BI reports, the authorities were able to gain insight into the situation. They realised there was a shortage of enumerators, which was a major hindrance to the progress of the census.

As the successful completion of the census was a priority, it was necessary to begin with contingency planning for enumerator reserves. This meant developing a comprehensive plan for the deployment and management of enumerators, including training, scheduling and logistics. The plan also needed a special focus on Belgrade. In this regard, two strategic decisions were made:

- the first one was to train additional enumerators specially for Belgrade. This was expected to help address the shortage of enumerators and increase the efficiency of the census process. Additional Belgrade training sessions followed a special, fast schedule, spanning over three days (starting on the 7th, 10th, and 14th day of the census). In total, 280 enumerators were trained (a hundred enumerators trained in each of the first two sessions and 80 in the last one). These staff

memberts successfully enumerated 87,000 persons in Belgrade, i.e. 5.2% of the total enumerated population of Belgrade;
- the second decision was to regroup the best enumerators from cities in Vojvodina and Central Serbia and send them to Belgrade. However, this was done after they had completed their original tasks in the respective areas. A total of 256 enumerators from 68 different municipalities were re-hired for Belgrade, and they managed to enumerate approximately 33,000 persons, i.e. 2.1% of the total enumerated population of Belgrade.

6.4. Help Desk Ticketing System

The SORS set up an IT support team to assist instructors and enumerators during the census implementation process. The team consisted of three levels of support: IT assistants (IT1), IT staff members at the SORS regional centres (IT2), and senior IT staff members at the SORS headquarters in Belgrade (IT3). Weekly online meetings were held to improve the process. The Population Census Help Desk Ticketing System was created to provide automated communication and store solutions to hardware and software problems.

Table 3. IT support team activities during the 2022 Census

IT support level	Activities
IT1	• Receiving phone calls about specific hardware or software problems in the field • Creating new tickets • Solving problems or forwarding them to a higher level of support • Sending automated notifications to a higher level, if needed
IT2	• Solving problems forwarded by IT1 and changing ticket status to resolved • Forwarding complex problems to a higher level of support • Sending automated notifications to a lower or higher level, if needed
IT3	• Solving problems forwarded by IT2 and changing ticket status to resolved • Sending automated notifications to a lower level, if needed

Source: authors' work.

The IT support team played a crucial role in the successful census implementation by using an automated communication system. This system enabled support in order to resolve tickets at the appropriate level or forward them to the next support level for further assistance. The process was streamlined with email notifications sent automatically to higher support levels when a ticket was submitted and to lower levels when the issue was resolved.

7. Use of machine learning in post-enumeration phase for occupation and economic activity classification

In the post-enumeration phase of a census, classifying occupations and economic activities, based on the International Standard Classification of Occupations (ISCO) and the Statistical Classification of Economic Activities in the European Community (NACE), usually involves large workload (and thus requires considerable human resources). Automatic classification through data entry applications is often not feasible due to the lack of unique descriptions for different professions and activities. To reduce the time required for the classification, machine learning (ML) has emerged as a natural solution (United Nations Economic Commission for Europe Statistics Wikis, n.d.). The task involves creating an algorithm that can classify activities and occupations on the basis of a written description the interviewer inputs listening to answers to open-ended questions.

ML classification relies on large datasets of descriptions and classifications of occupations, which consider the impact of various factors such as the education level, age and gender. Although occupation and activity descriptions are a primary factor in classification, the education level has the most significant influence on it. For instance, an economist with a high school education is likely to be classified differently from the one with a university degree, even if their occupation descriptions are the same. Gender also plays a role in determining professions and activities, as some may be predominantly associated with one gender, such as miners being mostly men. Age can also carry information, as more 'modern' occupations are more likely to be associated with younger respondents, while older respondents are more likely to have traditional occupations.

To build an ML model, data sets were created on the basis of information derived from the census and other statistical surveys, including various descriptions of occupations, activities, education levels, and age and gender of respondents.

7.1. Cloud (hardware) and data anonymising

The size of the dataset generated by the algorithms is extensive, requiring more powerful hardware than a standard PC to process the data efficiently and reduce model training time.

The SORS used a cloud for ML due to its scalability, security, and user-friendly interface. Anonymisation was done before moving the data to the cloud to protect sensitive information and secure privacy. Only necessary variables were extracted from the database and a unique row identification was created to later connect the original database with the rows extracted for data processing on the cloud. After the classification using ML, data were pulled out from the cloud and transferred to the

SORS data centre to be connected with the corresponding records via unique row identification. Appropriate security and access controls were implemented (including encryption, firewalls, and IAM (Identity Access Management) tools) to prevent unauthorised access or disclosure of sensitive data.

The system used on the cloud for the ML purposes was equipped with 64 CPUs, 256 GB of RAM, an Intel(R) Xeon(R) Platinum 8370 C CPU @ 2.80 GHz 2.79 GHz processor, and a 64-bit operating system, x64-based processor.

7.2. Data set for the ML model

The initial data set for the ML model was sourced from the previous census conducted in 2011, which comprised approximately 2.5 million descriptions of various occupations and activities. However, since this data was collected using the OCR technique (the technology that allows software to interpret text on scanned images), the text obtained from scanned questionnaires was mostly unusable for the model creation. Thus, another ML model was developed to identify relevant words that corresponded to the descriptions of occupations and activities. To create the ML model, a combination of two datasets from different sources was used. The first source was the Labour Force Survey (LFS), which collected data from approximately 30,000 respondents every three months. This survey provided a constant influx of new data and included new occupations and activities not represented in the 2011 census. The second source was the Central Register of Mandatory Social Insurance (CROSO), which contained about 2.5 million records of officially registered occupations and activities for all the citizens of the Republic of Serbia who had social insurance. The use of these two data sets provided an ideal training set for the ML model, as the descriptions of occupations and activities were entered through an application, eliminating errors resulting from the manual entry. Fields chosen for the data set layout were: AGE_LEVEL, SEX, ISCO_TXT, NACE_TXT, ISCO4D and NACE3D. The textual components of the data set were transformed into a vector form using the TF-IDF (term frequency – inverse document frequency) algorithm, which generates a vector for each word based on its frequency in the total dictionary. Due to hardware limitations, TF-IDF vectoriser settings remained at: min_df=5 (remove words that are mentioned in less than 5 descriptions), max_df=0.9 and use_idf=True. With max_df=0.9, we removed from the dictionary the words whose frequency of occurrence is greater than 90%. The rationale behind this is that words which occur very often, such as ‘and’, ‘the’, and ‘is’, do not contain much information about the content of a data set. By using max_df, we excluded such terms from our feature set.

The resulting vector contains approximately 13,000 members. However, the matrix size produced from the vector is enormous and requires substantial hardware resources. To address this challenge, dense matrices were converted to sparse matrices.

7.3. ML algorithms used during training

Two approaches, i.e. classic classification algorithms and neural networks were considered when decisions about the classification method were made. While the latter are known to offer higher accuracy, their computational demands and requirement for extensive fine-tuning were deemed impractical given the available timeframe. As such, the SORS opted for using the common ML algorithms instead.

Three algorithms were tested for accuracy: logistic regression, random forest classifier, and XGBoost. After initial testing, the random forest classifier was chosen due to its superior accuracy compared to the other two algorithms (61.08%).

The hyperparameters used during GridSearchCV with validation across three different mixtures of the training dataset are: {'n_estimators': [50, 60, 80, 100, 150, 160, 200], 'max_features': ['log2', 'sqrt'], 'max_depth': [80, 90, 100, 110, None], 'min_samples_split': [2, 5, 10, 20, 50], 'min_samples_leaf': [1, 2, 4], 'bootstrap' : [True, False]}.

The best hyperparameters obtained for the best model are: {'n_estimators': [160], 'max_features': ['log2'], 'max_depth': [None], 'min_samples_split': [2], 'min_samples_leaf': [1], 'bootstrap': [True]}.

The classification accuracy achieved 98% after training and validating on the training set. Classification accuracy testing was conducted on a sample checked by coders, and the reports were satisfactory.

The SORS's next step was to train the ML algorithm to correct the words from the previous census, thereby enriching the data set with 2.5 million new entries.

8. Census undercoverage and use of administrative data for item imputation

Given the purpose of the population census – to provide a comprehensive picture of the entire population, including citizens who did not participate in the census for whatever reason (out of country, did not want to participate, not found on the address), the SORS's efforts to produce high-quality census data do not end when the fieldwork is done. They continue during the data processing stage.

It turned out during fieldwork that the response rate in some urban areas was lower than in previous censuses. However, the fact that the SORS had access to all relevant administrative registers, including the Personal Identification Number

(PIN), made it possible to use a combined census method. The combined census in this case means that a database obtained through the direct census interviews is supplemented by administrative data for those citizens who did not participate in the census for whatever reason.

Most census topics are not fully covered by one administrative data source, but are likely to be partially represented in several different sources. Different approaches and methods are necessary to combine information and draw full information from a range of sources. In line with the census methodology, the administrative sources used to determine the permanent resident population in Serbia are: the recently established Central Population Register as well as administrative registers on the employed and unemployed, students and children attending kindergartens, retired people, taxpayers, beneficiaries of social assistance, beneficiaries of health insurance, etc. All the available administrative data were integrated into one master administrative dataset. Because the Central Population Register contains data on all citizens of Serbia, regardless of whether they live in the country or not, additional information from other listed records (called 'signs of life') was used to determine the groups of the population that definitely lived in Serbia on the Census Day, and as such corresponded with the definition of permanent residents. The record linkage of census and administrative data was performed by means of both the deterministic and probabilistic method.

Before matching the census and administrative data, preprocessing tasks were carried out to maximise the quality of the data linkage. These activities included:

- cleaning the matching variables (e.g. name, date of birth, address), which means that the variables were converted to an adequate format; null strings were ignored; abbreviations, punctuation marks, upper/lower cases, etc. were cleaned; and all the necessary transformations were carried out to standardise the variables;
- deriving new matching variables (for instance, month and year of birth);
- removing some records from dataset: (a) duplicates missed by the resolve multiple responses process, (b) individuals born after the Census Day, (c) records missing names and date of birth.

The second task was to choose variables for the record linkage. The parameters used for matching records from the census and administrative registers were PIN (where it was available and correct), or a combination of the following parameters: the municipality of birth, the date of birth, name, surname, parent's name, the municipality of residence and home address (in the cases where the PIN was missing or inaccurate). The high degree of certainty required for the linkage was achieved because of the presence of PIN in both datasets. About 80% of individuals in the census provided the correct PIN.

To improve the scalability of the linkage process, the creation of blocks was carried out using the following variables: name, surname and the date of birth, to limit the comparison of records to pairs or groups that likely correspond to the same entity. Subsequently, deterministic record linkage was applied to several subsets of the aforementioned parameters. As a result of this phase, many records were successfully matched, and the working datasets were significantly reduced.

Afterwards, some further analysis and probabilistic record linkage were done on the restricted datasets without records that were already matched. With smaller unmatched datasets, both from the census and administrative data, it was possible to perform some more detailed and time-consuming methods. Two of the algorithms that were used most frequently for comparing strings were the Soundex and the Levenshtein distance algorithm.

In the last stage, the matching of records between the census and administrative master datasets required significant clerical effort to find matches that could not be found by automatic means.

The process described above yielded two datasets: the first (A dataset), used for the imputation of the census data – individuals existing in administrative registers and considered as usual population but unaccounted for by the census (≈218,000) and the other (B dataset), consisting of individuals not found in the administrative data but present in the census dataset (≈50,000). In the A dataset, individuals were grouped in households wherever it was possible according to the address or common children living at the same address. Then associative clerical matching was used to check matched households containing unmatched persons. Unmatched households containing matched persons were also investigated to determine whether the households should also be matched. It was possible to make new person matches, or to break the existing person and household matches. This enabled the optimal within-household person linkage. Clerical work was also performed to further explore the dataset of individuals present in the census but not found in administrative registers.

9. Statistical Population Register

Over the past few decades, administrative registers have grown in importance as a source of data for official statistics. They offer numerous advantages, including cost reduction, improved data quality, and the ability to produce more frequent information.

Following trends prevailing among the EU member states, the transition from a traditional approach to the one relying on the use of administrative data for statistical purposes is a part of the SORS's vision of developing a more general

register-based statistical system. To fulfill this vision, bearing in mind that transition is a long-term and complex task, the SORS continues its work on the creation of conditions that would make it possible to conduct the next census (in 2031) on the basis of administrative data.

On the way to the introduction of a census based on register data, numerous activities have been planned. The first step after the completion of the 2022 Census is carrying out an exercise to simulate a 'register-based census' using administrative sources available at that time point, and to compare the administrative data with data obtained through the census.[2] As a result, the quality of both main registers and supplementary registers will be assessed, showing the potential shortcomings in terms of availability, coverage, concepts, reference date, accuracy, etc.

In this regard, it is planned to implement the following:

- the SORS is going to prepare a road map with the proposed targets and to define necessary steps to be taken on the basis of the main findings from the previous activity;
- afterwards, national workshops will be organised to bring together the representatives of relevant institutions to consider the possibilities, future steps, deadlines, roles and responsibilities in the process of introducing a register-based census;
- the conclusions arising from the workshops will be presented by the SORS at a conference;
- the final step will involve drawing up the action plan for the introduction of a register-based census and identifying all public administration bodies relevant for this activity.

Considering that it is not possible to collect all the necessary census variables (core topics) defined by the Recommendations (United Nations, 2015, 2017) from administrative sources, the SORS, relying on examples of good international practice, decided to overcome the lack of certain data by establishing a statistical population register (SPR).[3] The need for the SPR was already recognised in the previous census cycle, but creating it then was not feasible due to the absence of essential administrative registers like the Central Population Register. The establishment of the SPR has became possible thanks to the improvement of the national system of administrative registers in recent years.

The SPR will be established on the basis of the 2022 Census data and will be regularly updated using information from available administrative sources. The master file of the SPR will consist of the data from the 2022 Census. All the core

[2] Within the *IPA 2018 National Programme – EU for Development of Statistics in Serbia* project.
[3] Within the *IPA 2022 National Programme – EU support for efficient statistics and further strengthening statistical infrastructure* project.

variables defined by international recommendations, which are available and of satisfactory quality, will be regularly taken from administrative registers for updating the SPR. For example, the level of education will be updated from administrative sources for persons who continue receiving education; otherwise, a person's education status will remain the same as it was stated during the census. Variables defined as non-core according to international recommendations (such as means of transport, etc.), will be available only from the 2022 Census. So administrative registers will be used as sources for the statistical register, but the reverse process will not be possible, respecting the principle of 'one-way flow'.

The SPR will be one of the key preconditions for the 2031 census to be carried out as register-based. In addition, the SPR will be an up-to-date sample frame for running statistical surveys, and it will contribute to the annual production of selected population datasets foreseen by EU legislation.

10. Conclusions

The use of digital technology in censuses has become increasingly important due to the advantages it offers over traditional paper-based methods. In the SORS's experience, the digital census improved the accuracy, reduced costs, and speeded up the process of data collection, processing, and dissemination. Additionally, it allowed the adoption of advanced technologies such as geospatial technology, machine learning, record linkage or imputation, which, in turn, allowed a more efficient and effective analysis of the census data.

Compared to a register-based census, traditional census based on a digital technology provides more detailed and accurate information on the demographic properties, such as ethno-cultural characteristics or household and family composition. Traditional census may be more inclusive by ensuring that everyone is counted, regardless of their participation in government programmes or their legal status. However, traditional censuses are more expensive, time-consuming and labor-intensive than register-based censuses.

The SORS, on the basis of its experience from 2022, believes that a digital traditional census, without self-enumeration, that adopts administrative data for item imputation, is the most effective way to obtain precise and comprehensive information about a population, including its demographic characteristics, geographic distribution and the overall size.

References

Law on 2021 Census of Population, Households and Dwellings (Official Gazette of RS No. 9/20 of 4 February 2020). https://popis2022.stat.gov.rs/media/1596/law-on-the-2021-census.pdf.

United Nations. (2015). *Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing*. http://www.unece.org/publications/2020recomm.html.

United Nations. (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3*. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf.

United Nations. (2019). *Guidelines on the use of electronic data collection technologies in population and housing censuses*. https://unstats.un.org/unsd/demographic/standmeth/handbooks/data-collection-census-201901.pdf.

United Nations Economic Commission for Europe Statistic Wikis. (n.d.). *Machine Learning for Official Statistics*. https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home.

Zakon o izmienama zakona o popisu stanownisztwa, domaćinstawa i stanowa 2022. Godine (Official Gazette of RS No. 35/21 of 16 April 2021). https://popis2022.stat.gov.rs/media/1595/zakon-o-popisu-2021.pdf.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 71–79
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

XII Ogólnopolska Konferencja Naukowa im. Profesora Zbigniewa Czerwińskiego „Matematyka i informatyka na usługach ekonomii"

The 12th Professor Zbigniew Czerwiński National Scientific Conference 'Mathematics and IT at the services of economics'

22 września 2023 r. odbyła się XII Ogólnopolska Konferencja Naukowa im. Profesora Zbigniewa Czerwińskiego poświęcona zastosowaniom metod matematycznych i statystycznych oraz informatyki w ekonomii. Konferencja została zorganizowana przez Instytut Informatyki i Ekonomii Ilościowej (IIEI) Uniwersytetu Ekonomicznego w Poznaniu (UEP), a honorowy patronat nad nią objęli prof. dr hab. Maciej Żukowski, rektor UEP, oraz dr Dominik Rozkrut, prezes GUS. W ramach wydarzenia odbyły się zdalne sesje naukowe oraz sesja jubileuszowa z okazji 60-lecia Urzędu Statystycznego (US) w Poznaniu.

Komitetowi organizacyjnemu konferencji przewodniczyli dr hab. inż. Jarogniew Rykowski, prof. UEP, i dr hab. Krzysztof Węcel, prof. UEP. W pracach komitetu uczestniczyli: dr Marcin Bartkowiak, dr hab. Agata Filipowska, prof. UEP, dr hab. Helena Gaspars-Wieloch, prof. UEP, dr hab. Paweł Kliber, prof. UEP, mgr Natalia Lewandowska (sekretarz konferencji) i dr hab. Marcin Szymkowiak, prof. UEP. Komitet naukowy konferencji pracował pod kierunkiem prof. dr. hab. Krzysztofa Malagi, dyrektora IIEI, a w jego skład weszli: prof. dr hab. Witold Abramowicz, dr hab. Dorota Appenzeller, prof. UEP, prof. dr hab. Barbara Będowska-Sójka, prof. dr hab. inż. Wojciech Cellary, dr hab. Grażyna Dehnel, prof. UEP, dr hab. Krzysztof Echaust, prof. UEP, prof. dr hab. Elżbieta Gołata, prof. dr hab. Witold Jurek, prof. dr hab. Marian Matłoka, prof. dr hab. Emil Panek, dr hab. Elżbieta Rychłowska-Musiał, prof. UEP, i prof. dr hab. inż. Krzysztof Walczak.

Zasadniczym celem konferencji była integracja środowiska polskich ekonomistów zajmujących się badaniami operacyjnymi, ekonometrią, ekonomią matematyczną, informatyką i statystyką oraz praktyków. Spotkanie stało się forum dyskusji i wymiany doświadczeń naukowców, pracowników urzędów statystycznych i przedsiębiorców, zajmujących się zastosowaniami metod ilościowych i informatyki.

Tematem przewodnim konferencji były nowoczesne rozwiązania w szeroko rozumianej gospodarce i zarządzaniu związane z czwartą rewolucją przemysłową. Wie-

lu prelegentów omówiło zastosowania zaawansowanych metod ilościowych i informatycznych w przemyśle i administracji.

Referaty przedstawione na konferencji dotyczyły takich zagadnień, jak: ekonometria i modelowanie zjawisk makro- i mikroekonomicznych, badania operacyjne i ich wykorzystanie do celów statystyki publicznej, zastosowanie narzędzi informatycznych i nowych technologii w analizie danych ekonomicznych oraz zastosowanie metod ilościowych w przemyśle. W wydarzeniu – przy uwzględnieniu uczestników konferencji i gości przybyłych na jubileusz 60-lecia US w Poznaniu – wzięło udział blisko 180 osób. Byli to przedstawiciele świata nauki (13 uczelni i instytutów badawczych), pracownicy GUS i urzędów statystycznych, przedsiębiorcy, delegaci instytucji rządowych i samorządowych oraz studenci.

Pierwszej sesji plenarnej, która odbyła się w formie hybrydowej, przewodniczył Krzysztof Węcel. Jako pierwszy głos zabrał Dominik Rozkrut, który wygłosił referat *Statystyka publiczna w europejskim ekosystemie danych*. W wystąpieniu przywołał najważniejsze unijne akty legislacyjne mające istotny wpływ na rozwój statystyki publicznej, również w Polsce. Szczególną uwagę zwrócił na obowiązujące rozwiązania prawne w kontekście dostępu do danych wykorzystywanych przez krajowe urzędy statystyczne, na budowę bardziej złożonego ekosystemu interesariuszy oraz wyzwania, jakie nowe regulacje niosą dla oficjalnych systemów statystycznych. Podkreślił, że głównym celem obowiązujących i planowanych rozwiązań legislacyjnych w Unii Europejskiej jest wzrost zaufania do danych, zwiększenie ich dostępności oraz pokonywanie barier technicznych w ponownym wykorzystywaniu danych. W podsumowaniu wystąpienia zaznaczył, że w świetle nowych propozycji i rozwiązań prawnych coraz większego znaczenia nabiera zarządzanie danymi, które stwarza możliwości rozwoju statystyki publicznej.

Następnie głos zabrał dr Jacek Kowalewski, dyrektor US w Poznaniu, który w referacie *Wybrane aspekty funkcjonowania Urzędu Statystycznego w Poznaniu* przybliżył uczestnikom konferencji historię, cele i osiągnięcia kierowanej przez siebie jednostki. Podkreślił znaczenie pracowników jako podstawowego kapitału instytucji, a także scharakteryzował w liczbach urząd i jego działalność.

Elżbieta Gołata i Grażyna Dehnel w wystąpieniu *Jubileusz 60-lecia Urzędu Statystycznego z perspektywy Uniwersytetu Ekonomicznego w Poznaniu* opisały przeszłą, bieżącą i planowaną współpracę UEP z US w Poznaniu, podkreślając wagę długoletniego współdziałania między tymi instytucjami na płaszczyźnie organizacyjnej i naukowej.

Druga sesja plenarna, poprowadzona w formie hybrydowej przez dr Arletę Olbrot z US w Poznaniu, była sesją jubileuszową zorganizowaną z okazji 60-lecia powstania

US w Poznaniu. W uroczystości wzięło udział 150 gości, a wśród nich obecni i emerytowani pracownicy urzędu, przedstawiciele instytucji rządowych i samorządowych, urzędów statystycznych, uczelni wyższych, Polskiego Towarzystwa Statystycznego, Poznańskiego Towarzystwa Przyjaciół Nauk, Komendy Wojewódzkiej Państwowej Straży Pożarnej i Komendy Wojewódzkiej Policji.

Słowo powitalne wygłosił Jacek Kowalewski. Następnie uczestnicy spotkania obejrzeli film *60 lat Urzędu Statystycznego w Poznaniu*, po którym wystąpili zaproszeni goście. Z okazji jubileuszu urzędu dyrektorowi Kowalewskiemu gratulacje złożyli oraz przekazali życzenia pomyślności i sukcesów dla pracowników urzędu: Maciej Żukowski, rektor UEP, Iwona Matuszczak-Szulc, dyrektor Wydziału Rozwoju Miasta i Współpracy Międzynarodowej Urzędu Miasta Poznania, Jowita Maćkowiak, dyrektor Wojewódzkiego Biura Planowania Przestrzennego w Poznaniu, Magdalena Wegner, dyrektor US w Szczecinie, Roman Fedak, dyrektor US w Zielonej Górze, prof. dr hab. Józef Orczyk, przewodniczący Wojewódzkiej Rady Rynku Pracy, prof. dr hab. Krzysztof Malaga, a w imieniu pracowników Katedry Statystyki UEP – Grażyna Dehnel, kierownik katedry, i Elżbieta Gołata, prorektor ds. nauki i współpracy z zagranicą. W czasie uroczystości odczytano także listy gratulacyjne, które nadesłali: Michał Zieliński, wojewoda wielkopolski, Jan Grabkowski, starosta poznański, Jacek Jaśkowiak, prezydent Poznania, insp. Robert Kasprzyk, komendant miejski policji w Poznaniu, Krzysztof Skrzypczak, dyrektor Okręgowego Urzędu Miar w Poznaniu, prof. dr hab. Krzysztof Szoszkiewicz, rektor Uniwersytetu Przyrodniczego w Poznaniu, dr hab. Artur Zimny, prof. ANS, rektor Akademii Nauk Stosowanych w Koninie, i prof. dr hab. Mirosław Krzyśko z Zakładu Statystyki Matematycznej i Analizy Danych Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Podczas uroczystości siedmioro zasłużonych pracowników urzędu otrzymało medale za długoletnią służbę. Dekoracji dokonał Jerzy Fleming, radca Wojewody Wielkopolskiego. Z kolei Dominik Rozkrut odznaczył 73 pracowników US w Poznaniu odznakami honorowymi „Za Zasługi dla Statystyki RP". Za szczególny wkład w rozwój statystyki publicznej w regionie odznaczenia od prezesa GUS otrzymali również: Maciej Żukowski, Krzysztof Malaga i dr Aleksandra Witkowska, emerytowany pracownik naukowy UEP.

Podczas konferencji odbyło się pięć sesji tematycznych, w trakcie których wygłoszono 15 referatów przygotowanych przez przedstawicieli: Akademii Górniczo-Hutniczej (AGH) w Krakowie, Akademii Kaliskiej im. Prezydenta Stanisława Wojciechowskiego, Akademii Mazowieckiej w Płocku, Politechniki Warszawskiej (PW), Szkoły Głównej Handlowej w Warszawie (SGH), Uniwersytetu Ekonomicznego

(UE) w Katowicach, UE w Krakowie, UEP, Uniwersytetu Warmińsko-Mazurskiego w Olsztynie (UWM), Uniwersytetu Warszawskiego (UW), Uniwersytetu Zielonogórskiego (UZ), Urzędu Miasta Ustka, US w Poznaniu, US w Zielonej Górze i Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie (ZUT). Ponadto w ramach dwóch sesji studenckich zaprezentowano pięć referatów studentów UEP.

W pierwszej sesji naukowej – poświęconej ekonometrii, która odbywała się równolegle z sesją jubileuszową US w Poznaniu, a której przewodniczyła dr hab. Agata Kliber, prof. UEP – wygłoszono trzy referaty. Sesję rozpoczął Błażej Supron (ZUT) referatem *Energia odnawialna i nieodnawialna a emisja CO_2 w krajach Grupy Wyszehradzkiej: dowody empiryczne z zastosowania modelu NARDL-PMG*. Celem przeprowadzonego przez niego badania było ustalenie krótko- i długoterminowego oddziaływania produkcji energii ze źródeł odnawialnych i nieodnawialnych oraz wzrostu gospodarczego na emisję CO_2 w Czechach, Polsce, na Słowacji i Węgrzech w latach 1991–2021. W badaniu zastosowano asymetryczny model NARDL-PMG dla danych panelowych. Prelegent wykazał istnienie asymetrycznych długookresowych zależności między emisją CO_2 a wzrostem gospodarczym oraz produkcją energii ze źródeł nieodnawialnych. Analogiczne zależności między emisją CO_2 a zużyciem energii ze źródeł odnawialnych okazały się symetryczne, a w krótkim okresie – nieistotne statystycznie.

Jacek Wolak i Justyna Tora (AGH) w referacie *Charakterystyka popytu na napoje słodzone w gospodarstwach domowych w Polsce* ukazali wpływ zmiany cen napojów bezalkoholowych na przyszłą konsumpcję napojów słodzonych. Analizy dokonali na podstawie danych jednostkowych pochodzących z badań budżetów gospodarstw domowych przeprowadzonych przez GUS w latach 2011 i 2020.

Ostatni referat w tej sesji – *Pomiar i predykcja nierówności społecznych z wykorzystaniem danych z rynku nieruchomości* – został przygotowany przez Tomasza Stachurskiego, Tomasza Żądłę i Alicję Wolny-Dominiak (UE w Katowicach). W wystąpieniu omawiano polski rynek nieruchomości i zaproponowano metody predykcji czterech mierników dyspersji opartych na kwantylach. Analizowano lata 2015–2021, a poziomem agregacji danych były powiaty.

Drugiej sesji równoległej, poświęconej badaniom operacyjnym, przewodniczyła Helena Gaspars-Wieloch. Pierwszy referat – *Efektywność badawczo-rozwojowa instytutów Polskiej Akademii Nauk oraz czynniki zewnętrzne mające na nią wpływ* – zaprezentowali Artur Prędki (UE w Krakowie) i Łukasz Brzezicki (Urząd Miasta Ustka). Celem badania omówionego w wystąpieniu było dokonanie oceny efektywności badawczo-rozwojowej instytutów Polskiej Akademii Nauk (PAN) w 2019 r.

i określenie czynników zewnętrznych, które istotnie na nią wpływają. Warto dodać, że to pierwsza ocena efektywności działalności badawczo-rozwojowej instytutów PAN po wprowadzeniu reformy systemu nauki i szkolnictwa wyższego w 2018 r.

Daniel Konrad Kaszyński (SGH) w referacie *Metody sprawiedliwej oceny zdolności kredytowej* omówił modele oceny zdolności kredytowej, które są coraz częściej stosowane w praktyce. Zauważył, że coraz powszechniejsze wykorzystywanie produktów kredytowych prowadzi do popularyzacji modeli oceny zdolności kredytowych zarówno w procesie podejmowania decyzji kredytowej, jak i w monitoringu kredytowym. Obok technicznych podejść do ewaluacji działania modeli oceny zdolności kredytowej (moc dyskryminacyjna, jakość kalibracji i stabilność modeli) prelegent przedstawił miary identyfikujące pojawianie się zagadnień stronniczości algorytmicznej i metody jej redukcji.

Ostatni referat w sesji – *Potencjał analityczny badania koniunktury w gospodarstwie rolnym* – przedstawiła Agnieszka Jankowska (US w Zielonej Górze). Jej wystąpienie, oprócz omówienia metodologicznych i poznawczych aspektów badania koniunktury w gospodarstwach rolnych, obejmowało prezentację wskaźników oraz ich analizę pod kątem diagnozowanych i prognozowanych zjawisk ekonomicznych w rolnictwie.

Trzeciej sesji naukowej, poświęconej ekonometrii, przewodniczył Paweł Kliber. Sesję otworzył Artur Wyszyński (UWM) referatem *Wpływ inwestorów sektora publicznego na wyniki finansowe profesjonalnych klubów piłki nożnej*, w którym na przykładzie polskiej Ekstraklasy określił zakres własności państwowej i dokonał oceny wyników finansowych państwowych i prywatnych przedsiębiorstw prowadzących kluby piłki nożnej. Przedmiotem badań był panel składający się z 98 podmiotów zawierający 26 przedsiębiorstw prowadzących kluby piłki nożnej, których męskie drużyny w sześciu sezonach (od 2016/2017 do 2021/2022) uczestniczyły w rozgrywkach Ekstraklasy.

Pytanie *Czy sztuczna inteligencja może poprawić dobrostan człowieka na emeryturze?* postawiła dr Alicja Jajko-Siwek (UEP). Celem badania omawianego w referacie było sprawdzenie możliwości posłużenia się w modelowaniu dobrostanu emerytalnego metodami uczenia maszynowego, a dokładniej – wzmacnianymi gradientowo algorytmami opartymi na drzewach decyzyjnych i algorytmami wywodzącymi się z teorii gier. Analizy przeprowadzono na podstawie danych z panelowego badania SHARE, dotyczącego zdrowia, starzenia się i emerytur osób w wieku 50 lat i więcej w Europie. Oprócz oceny przydatności analizowanych metod uczenia maszynowego do badania dobrostanu rozważano użyteczność rankingów ważności zmiennych –

algorytmów wbudowanych w poszczególne metody i metody SHAPleya. Wyniki badania wskazują, które obszary życia osób starszych mają dla nich największe znaczenie, a tym samym jakie działania należałoby podjąć, aby poprawić dobrostan osób na emeryturze.

Sesję zakończyło wystąpienie *The analysis of selected properties of expert system algorithms for forecasting the telecommunications market* Pawła Kaczmarczyka (Akademia Mazowiecka w Płocku). Prelegent wskazał algorytmy systemów ekspertowych, których można użyć do predykcji rynku telekomunikacyjnego. Omówione w wystąpieniu badanie dotyczyło wpływu wartości parametrów i długości serii na wydajność algorytmów i jakość prognozy.

Kolejną sesję naukową, poświęconą statystyce, poprowadziła Grażyna Dehnel. Pierwszy referat – *Sponsored content in contextual bandits. Targeting not at random case* – wygłosił Hubert Drążkowski (absolwent PW), który zaproponował nowe narzędzie analizy interakcji między systemem rekomendacyjnym i sponsorem. Zauważył, że ignorowanie faktu sponsoringu może prowadzić do gorszych wyników i mniejszego zaangażowania uczestników, oraz wskazał na podobieństwo problemu do sekwencyjnego podejmowania decyzji.

W referacie *A simple covariate distribution balance method with application to non-probability samples and observational data* Maciej Beręsewicz (UEP) zaproponował prostą metodę balansowania rozkładów na potrzeby badania związków przyczynowo-skutkowych opartą na podejściach Deville'a i Särndala oraz Harmsa i Duchesne'a. Autor porównał ją do innych metod opisanych w literaturze przedmiotu i przedstawił wyniki badań symulacyjnych.

Referat *Edukacja w zakresie kontroli ujawniania danych – potrzeby i wyzwania*, który przygotowali Andrzej Młodak (Akademia Kaliska im. Prezydenta Stanisława Wojciechowskiego, US w Poznaniu), Tomasz Klimanek i Tomasz Józefowski (UEP, US w Poznaniu), dotyczył niektórych podjętych w ostatnim czasie działań związanych z metodami i narzędziami kontroli ujawniania danych (ang. *statistical disclosure control* – SDC). Działania te obejmują specjalistyczne warsztaty szkoleniowe dla członków zespołu odpowiedzialnego za SDC, w tym obecnych i przyszłych ekspertów w tej dziedzinie w GUS, a także szkolenia przygotowawcze dla nowych pracowników urzędów statystycznych. Wskazane zostały również najistotniejsze problemy i wyzwania – aktualne i przyszłe – z zakresu edukacji dotyczącej efektywnej ochrony danych.

Ostatnią sesję naukową, poświęconą ekonomii matematycznej, prowadziła dr Karolina Sobczak (UEP). W wystąpieniu *Przyrost różnorodności gospodarczej*

a odsetek innowatorów w populacji producentów Elżbieta Pliś (UE w Krakowie) określiła, jak potencjalny wzrost różnorodności gospodarczej związany z wprowadzaniem innowacji zmienia strukturę systemu gospodarczego, a w szczególności – jak wpływa na odsetek innowatorów w populacji producentów. Rozwój gospodarczy, do którego przyczyniają się innowacje, poddano badaniu mikroekonomicznemu. Analiza strategii ewolucyjnie stabilnych pozwoliła stwierdzić, że w niektórych przypadkach struktura populacji producentów składającej się zarówno z innowatorów, jak i imitatorów pozostaje stabilna, a na odsetek innowatorów ma wpływ zmiana jakościowa zachodząca w systemie produkcji mierzona przyrostem różnorodności związanej z preferencjami konsumentów.

Sylwia Radomska (doktorantka, UW) w referacie *Investment in human capital: an optimal taxation approach* omówiła dwa źródła inwestowania w kapitał ludzki: prywatne i publiczne. Wskazała ich wady i zalety, a następnie zaproponowała nowe podejście – wykorzystujące wnioskowanie normatywne (ewaluacja ex ante) – do badania dwóch standardowych instrumentów, czyli stypendiów edukacyjnych i pożyczek zależnych od wysokości dochodu, uwzględniające altruistyczne dynastie gospodarstw domowych, ich nieobserwowalną heterogeniczność i niepewność dochodów. Przedstawiła także wpływ własności opracowanego modelu na wyniki dotyczące dobrobytu i nierówności społecznych dzięki porównaniu dwóch sytuacji: gdy subsydia edukacyjne i pożyczki są udzielane i gdy nie są dostępne.

Ewa Synówka (UZ, US w Zielonej Górze) zaprezentowała referat *Koniunktura w gospodarstwach rolnych a zmiany wartości wybranych wskaźników makroekonomicznych – badanie przyczynowości w sensie Grangera*. Prelegentka dokonała analizy przyczynowości w sensie Grangera między wskaźnikami koniunktury w gospodarstwach rolnych, wskaźnikami systemu rachunków narodowych i przeciętnymi cenami skupu produktów rolnych. Badaniem omówionym w wystąpieniu objęto polską gospodarkę w latach 2012–2022. Wykorzystano w nim półroczne dane opracowane przez Ośrodek Badań Koniunktury US w Zielonej Górze na podstawie ankiet koniunktury w gospodarstwach rolnych, a także kwartalne i półroczne dane z Dziedzinowych Baz Wiedzy i Banku Danych Makroekonomicznych opublikowane przez GUS.

Równolegle odbywały się dwie sesje studenckie, którym przewodniczyli Krzysztof Walczak i Elżbieta Rychłowska-Musiał. Cyryl Marek Leszczyński (UEP) w referacie *Uczenie przez wzmacnianie w środowisku Unity3D* omówił badanie, którego przedmiotem była implementacja algorytmów uczenia przez wzmacnianie w trójwymiarowym środowisku silnika Unity3D, najczęściej wykorzystywanego do usprawniania

produkcji gier. Pomimo wypunktowanych w prezentacji niedoskonałości potencjał Unity3D w roli takiego narzędzia jest bardzo duży, zwłaszcza w przypadku problemów wymagających dobrego wizualnego odzwierciedlenia rzeczywistości (np. uczenie robota z sensorem optycznym).

Wystąpienie Jeremiego Ranosza, Stanisława Kumora, Julii Głowaczewskiej, Patryka Garwola i Mateusza Kaczmarka (UEP) pt. *Wspieranie nauki przedmiotów STEM za pomocą zintegrowanego systemu opartego o gry 3D* dotyczyło metody integrującej aspekty techniczne i społeczne, wspomagającej nauczanie przedmiotów z grupy STEM (*science*, *technology*, *engineering*, *mathematics*) w szkołach podstawowych. Łączy ona gry edukacyjne 3D z tradycyjnymi narzędziami dydaktycznymi. Aspekt techniczny uwzględnia wieloplatformowe gry edukacyjne, a w aspekcie społecznym promuje się wymianę informacji między nauczycielami, uczniami i twórcami gier. Celem zaprezentowanej metody jest zwiększenie efektywności procesu kształcenia, a zarazem utrzymanie uwagi uczniów.

Referat *Wykorzystanie nowoczesnych technologii w kształtowaniu poglądów ekologicznych wśród najmłodszych* wygłosili Natalia Sylwia Marszał, Agnieszka Popiel i Jakub Mikołaj Ścieszka (UEP). Studenci przedstawili propozycję gry edukacyjnej o tematyce ekologicznej wykorzystującej zabawki, które pozwalają na przeniesienie fizycznej figurki do świata wzbogaconej rzeczywistości, tj. łączącej świat rzeczywisty z generowanym komputerowo za pomocą smartfona. W założeniu seria zabawek ma być wykonana z plastiku pochodzącego z recyklingu z wykorzystaniem metod druku 3D. Bohater gry wchodzi w interakcję z wirtualnym środowiskiem, co ma na celu zwiększenie świadomości ekologicznej graczy również w codziennym życiu.

Krzysztof Kaczmarek, Jan Matera i Aleksandra Panek (UEP) pod kierownictwem Alicji Jajko-Siwek w wystąpieniu *Związek pomiędzy zarobkami a konsumowaniem sztuki wysokiej* omówili badanie mające dać odpowiedź na pytanie, czy nadal istnieje związek między statusem finansowym a uczestnictwem w życiu kulturalnym, a dokładniej – między wysokością zarobków a konsumpcją w sferze kultury. Uzyskane wyniki wykazały występowanie dodatniej korelacji między logarytmem zarobków a korzystaniem z dóbr kultury, jak również zróżnicowanie siły tej zależności w badanych latach.

Ostatni referat w sesji studenckiej – *Modelling gold price volatility: a Monte Carlo simulation approach with financial econometric insights* – wygłosił Zakarias Larsson (UEP), który zaprezentował wyniki badania zmienności cen złota z wykorzystaniem symulacji Monte Carlo i modeli ekonometrycznych. Prelegent powiązał zmienność cen z szeregiem parametrów rynkowych, takich jak inflacja, wskaźniki makroekonomiczne i stopy procentowe.

W trakcie konferencji przeprowadzono konkurs na najlepszy referat w sesji studenckiej. Do głosowania – online lub anonimowo na kartkach zbieranych w sekretariacie – byli uprawnieni wszyscy uczestnicy konferencji. Za najlepszy uznano referat *Wykorzystanie nowoczesnych technologii w kształtowaniu poglądów ekologicznych wśród najmłodszych* Natalii Sylwii Marszał, Agnieszki Popiel i Jakuba Mikołaja Ścieszki.

Krzysztof Malaga
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Poland

Jarogniew Rykowski
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Poland

Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej; Urząd Statystyczny w Poznaniu, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics; Statistical Office in Poznań, Poland

Krzysztof Węcel
Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej, Polska
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Poland

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(10), 80–81
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

WYDAWNICTWA GUS. WRZESIEŃ 2023
PUBLICATIONS OF STATISTICS POLAND. SEPTEMBER 2023

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

Sytuacja makroekonomiczna w Polsce na tle procesów w gospodarce światowej w 2022 r.
Macroeconomic situation in Poland in the context of the world economic processes in 2022

Najnowsza (dziesiąta) edycja opracowania analitycznego poświęconego zjawiskom makroekonomicznym w Polsce w kontekście uwarunkowań gospodarki krajów UE i świata. Dokonano w nim analizy podstawowych zagadnień społeczno-gospodarczych.

Język: polski (przedmowa, spis treści, tytuły rozdziałów i podrozdziałów oraz tablice i wykresy również w języku angielskim)
Language: Polish (preface, contents, titles of chapters and subchapters, tables and charts available also in English)
Seria: Analizy statystyczne
Series: Statistical analyses
Dostępne wersje: drukowana i elektroniczna
Available in: printed and electronic form

We wprowadzeniu omówiono zmiany w podejściu do polityki makroekonomicznej związane z wybuchem wojny w Ukrainie i ze wzrostem inflacji. Zasadnicza część publikacji składa się z czterech rozdziałów poświęconych: sytuacji makroekonomicznej (ze szczególnym uwzględnieniem wzrostu gospodarczego oraz kształtowania się cyklu koniunkturalnego w głównych gospodarkach światowych i unijnych), rynkowi pracy i sytuacji dochodowej gospodarstw domowych, procesom w obszarze finansów publicznych oraz sytuacji na rynkach finansowych. Uzupełnieniem są dwa aneksy, obejmujące m.in. problematykę zastosowania danych KLEMS do badań produktywności w sektorach.

W publikacji wykorzystano dane pochodzące z wielu źródeł, przede wszystkim GUS, a także Komisji Europejskiej, Międzynarodowego Funduszu Walutowego, Organizacji Współpracy Gospodarczej i Rozwoju, Międzynarodowej Organizacji Pracy, Konferencji Narodów Zjednoczonych ds. Handlu i Rozwoju, Banku Świato-

wego, Europejskiego Urzędu Nadzoru Bankowego, Narodowego Banku Polskiego, Komisji Nadzoru Finansowego, Giełdy Papierów Wartościowych w Warszawie i Ministerstwa Finansów. Zamieszczono również wyniki badań o charakterze eksperymentalnym, niestanowiące oficjalnych danych statystycznych.

We wrześniu br. ukazały się ponadto:

- *Bezrobocie rejestrowane 1–2 kwartał 2023 r.*;
- „Biuletyn statystyczny" nr 8/2023;
- *Budżety gospodarstw domowych w 2022 r.*;
- *Ceny robót budowlano-montażowych i obiektów budowlanych (lipiec 2023 r.)*;
- *Fizyczne rozmiary produkcji zwierzęcej w 2022 r.*;
- *Koniunktura w przetwórstwie przemysłowym, budownictwie, handlu i usługach 2000–2023 (wrzesień 2023 r.)*;
- *Kultura i dziedzictwo narodowe w 2022 r.*;
- *Nakłady i wyniki przemysłu – 1–2 kwartał 2023 r.*;
- *Produkcja ważniejszych wyrobów przemysłowych w sierpniu 2023 r.*;
- *Równoległy oraz wyprzedzający zagregowany wskaźnik koniunktury, zegar koniunktury – szereg do lipca 2023 r.*;
- „Statistics in Transition new series" nr 4/2023;
- *Sytuacja demograficzna Polski do 2022 r.*;
- *Sytuacja społeczno-gospodarcza kraju w sierpniu 2023 r.*;
- „Sytuacja społeczno-gospodarcza województw" nr 2/2023;
- *Telekomunikacja 2022*;
- *Transport – wyniki działalności w 2022 r.*;
- *Uniwersytety trzeciego wieku w roku akademickim 2021/2022*;
- „Wiadomości Statystyczne. The Polish Statistician" nr 9/2023;
- *Zatrudnienie i wynagrodzenia w gospodarce narodowej w pierwszym półroczu 2023 r.*

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

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